Dual Mutual Information Constraints for Discriminative Clustering

Hongyu Li[1], **Lefei Zhang** [1,2], **Kehua Su** [1*]

[1]School of Computer Science, Wuhan University, Wuhan, 430072, P. R. China
[2]Hubei Luojia Laboratory, Wuhan 430072, P. R. China
{hongyuli, zhanglefei, skh}@whu.edu.cn

Abstract

Deep clustering is a fundamental task in machine learning and data mining that aims at learning clustering-oriented feature representations. In previous studies, most of deep clustering methods follow the idea of self-supervised representation learning by maximizing the consistency of all similar instance pairs while ignoring the effect of feature redundancy on clustering performance. In this paper, to address the above issue, we design a dual mutual information constrained clustering method named DMICC which is based on deep contrastive clustering architecture, in which the dual mutual information constraints are particularly employed with solid theoretical guarantees and experimental validations. Specifically, at the feature level, we reduce the redundancy among features by minimizing the mutual information across all the dimensionalities to encourage the neural network to extract more discriminative features. At the instance level, we maximize the mutual information of the similar instance pairs to obtain more unbiased and robust representations. The dual mutual information constraints happen simultaneously and thus complement each other to jointly optimize better features that are suitable for the clustering task. We also prove that our adopted mutual information constraints are superior in feature extraction, and the proposed dual mutual information constraints are clearly bounded and thus solvable. Extensive experiments on five benchmark datasets show that our proposed approach outperforms most other clustering algorithms. The code is available at https://github.com/Li-Hyn/DMICC.

Introduction

As a fundamental task in unsupervised learning, data clustering plays an important role in various artificial intelligence applications. The objective of clustering is to assign the instances into certain groups, such that similar samples belong to the same cluster, whereas the dissimilar ones are distributed in different clusters. Although some data clustering techniques have achieved promising results (Nie et al. 2011, 2016; Chen et al. 2022; Xu et al. 2022), conventional clustering methods usually result in poor performance on complex and high-dimensional data (e.g., image clustering) due to the weak capability of feature representations and the subsequent inefficiency similarity measurement among the instances. Deep neural networks (DNNs) can be used to transform the data into more clustering-friendly representations due to their inherent property of highly nonlinear transformation (Wen et al. 2021; Wang et al. 2021a; Dizaji et al. 2017; Huang et al. 2021; Ma et al. 2019). Conversely, as an alternative form of self-supervised feature learning, contrastive learning has been recently applied to the field of clustering (Li et al. 2021).

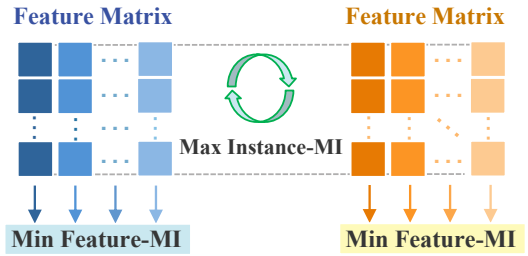


Figure 1: Dual mutual information (MI) constraints. We perform MI constraints at the instance and feature levels to obtain better clustering efficacy.

Although these existing data clustering algorithms have achieved promising progress, we believe that the following essential factors should be crucial for clustering performance (i.e., the minimization of redundancy among features and maximization of similarities among instances). Previous works (Chang et al. 2017; Ji, Vedaldi, and Henriques 2019; Wang et al. 2021b; Do, Tran, and Venkatesh 2021; Ji et al. 2021, 2022) have focused more on the learning of inter-instance similarities while ignoring the redundancy of correlation among instance features. Although this point has been noted in some recent studies (Tao, Takagi, and Nakata 2021), the approach they attempt to address redundancy issues fails to reach the feature information level. Therefore, advanced strategies can be explored for better feature deredundancy to improve clustering effectiveness.

In this paper, we explore the essence of the issue and propose a deep comparative learning clustering technique to obtain clustering-friendly features by performing two mutual information (MI) constraints. We then apply a simple k-means strategy to achieve better clustering performance. Specifically, we add the corresponding MI constraints at the

*Corresponding author.

feature and instance levels. The core of the proposed MI constraints is shown in Fig. 1. At the feature level, we reduce the redundancy among features by minimizing the MI across all the dimensionalities to encourage the neural network to extract more discriminative features. In other words, we perform an MI minimization constraint at the feature level to remove the redundant information among feature dimensions. In comparison with the study of (Tao, Takagi, and Nakata 2021), we propose a more flexible approach by feature-level MI (FMI). We mainly focus on whether the feature-level information used for clustering is sufficiently discriminative. To the best of our knowledge, we are the first to propose a method for FMI minimization in contrastive clustering. At the instance level, we extend the discriminative model presented by (Wu et al. 2018) with an MI maximization constraint among instance pairs. This constraint allows the original images to be consistent at the instance level despite the applied augmentation methods, thereby obtaining more unbiased and robust representations and achieving our goal of maximizing similarities among instances. Our model is superior in instance similarity enhancement compared with the pure discriminative model. The two MI constraints do not exist separately; they complement each other and work jointly. Our design achieves a stable improvement in clustering effectiveness.

The rest of the paper is organized as follows. We first summarize the recent works of deep clustering and MI in the following section. Then, we present a detailed description of the dual MI constrained clustering (DMICC) method. The experimental results and analysis of five benchmark datasets are reported in Section Experiments, followed by conclusions and future remarks. The main contributions can be summarized as follows.

- At the feature level, we propose a novel feature-level MI (FMI) minimization constraint for information de-redundancy, which leads to more discriminative features. Moreover, we provide clear proof that the FMI strategy obviously outperforms the simple feature orthogonal approach.
- At the instance level, we introduce instance-level MI (IMI) maximization constraint that can be combined with discriminative methods to further enhance the consistency of all the similar instance pairs, which results in more unbiased and robust feature representations.
- We theoretically show the boundedness of the employed MI constraints, thereby demonstrating that our constraints are solvable.

Related Works

Deep Clustering

Deep clustering aims at clustering unstructured data or high-dimensional data with deep neural networks (DNNs). Recently, some clustering methods (Ma et al. 2018; Xie et al. 2020) based on representation learning have also been proposed. These methods outperform traditional algorithms and, in some cases, capture supervised learning results. (Van Gansbeke et al. 2020) achieves wonderful results with end-to-end learning which combines feature learning and clustering. At the same time, some works have realized that learning discriminative instance representations in deep clustering is important to the final clustering works, like (Bojanowski and Joulin 2017; Caron et al. 2018; Donahue, Krähenbühl, and Darrell 2017). Especially, based on the observation of the results of the ImageNet dataset, (Wu et al. 2018) found that the apparent similarity comes not from the semantic annotations, but the images themselves, that is to say, discriminating individual instance classes leads to learning representations that retain similarities among data.

Through the existing literature, we can identify that the key to get good clustering results lies in instance similarity. Many recent studies have also applied MI to the field of deep clustering and obtained good clustering performance. (Hu et al. 2017) extended the MI maximization clustering algorithm to deep clustering by DNNs. (Ji, Vedaldi, and Henriques 2019) directly learned semantic labels without learning representations based on MI between image pairs.

All of the above methods have achieved clustering improvement by enhancing instance similarities, and many existing studies have demonstrated that feature de-redundancy of instances also leads to clustering improvement based on the aforementioned works (i.e., (Tao, Takagi, and Nakata 2021)). Our work is motivated by exploring the relationship between clustering and MI, and we believe that the improvement of clustering performance comes from two efforts (i.e., the minimization of redundancy among feature dimensions and the maximization of similarities among instances). The essence of reducing feature redundancy is that each feature needs to be more independent, and the MI of these features should thus be minimized.

MI

MI methods have been widely used in several domains, such as unsupervised feature representation (Wen et al. 2020; Kong et al. 2020) and feature selection (Roy et al. 2020; Schnapp and Sabato 2021), which allow learning view-invariant representations and thus obtaining invariant semantic information important for downstream tasks. The MI measures the dependencies between random variables.

Given two random variables x and y, their probability density functions are denoted as $p(x)$ and $p(y)$. From the perspective of the original definition, the MI can be expressed as follows.

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy. \tag{1}$$

For MI, bounds are crucial. When we utilize MI, we need to demonstrate its boundedness. However, the MI bounds are difficult to calculate directly; some common methods turn to maximize the InfoNCE loss of the lower bound of the MI (Chen et al. 2020b; He et al. 2020).

As for the application of MI in the field of clustering, such as the method proposed by (Do, Tran, and Venkatesh 2021), which combines joint representation learning with clustering to achieve good performance, and the method named DFC (Pang et al. 2020) proposes extended MI (EMI) between input data. Different from previous approaches, our approach

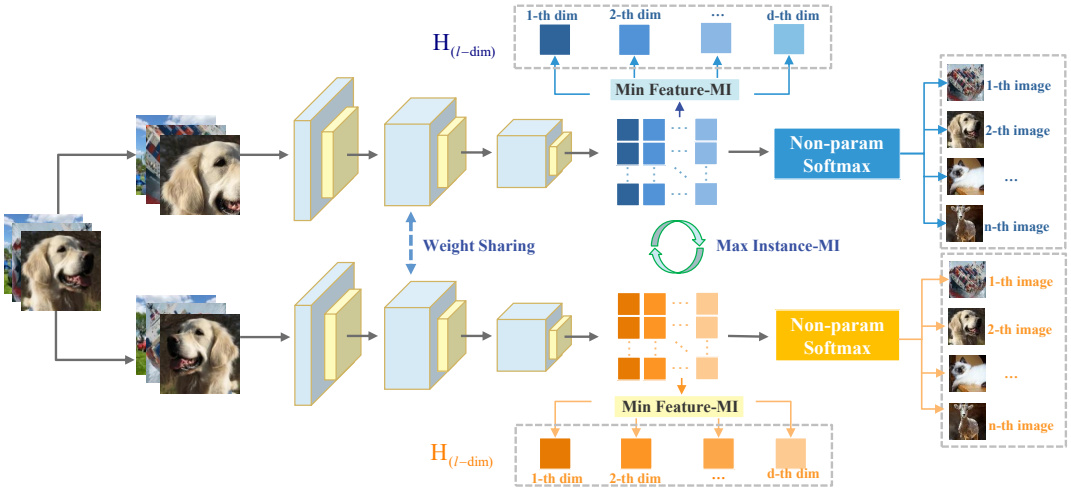


Figure 2: Framework of the proposed DMICC approach. We initially construct data pairs by two data augmentations. Then, we use a parameter-shared network for extracting features from different data augmentations. We perform IMI constraints on the features extracted from the two branches with non-parametric SoftMax and FMI constraints on each branch to learn discriminative features.

innovatively proposes dual MI constraints at two levels by maximizing IMI and minimizing FMI to jointly contribute to our clustering goal. Furthermore, we prove the boundedness of the proposed dual MI constraints.

Method

As shown in Fig. 2, our approach mainly consists of three modules: the feature extraction network, the feature redundancy-minimization module, and the instance similarity-maximization module, where the instance similarity-maximization module is made of two components (i.e., instance discrimination backbone and IMI constraint). In brief, we initially construct data pairs through the feature extraction network and extract features from the augmented samples, and utilize the IMI constraint to keep the extracted features more unbiased and robust. Then, we learn the representations that capture instance similarities via the instance discrimination backbone. The feature redundancy-minimization module also further contributes to the effect of obtaining more independent representations. After receiving the final representations, we utilize simple k-means to output the final clustering results. We will elaborate the two main components in turn and present the proposed objective function at the end.

Feature Redundancy-Minimization Module

FMI Constraint Our work focuses on feature redundancy and innovatively implements an MI minimization constraint at the feature level. We aim to effectively reduce the redundancy among feature dimensions in this way to obtain more discriminative features that contribute to the final clustering efficacy.

In information theory, entropy is a measure of uncertainty, and here we denote it as H. For a random variable x, its entropy can be expressed as $\mathrm{H(x)} = -\Sigma_{\mathrm{x}} p(\mathrm{x}) \log p(\mathrm{x})$. $p(\mathrm{x})$ denotes the probability of x occurring. When we are concerned with the direct uncertainty relationship between two random variables, conditional entropy can be expressed as $H(Y \mid X) = \sum_{x,y} p\left(x_i y_j\right) I\left(y_j \mid x_i\right)$. By combining the definition of entropy H and conditional entropy, we can obtain the formula for the MI of the following form with calculation and simplification.

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{2}$$

We denote the two feature matrices as F^1 and F^2 and use f_l^1 and f_l^2 to denote the lth column of F^1 and F^2, respectively. Take one feature matrix F as an example. In the DNNs, we denote the size of F as $[b, d]$, where b is the number of the batch size, and d represents the number of dimensions. $F = [f_1, f_2, ..., f_d]$. We achieve the constraint of MI minimization at the feature level, that is,

$$\min \frac{1}{d^2} \sum_{i=1,j=1}^{d} I\left(f_i, f_j\right). \tag{3}$$

According to Eq.(2), we initially need to denote the joint probability distribution $P\left(f_i, f_j\right)$ of feature dimension f_i and f_j and the marginal probability distribution $P(f_i)$ and $P(f_j)$. Through our observation, the degree of dimensional correlation and the joint probability distribution $P\left(f_i, f_j\right)$ are strongly correlated. That is, the value of the joint probability distribution is higher where the dimensional correlation is significant. Thus, we assume that the joint probability distribution $P\left(f_i, f_j\right)$ can be replaced by the dimensional correlation value to some extent. We can initially construct the covariance matrix $cov(f_i, f_j)$ and normalize it, with the final result being our joint probability distribution $P\left(f_i, f_j\right)$.

After completing the normalization of F, we simply multiply F with its transpose F^T to obtain the matrix C, with a size of $[d, d]$.

With the above assumptions, the joint probability distribution $P(f_i, f_j)$ can be stated as $P(f_i, f_j) = \frac{C(f_i,f_j)}{\text{sum}(C)}$. We denote the sum of the matrix C as $sum(C)$. The marginals $P(f_i) = \sum_{j=1}^{d} P(f_i, f_j)$ and $P(f_j) = \sum_{i=1}^{d} P(f_i, f_j)$ can be obtained by summing over the rows and columns of the matrix C.

In summary, our FMI constraint loss $\mathcal{L}_{FMI}$ can be represented as:

$$\frac{1}{d^2}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{C(f_i,f_j)}{\text{sum}(C)}\cdot\log\frac{C(f_i,f_j)\,\text{sum}(C)}{\eta^2\cdot\sum_{j=1}^{d}C(f_i,f_j)\sum_{i=1}^{d}C(f_i,f_j)}, \tag{4}$$

where η is a balance hyperparameter used to relax the marginal probability distribution, such that $\mathcal{L}_{FMI}$ can be relaxed either.

Proof of Method Validity We wish to demonstrate that the method with MI constraint at the feature level is superior to the method with simple feature orthogonality. For this purpose, we compare the two on the same baseline. That is, we uniformly restrict values between $[0, 1]$ after feature orthogonality. Generally, our hyperparameter η takes a value that satisfies the range of $\left[-\sqrt{\varepsilon}, -\sqrt{\frac{d}{\varepsilon e^{\varepsilon d^2(1+\delta)}}}\right] \cup \left[\sqrt{\frac{d}{\varepsilon e^{\varepsilon d^2(1+\delta)}}}, \sqrt{\varepsilon}\right]$, where ε and δ represent small constants aiming to keep the matrix non-negative. Mathematically, the objective is as follows:

$$\|I(f_i,f_j)\| \leqslant \|FO(f_i,f_j)\|, \tag{5}$$

where $i \neq j$. Eq. (5) can be further expanded as:

$$\left\|\frac{C(f_i,f_j)}{\text{sum}(C)}\log\frac{C(f_i,f_j)\,\text{sum}(C)}{\eta^2\cdot\sum_{j=1}^{d}C(f_i,f_j)\sum_{i=1}^{d}C(f_i,f_j)}\right\| \leqslant \left\|\frac{f_i^{\top}f_j}{\Omega}\right\|, \tag{6}$$

where Ω is a scaling parameter, helping the right side achieve regularization to keep both sides comparable. The formula $\left\|\frac{C(f_i,f_j)}{\text{sum}(c)}\right\|$ can be scaled into $\frac{1}{d^2(1+\delta)}$ with the assumption of consistent distribution when converged. We can easily obtain $\frac{C(f_i,f_j)\cdot\text{sum}(C)}{\eta^2\sum_{j=1}^{d}C(f_i,f_j)\sum_{i=1}^{d}C(f_i,,f_j)} \geqslant 1$ through the range of the hyperparameter η. Eq.(6) can be translated into the following objective:

$$\left(\frac{C(f_i,f_j)\cdot\text{sum}(C)}{\sum_{j=1}^{d}C(f_i,f_j)\sum_{i=1}^{d}C(f_i,f_j)}\right)_{\max} \leqslant \left(\eta^2 e^{\varepsilon d^2(1+\delta)}\right)_{\min}. \tag{7}$$

We can easily obtain that $\left(\eta^2 e^{\varepsilon d^2(1+\delta)}\right)_{\min} = \frac{d}{\varepsilon}$. The equation on the left can be performed with the following scaling to find the upper limit:

$$\frac{C(f_i,f_j)\cdot\text{sum}(C)}{\sum_{j=1}^{d}C(f_i,f_j)\cdot\sum_{i=1}^{d}C(f_i,f_j)} \leqslant \frac{\text{sum}(C)_{\max}}{\left(\sum_{i=1}^{d}C(f_i,f_j)\right)_{\min}} \leqslant \frac{d}{\varepsilon}. \tag{8}$$

Therefore, Eq.(5) holds. That is to say, the approach with FMI constraint is superior to the method with simple feature orthogonality. Please refer to the supplementary material for detailed proof.

Algorithm 1: DMICC

Input: Dataset X; training epochs E; Batch size b; temperature parameter τ; cluster number N; hyper-parameters λ_1, λ_2; structure of $\mathcal{T}$, f.
Output: The clustering result O.

for $epoch = 1$ in E **do**
  sample a batch $\{x_i\}_{i=1}^{b}$ from X
  two augmentations on the same batch of images as $T^1, T^2 \sim \mathcal{T}$
  compute the representations F^1 and F^2 by $v_i^1 = f\left(T^1(x_i)\right), v_i^2 = f\left(T^2(x_i)\right)$
  compute Feature-level MI constraint loss $\mathcal{L}_{FMI}$ through Eq.(4)
  compute Instance Discrimination loss $\mathcal{L}_{ID}$ through Eq.(10)
  compute Instance-level MI constraint loss $\mathcal{L}_{IMI}$ through Eq.(13)
  update f through gradient descent to minimize $\mathcal{L}$ in Eq.(14)
end for
Obtain the clustering results O with the final F^1 or F^2 by simple k-means of N clusters
return O

Instance Similarity-Maximization Module

Instance discrimination Backbone We introduce the approach of instance discrimination proposed by (Wu et al. 2018). The key idea of instance discrimination is that every instance is assumed to represent a distinct class. Suppose we have n images $x_1, ..., x_n$ in n class and their corresponding features $v_1, ..., v_n$. v_i can also replace the weight vector. v is enforced $\|v\| = 1$ via L_2-normalization layer. The probability $P(i \mid v)$ is defined as follows:

$$P(i \mid v) = \frac{\exp\left(v_i^T v/\tau\right)}{\sum_{j=1}^{n}\exp\left(v_j^T v/\tau\right)}, \tag{9}$$

where τ is a temperature parameter controlling the concentration level of the distribution (Hinton, Vinyals, and Dean 2015).

The work focuses on learning an embedding function $v = f_\theta(x)$, where f_θ is modeled as a DNN with parameter θ. The objection function aims to maximize joint probability $\prod_{i=1}^{n} P_\theta(i \mid f_\theta(x_i))$ as:

$$\mathcal{L}_{ID}(\theta) = -\sum_{i=1}^{n}\log P(i \mid f_\theta(x_i)). \tag{10}$$

IMI Constraint Although instance discrimination has yielded relatively good instance similarity, MI theory can further enhance instance similarity. Inspired by previous works (Lin et al. 2021; Ji, Vedaldi, and Henriques 2019), we attempt to introduce contrastive learning to maximize the MI constraint on the augmented instance pairs generated from the same set of images, resulting in more unbiased and robust feature representations.

We denote the original image sets as X, and the two augmentations as X^1, X^2 with corresponding representations matrices F^1, F^2. $F^1 = [v_1^1; v_2^1; ...; v_b^1]$ and $F^2 =$

$[v_1^2; v_2^2; ...; v_b^2]$, where $v_i^1 = f_\theta\left(x_i^1\right)$ and $v_i^2 = f_\theta\left(x_i^2\right)$. Our goal is to maximize the MI of the corresponding representations:

$$\max I\left(F^1, F^2\right). \tag{11}$$

Given that we construct a model with shared parameters, the sizes of feature representations on two branches are both $[b, d]$. The joint distribution is donated as $P = \frac{1}{b}\sum_{i=1}^{b} F^1 \cdot \left(F^2\right)^\top$. The marginals $P\left(v_i^1\right) = \sum_{j=1}^{d} P\left(v_i^1, v_j^2\right)$ and $P\left(v_j^2\right) = \sum_{i=1}^{d} P\left(v_i^1, v_j^2\right)$ can be obtained by summing up the rows or the columns. In view of the symmetric problems, P is symmetrized using $\left(P + P^\top\right)/2$. The IMI can be obtained as follows:

$$I\left(F^1, F^2\right) = \sum_{i=1}^{d}\sum_{j=1}^{d} P_{v_i^1 v_j^2} \ln \frac{P_{v_i^1 v_j^2}}{P_{v_i^1} \cdot P_{v_j^2}}. \tag{12}$$

Combined with Eq.(12), the IMI constraint loss can be expressed as follows:

$$\mathcal{L}_{IMI} = -\sum_{i=1}^{d}\sum_{j=1}^{d} P_{v_i^1 v_j^2} \ln \frac{P_{v_i^1 v_j^2}}{\gamma^2 P_{v_i^1} \cdot P_{v_j^2}}, \tag{13}$$

γ is a non-zero constant. In our experiments, original MI solutions may receive trivial solutions. Therefore, we consider further increasing the share of entropy by relaxing the edge distribution to avoid trivial solutions.

Objective Function

Without loss of generality, we propose the following objective function:

$$\mathcal{L} = \mathcal{L}_{ID} + \lambda_1 \mathcal{L}_{FMI} + \lambda_2 \mathcal{L}_{IMI}, \tag{14}$$

where $\mathcal{L}_{ID}$, $\mathcal{L}_{IMI}$, and $\mathcal{L}_{FMI}$ are the instance discriminant loss, instance similarity maximization loss, and feature redundancy minimization loss, respectively. Parameters λ_1 and λ_2 are the balance hyperparameters of $\mathcal{L}_{FMI}$ and $\mathcal{L}_{IMI}$, respectively.

Boundedness Proof of the Objective Function

To prove that our objective function is bounded and solvable, we initially need to rewrite the objective function Eq. (14) as

$$\min_\theta \mathcal{L}_{ID}(\boldsymbol{\theta}) + \lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{d^2} I\left(f_i, f_j\right) - \lambda_2 \sum_{i=1}^{d}\sum_{j=1}^{d} I\left(F^1, F^2\right). \tag{15}$$

Obviously, the last two terms are completely independent of θ when we take θ as the minimization parameter. First, we need to prove that the latter two have upper bounds. We prove that $\lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{d^2} I\left(f_i, f_j\right)$ has an upper bound in brief. Generally, our hyperparameters η and γ take values that satisfy the same range. The item $\lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{d^2} I\left(f_i, f_j\right)$ is bounded, which is equivalent to proving that $I\left(f_i, f_j\right)$ is bounded, that is,

$$\begin{aligned}
&\left\| \frac{C_{(f_i, f_j)}}{\mathrm{sum}(C)} \cdot \log \frac{C\left(f_i, f_j\right) \cdot \mathrm{sum}(C)}{\eta^2 \cdot \sum_{j=1}^{d} C\left(f_i, f_j\right) \cdot \sum_{i=1}^{d} C\left(f_i, f_j\right)} \right\| \\
&\leq \frac{\left\| C\left(f_i, f_j\right) \right\|}{\eta^2 \cdot \left\| \sum_{j=1}^{d} C\left(f_i, f_j\right) \right\| \cdot \left\| \sum_{i=1}^{d} C\left(f_i, f_j\right) \right\|} + 1 \\
&\leq \frac{1}{\eta^2 \cdot \left\| \sum_{i=1}^{d} C\left(f_i, f_j\right) \right\|} + 1 \leqslant \frac{1}{\eta^2 \varepsilon d} + 1.
\end{aligned} \tag{16}$$

Eq.(16) shows that $I\left(f_i, f_j\right)$ is bounded, which, in turn, shows that $\left\|\lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{d^2} I\left(f_i, f_j\right)\right\|$ is bounded. The proof of boundedness of $\left\|\lambda_2 \sum_{i=1}^{d}\sum_{j=1}^{d} I\left(F^1, F^2\right)\right\|$ is similar. We can naturally obtain that $\left\|\lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{d^2} I\left(f_i, f_j\right)\right\| + \left\|\lambda_2 \sum_{i=1}^{d}\sum_{j=1}^{d} I\left(F^1, F^2\right)\right\| \leq M$, where M represents the bound. According to the triangle inequality $\|x - y\| \leqslant \|x\| + \|y\|$ that considers all norm equivalence, we can obtain the following:

$$\left\| \lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{d^2} I\left(f_i, f_j\right) - \lambda_2 \sum_{i=1}^{d}\sum_{j=1}^{d} I\left(F^1, F^2\right) \right\|_\infty \leq M. \tag{17}$$

Therefore, $\lambda_1 \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{1}{d^2} I\left(f_i, f_j\right) - \lambda_2 \sum_{i=1}^{d}\sum_{j=1}^{d} I\left(F^1, F^2\right)$ is bounded.

Experiments

Datasets

We evaluate the performance of our DMICC approach on five publicly available datasets, including CIFAR-10/100 (Krizhevsky, Hinton et al. 2009), STL-10 (Coates, Ng, and Lee 2011), ImageNet-10 (Deng et al. 2009) and ImageNet-Dogs (Deng et al. 2009). The number of images, clusters, and image size are presented in Table 1.

Dataset	Images	clusters	Image Size
CIFAR-10	50,000	10	32×32×3
CIFAR-100	50,000	20	32×32×3
STL-10	13,000	10	96×96×3
ImageNet-10	13,000	10	224×224×3
ImageNet-Dogs	19,500	15	224×224×3

Table 1: Datasets used in our experiments.

Comparison Methods

We compare our approach with 15 representative state-of-the-art methods on five challenging image benchmarks, namely, clustering with k-means (MacQueen 1967), DAE (Vincent et al. 2010), DeCNN (Zeiler et al. 2010), VAE (Kingma and Welling 2013), DEC (Xie, Girshick, and Farhadi 2016), DAC (Chang et al. 2017), DDC (Chang et al. 2019), DCCM (Wu et al. 2019), IIC (Ji, Vedaldi, and Henriques 2019), ID (Wu et al. 2018), PICA (Huang, Gong,

Dataset	CIFAR-10			CIFAR-100			STL-10			ImageNet-10			ImageNet-Dogs		
Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
k-means	22.9	8.7	4.9	13.0	8.4	2.8	19.2	12.5	6.1	24.1	11.9	5.7	10.5	5.5	2.0
DAE	29.7	25.1	16.3	15.1	11.1	4.6	30.2	22.4	15.2	30.4	20.6	13.8	19.0	10.4	7.8
DeCNN	28.2	24.0	17.4	13.3	9.2	3.8	29.9	22.7	16.2	31.3	18.6	14.2	17.5	9.8	7.3
VAE	29.1	24.5	16.7	15.2	10.8	4.0	28.2	20.0	14.6	33.4	19.3	16.8	17.9	10.7	7.9
DEC	30.1	25.7	16.1	18.5	13.6	5.0	35.9	27.6	18.6	38.1	28.2	20.3	19.5	12.2	7.9
DAC	52.2	39.6	30.6	23.8	18.5	8.8	47.0	36.6	25.7	52.7	39.4	30.2	27.5	21.9	11.1
DDC	52.4	42.4	32.9	-	-	-	48.9	37.1	26.7	57.7	43.3	34.5	-	-	-
DCCM	62.3	49.6	40.8	32.7	28.5	17.3	48.2	37.6	26.2	71.0	60.8	55.5	38.3	32.1	18.2
IIC	61.7	51.1	41.1	25.7	22.5	11.7	59.6	49.6	39.7	-	-	-	-	-	-
ID	44.0	30.9	22.1	26.7	22.1	10.8	51.4	36.2	28.5	63.2	47.8	42.0	36.5	24.8	17.2
PICA	69.6	59.1	51.2	33.7	31.0	17.1	71.3	61.1	53.1	87.0	80.2	76.1	35.2	35.2	20.1
DRC	72.7	62.1	54.7	36.7	35.6	20.8	74.7	64.4	56.9	88.4	83.0	79.8	38.9	38.4	23.3
IDFD	81.5	71.1	66.3	42.5	42.6	26.4	75.6	64.3	57.5	95.4	89.8	90.1	59.1	54.6	41.3
DCDC	69.9	58.5	50.6	34.9	31.0	17.9	73.4	62.1	54.7	87.9	81.7	78.7	36.5	36.0	20.7
EDESC	62.7	46.4	-	38.5	37.0	-	74.5	68.7	-	-	-	-	-	-	-
DMICC	82.8	74.0	69.0	46.8	45.2	29.1	80.0	68.9	62.5	96.2	91.7	91.6	58.7	58.1	43.8

Table 2: Clustering performance on five publicly available image benchmarks in terms of ACC, NMI, and ARI. The results of our approach is presented as DMICC.

and Zhu 2020), IDFD (Tao, Takagi, and Nakata 2021), DRC (Zhong et al. 2020), DCDC (Dang et al. 2021), and EDESC (Cai et al. 2022).

For DAE, DeCNN, VAE, IDFD, and our method, clustering results are obtained via k-means on the features extracted from images. The results of ID and IDFD are cited from (Tao, Takagi, and Nakata 2021), and the results of DCDC, IIC, and EDESC are cited from (Dang et al. 2021; Van Gansbeke et al. 2020; Cai et al. 2022), respectively. We cite the rest results from (Huang, Gong, and Zhu 2020) here.

Evaluation Metrics

For all quantitative evaluations, we use three widely used clustering metrics, namely, accuracy (ACC), normalized MI (NMI), and adjusted Rand index (ARI). All of these metrics scale from 0 to 1, and higher values indicate better efficacy. We convert the results into percentages.

Implementation Details

We refer to the structure of the ResNet and modify the input layer to adapt to different inputs when working on the five standard datasets including two large ImageNet subsets. Following previous works (Khosla et al. 2020; Chen et al. 2020a), we set the dimensionality of the latent feature vector d to 128, and the temperature coefficient τ in the instance discrimination to 2. We use a stochastic gradient descent optimizer at momentum β = 0.9. The learning rate was initialized to 0.05 and then gradually decreased after the first 600 epochs by a factor of 0.5 per 350 epochs. The weight decay 5e-4 is used for all the datasets. The total number of epochs is set to 5000 and the batch size is set to 128. For some large datasets, we try to extend the epochs to 7000 to ensure convergence. To report the stable performance of the approach, we train our model in all datasets with five trials and display the average best results. All experiments are conducted on an NVIDIA RTX 2080Ti GPU. The setting of balance parameters λ_1 and λ_2 are as follows. In CIFAR-10/100, we set the balance parameters λ_1 as 1e-2 and λ_2 as 1e-4; in STL-10, we set λ_1 as 1e-5 and λ_2 as 1e-6; in ImageNet-10/Dogs, we set λ_1 as 1e-5 and λ_2 as 1e-7, respectively.

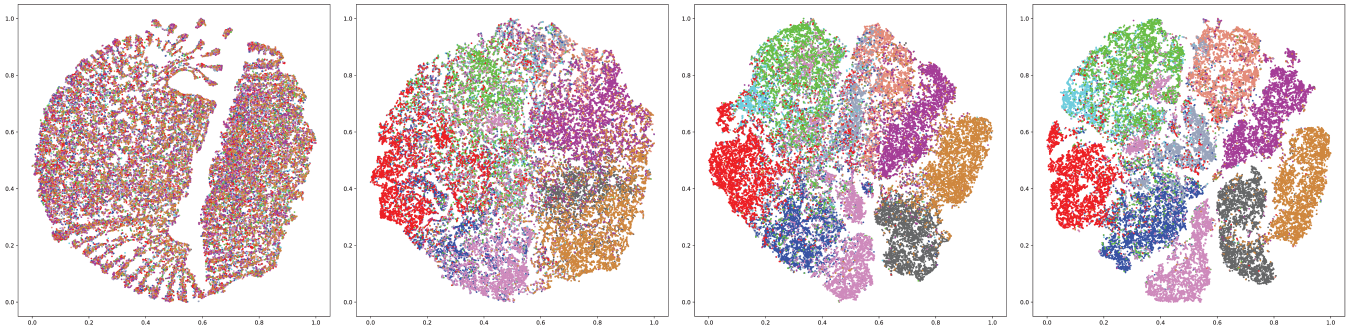
Dataset	ID (original)	ID (tuned)	ID +IMI	ID +FMI	Our method
CIFAR-10	44.0	77.0	79.7	82.0	82.8
	30.9	68.2	69.3	72.8	74.0
	22.1	61.6	64.1	67.4	69.0
ImageNet-10	63.2	93.7	94.2	95.9	96.2
	47.8	86.7	87.3	91.5	91.7
	42.0	86.5	87.5	91.1	91.6

Table 3: Ablation study on the benchmark datasets CIFAR-10 and ImageNet-10.

Experimental Result

To demonstrate the superiority of the proposed method, we show the performance of our DMICC and baselines on the five benchmark datasets in Table 2. The final clustering performance shows that our method consistently outperforms most other methods on the five benchmark datasets.

On the basis of the results shown in Table 2, our method surpasses most methods when the simplest k-means clustering method is applied. The main improvements of our approach can be attributed to the FMI and IMI constraints. Therefore, 1) the FMI constraint is innovatively designed to keep the independence of each feature dimension by reducing the MI among feature dimensions. Thus, the proposed FMI constraint can avoid errors induced by the redundant information among feature dimensions, leading to better clustering performance, especially for fine-grained datasets. 2) The IMI constraint enhances the consistency of all similar instance pairs by maximizing the MI of instance pairs, resulting in more unbiased and robust feature representations. In brief, the unified MI constraints at the instance level and feature level bring effective improvement.



(a) 0 epoch (ACC=0.032) (b) 100 epoch (ACC=0.117) (c) 800 epoch (ACC=0.645) (d) 2400 epoch (ACC=0.807)

Figure 3: Evolution of cluster assignment by t-SNE visualizations of the feature vectors learned by our method on CIFAR-10 dataset. The color indicates the cluster assignments obtained by our proposed DMICC algorithm.

Ablation Study

Effectiveness of FMI Constraints We further conduct ablation studies to verify the effectiveness of FMI constraints and report the results in Table 3 with CIFAR-10 and ImageNet-10 datasets as examples. ID(original) represents the ID method without tuning, ID(tuned) represents our baseline and ID+FMI represents our FMI constraints based on the baseline. Our experimental results indicate that 1) in comparison with the baseline, our proposed method of FMI constraints yields a great improvement on the benchmark datasets, which has 0.5%-5.8% performance improvement in each metric. 2) Our FMI constraints based on the baseline (i.e., ID+FMI) have surpassed most of the current methods, verifying the theoretical validity described in our algorithm.

Effectiveness of IMI Constraints The definitions of ID(original) and ID(tuned) are consistent with the previous definitions. ID+IMI indicates our proposed IMI constraints based on ID(tuned). As summarized in Table 3, the ID+IMI method has certain performance improvement compared with the baseline, indicating that the introduction of contrastive learning and the maximization of IMI have positively affected clustering performance. It is also a valid complement to the baseline method.

Hyperparameter Analysis

Furthermore, we explore the effects of hyperparameters λ_1 and λ_2 on the experimental results. As shown in Eq. (14), we introduce two hyperparameters λ_1 and λ_2 to make a trade-off among instance discrimination, FMI constraints, and IMI constraints. With other experimental settings in agreement, our results are shown in Fig. 4. We observe that 1) for a certain λ_1, a change in λ_2 significantly affects the experimental results. Conversely, λ_1 only needs a simple adjustment to reach good results, which is more beneficial for the final results. 2) λ_1 and λ_2 act together to influence the experimental results, and the performance of the method is good in a wide range of λ_1 and λ_2.

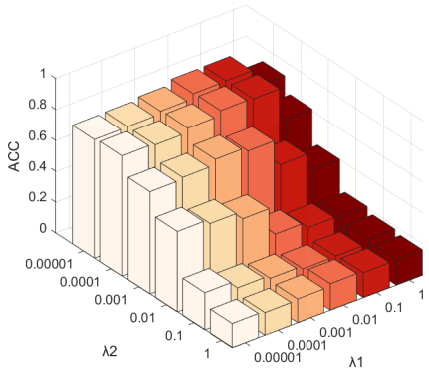


Figure 4: Sensitivity of our proposed DMICC with the variation of λ_1 and λ_2 in the CIFAR-10 dataset.

t-SNE Visualization of Clustering Results

To illustrate the convergence effect of our model more visually, we use the t-SNE to visualize all the instances at four time points during the training process from initialization to half of the target epochs number. Fig.3 shows the scatter plots with the different colors representing various clustering labels. In this figure, the clustering assignments are becoming distinguishable as the number of training sessions increases, which clearly indicates that the proposed DMICC indeed generates discriminative and clustering-friendly feature representations to support the subsequent k-means to achieve better performance.

Conclusion

In this paper, we propose a novel clustering method based on MI. Our method utilizes a combination of fine-grained FMI constraint and coarse-grained IMI constraint to construct a unified cluster-oriented comparative learning framework. To the best of our knowledge, we are the first to propose a method for FMI minimization in contrastive clustering. The proposed DMICC approach shows promising performance in clustering. In the future, we intend to further develop our approach to incorporate pseudo-label generation, such that the representation layer can be associated with the clustering layer. Besides, we also consider extending the model in order to make it adaptable to multi-view clustering.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants No. 62076188 and No. 62272354, and the Special Fund of Hubei Luojia Laboratory under Grant No. 220100014.

References

Bojanowski, P.; and Joulin, A. 2017. Unsupervised Learning by Predicting Noise. In *ICML*, volume 70, 517–526.

Cai, J.; Fan, J.; Guo, W.; Wang, S.; Zhang, Y.; and Zhang, Z. 2022. Efficient Deep Embedded Subspace Clustering. In *CVPR*, 1–10.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep Clustering for Unsupervised Learning of Visual Features. In *ECCV*, volume 11218, 139–156.

Chang, J.; Guo, Y.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2019. Deep discriminative clustering analysis. *arXiv preprint arXiv:1905.01681*.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep Adaptive Image Clustering. In *ICCV*, 5880–5888.

Chen, M.; Huang, L.; Wang, C.; Huang, D.; and Yu, P. S. 2022. Multiview Subspace Clustering With Grouping Effect. *IEEE Trans. Cybern.*, 52(8): 7655–7668.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, volume 119, 1597–1607.

Coates, A.; Ng, A. Y.; and Lee, H. 2011. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *AISTATS*, volume 15, 215–223.

Dang, Z.; Deng, C.; Yang, X.; and Huang, H. 2021. Doubly Contrastive Deep Clustering. *CoRR*, 2103.05484.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Dizaji, K. G.; Herandi, A.; Deng, C.; Cai, W.; and Huang, H. 2017. Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entropy Minimization. In *ICCV*, 5747–5756.

Do, K.; Tran, T.; and Venkatesh, S. 2021. Clustering by Maximizing Mutual Information Across Views. In *ICCV*, 9908–9918.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2017. Adversarial Feature Learning. In *ICLR*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *CVPR*, 9726–9735.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:*, 1503.02531.

Hu, W.; Miyato, T.; Tokui, S.; Matsumoto, E.; and Sugiyama, M. 2017. Learning Discrete Representations via Information Maximizing Self-Augmented Training. In *ICML*, volume 70, 1558–1567.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep semantic clustering by partition confidence maximisation. In *CVPR*, 8849–8858.

Huang, Z.; Ren, Y.; Pu, X.; Pan, L.; Yao, D.; and Yu, G. 2021. Dual self-paced multi-view clustering. *Neural Netw.*, 140: 184–192.

Ji, G.-P.; Fu, K.; Wu, Z.; Fan, D.-P.; Shen, J.; and Shao, L. 2021. Full-Duplex Strategy for Video Object Segmentation. In *ICCV*, 4922–4933.

Ji, G.-P.; Xiao, G.; Chou, Y.-C.; Fan, D.-P.; Zhao, K.; Chen, G.; and Van Gool, L. 2022. Video polyp segmentation: A deep learning perspective. *MIR*, 19(6): 531–549.

Ji, X.; Vedaldi, A.; and Henriques, J. F. 2019. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *ICCV*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NeurIPS*, 33: 18661–18673.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kong, L.; de Masson d'Autume, C.; Yu, L.; Ling, W.; Dai, Z.; and Yogatama, D. 2020. A Mutual Information Maximization Perspective of Language Representation Learning. In *ICLR*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. Technical report, CIFAR.

Li, Y.; Hu, P.; Liu, J. Z.; Peng, D.; Zhou, J. T.; and Peng, X. 2021. Contrastive Clustering. In *AAAI*, 8547–8555.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. COMPLETER: Incomplete Multi-View Clustering via Contrastive Prediction. In *CVPR*, 11174–11183.

Ma, J.; Zhang, Y.; Zhang, L.; Du, B.; and Tao, D. 2019. Pseudo Supervised Matrix Factorization in Discriminative Subspace. In *IJCAI*, 4554–4560.

Ma, S.; Zhang, L.; Hu, W.; Zhang, Y.; Wu, J.; and Li, X. 2018. Self-Representative Manifold Concept Factorization with Adaptive Neighbors for Clustering. In *IJCAI*, 2539–2545.

MacQueen, J. 1967. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, 281–297.

Nie, F.; Wang, X.; Jordan, M. I.; and Huang, H. 2016. The Constrained Laplacian Rank Algorithm for Graph-Based Clustering. In *AAAI*, 1969–1976.

Nie, F.; Zeng, Z.; Tsang, I. W.; Xu, D.; and Zhang, C. 2011. Spectral Embedded Clustering: A Framework for In-Sample and Out-of-Sample Spectral Clustering. *IEEE Trans. Neural Netw.*, 22(11): 1796–1808.

Pang, Y.; Chen, F.; Huang, S.; Ge, Y.; Wang, W.; and Zhang, T. 2020. Deep Fuzzy Clustering with Weighted Intra-class Variance and Extended Mutual Information Regularization. In *ICDM*, 464–471.

Roy, P.; Sharmin, S.; Ali, A. A.; and Shoyaib, M. 2020. Discretization and Feature Selection Based on Bias Corrected

Mutual Information Considering High-Order Dependencies. In *PAKDD*, volume 12084, 830–842.

Schnapp, S.; and Sabato, S. 2021. Active Feature Selection for the Mutual Information Criterion. In *AAAI*, 9497–9504.

Tao, Y.; Takagi, K.; and Nakata, K. 2021. Clustering-friendly Representation Learning via Instance Discrimination and Feature Decorrelation. In *ICLR*.

Van Gansbeke, W.; Vandenhende, S.; Georgoulis, S.; Proesmans, M.; and Van Gool, L. 2020. Scan: Learning to classify images without labels. In *ECCV*, 268–285.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.; and Bottou, L. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11(12).

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2021a. Generative Partial Multi-View Clustering With Adaptive Fusion and Cycle Consistency. *IEEE Trans. Image Process.*, 30: 1771–1783.

Wang, Z.; Li, Z.; Wang, R.; Nie, F.; and Li, X. 2021b. Large Graph Clustering With Simultaneous Spectral Embedding and Discretization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(12): 4426–4440.

Wen, J.; Zhang, Z.; Zhang, Z.; Zhu, L.; Fei, L.; Zhang, B.; and Xu, Y. 2021. Unified Tensor Framework for Incomplete Multi-view Clustering and Missing-view Inferring. In *AAAI*, 10273–10281.

Wen, L.; Zhou, Y.; He, L.; Zhou, M.; and Xu, Z. 2020. Mutual Information Gradient Estimation for Representation Learning. In *ICLR*.

Wu, J.; Long, K.; Wang, F.; Qian, C.; Li, C.; Lin, Z.; and Zha, H. 2019. Deep Comprehensive Correlation Mining for Image Clustering. In *ICCV*, 8149–8158.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *CVPR*, 3733–3742.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *ICML*, 478–487.

Xie, Y.; Liu, J.; Qu, Y.; Tao, D.; and Ma, L. 2020. Robust Kernelized Multiview Self-Representation for Subspace Clustering. *IEEE Trans Neural Netw Learn Syst.*, PP(99): 1–14.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-Level Feature Learning for Contrastive Multi-View Clustering. In *CVPR*, 16051–16060.

Zeiler, M. D.; Krishnan, D.; Taylor, G. W.; and Fergus, R. 2010. Deconvolutional networks. In *CVPR*, 2528–2535.

Zhong, H.; Chen, C.; Jin, Z.; and Hua, X.-S. 2020. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*.