文章编号:2096-1472(2026)10-0055-07
DOI:10.19644/j.cnki.issn2096-1472.2026.010.009

DE3WC:深度嵌入三支聚类

吴昊哲[1],王平心[2]

(1.江苏科技大学计算机学院,江苏 镇江 212003;
2.江苏科技大学理学院,江苏 镇江 212003)
✉ wuhaozhe2025@163.com; pingxin_wang@hotmail.com

摘 要:针对传统聚类方法在特征判别性和样本不确定性处理方面的不足,提出了一种融合深度嵌入特征学习、对比学习、焦点损失与三支决策机制的深度聚类方法 DE3WC。该方法基于改进的深度嵌入聚类框架,在特征提取阶段引入对比损失以增强特征判别性,在聚类阶段结合三支决策思想划分核心域、边界域与琐碎域,进一步引入焦点损失重点优化边界域样本,从而提升聚类性能与模型鲁棒性。在 5 个公开数据集上的实验结果表明,DE3WC 为复杂数据的深度聚类任务提供了一种有效、可推广的新思路。

关键词:深度聚类;三支决策;对比学习;焦点损失
中图分类号:TP391 **文献标识码**:A

DE3WC: Deep Embedded Three-Way Clustering

WU Haozhe[1], WANG Pingxin[2]

(1. *School of Computer, Jiangsu University of Science and Technology, Zhenjiang* 212003, *China*;
2. *School of Science, Jiangsu University of Science and Technology, Zhenjiang* 212003, *China*)
✉ wuhaozhe2025@163.com; pingxin_wang@hotmail.com

Abstract: Addressing the limitations of traditional clustering methods in feature discrimination and handling of sample uncertainty, this paper proposes a deep clustering method, DE3WC, which integrates deep embedding feature learning, contrastive learning, focal loss, and a three-way decision mechanism. The method is built upon an improved deep embedded clustering framework. In the feature extraction stage, contrastive loss is introduced to enhance feature discrimination. In the clustering stage, the concept of three-way decision is incorporated to partition samples into core, boundary, and trivial regions, and focal loss is further introduced to focus on optimizing the boundary region samples, thereby improving clustering performance and model robustness. Extensive experiments conducted on five public datasets demonstrate that DE3WC provides an effective and generalizable new approach for deep clustering tasks with complex data.

Key words: deep clustering; three-way decision; contrastive learning; focal loss

0 引言(Introduction)

聚类作为无监督学习的一种方法,能够在数据集缺乏原始类别标签的时候挖掘出潜在结构的信息。聚类的目标是将具有较高相似度的样本划分到同一簇中,而将相似度较低的样本分配到不同簇中,从而在同一簇内保持较强的相似性,在不同簇间体现出显著的差异性。传统的聚类方法包括 K-means 聚类[1]、层次聚类[2]和谱聚类[3]等多种方法。K-means 聚类方法通过迭代更新簇中心,最小化簇内数据点与簇中心的距离来达到聚类目标。K-means 方法具有计算高效性和易于实现的优点,但由于对初始值敏感且容易陷入局部最优解,使得它在处理复杂数据结构时表现不佳;层次聚类方法通过自底向上或自顶向下的策略逐步合并或拆分数据簇,特别适合于揭示数据的多层次结构,但它的计算复杂度较高,且在处理大规模数据集时效率有限;谱聚类基于数据点之间的相似性矩阵,通过矩阵的谱分解方法实现聚类,具有捕捉非凸簇结构的显著优势,但是它的计算成本随数据规模增加显著增长。聚类作为机器学习中的常用技术已广泛应用于医学图像分割[4]、生物学研究[5]、异常检测、推荐系统和社会科学等众多实际场景中。

三支聚类作为一种新的聚类理论,引起了研究者的广泛关注。三支聚类借助三支决策理论,打破了传统的归属与不归属的双向决策,引入了“可能归属”的概念。将数据空间划分为核心域、边界域和琐碎域,以有效处理数据中的不确定性和模糊性。Yao[6]在研究决策粗糙集的基础上提出了三支决策概念,将实体集划分成正域、负域和边界域,分别对应接受、拒绝以及

收稿日期:2025-11-30
基金项目:国家自然科学基金项目资助(62076111,61773012);江苏省高校自然科学基金项目资助(15KJB110004)

www.rjgczz.com

不承诺的决策规则。这种划分方式较二支决策更符合人类认知模式，能够有效规避样本信息不充分所带来的决策风险。Yu 等[7]将三支决策理论与聚类方法结合，提出了三支聚类框架，该框架用核心域和边界域描述簇结构，定义了属于、不属于和可能属于三种关系。Wang 等[8]结合数学形态学思想，提出了基于收缩与扩张策略的三支聚类理论，并进一步提出了三向 K-means 聚类算法，旨在优化聚类簇之间的边界。Yao[9]在几何学中引入三支决策，深入研究它的几何结构特性。Wen 等[10]针对不完备数据提出了基于三支决策的方法，并展示了该方法在实际应用中的有效性。Hao 等[11]在尺度决策表的背景下，通过三支决策模型选取最优标记，有效地提高了决策精度。D3WC(Deep Three-Way Clustering)[12]方法使用了深度学习网络，融合了证据理论与信息粒化概念，通过信息粒的构建和证据融合技术，显著提升了聚类结果的鲁棒性和准确性。

随着深度学习技术的迅速发展，深度聚类方法在处理高维和复杂数据的过程中表现出了卓越的性能。这类方法通过有效融合深度学习网络与传统聚类策略，从原始高维数据中学习和提取低维的特征表示。其中，深度嵌入聚类(Deep Embedded Clustering，DEC)[13]是一种经典方法，它首先利用自编码器预训练特征，再结合学生 t 分布对特征空间进行软聚类，从而提升了聚类效果。改进的深度嵌入聚类[14]在 DEC 的基础上，通过联合优化自编码器的重构损失和聚类损失，进一步提升了特征表示的质量与聚类的准确性。另外，深度模糊 K-means 聚类方法则引入了模糊理论，赋予数据点对多个簇的隶属度，更有效地处理了聚类过程中边界区域模糊的问题。

本文方法创新性地结合了深度嵌入聚类、对比学习、焦点损失以及三支决策理论，构建了一个全新的深度聚类框架，如图 1 所示。具体而言，本文方法首先在 IDEC 框架基础上，在网络层中加入对比损失来优化重构，在聚类层利用三支决策理论，将聚类结果划分为核心域、边界域和琐碎域，然后引入焦点损失，着重优化难以分类的边界域样本，进而提高聚类模型对困难样本的处理能力。这种方法对样本和聚类之间的关系提供了更全面的理解，更贴近现实世界数据中的模糊性和不确定性，采用差异化的优化策略实现了更精细、更准确的聚类结果。实验验证表明，本文方法在聚类准确性、鲁棒性以及数据内在结构的识别能力方面均表现出显著的优势，体现了它在实际应用中的巨大潜力与优势。

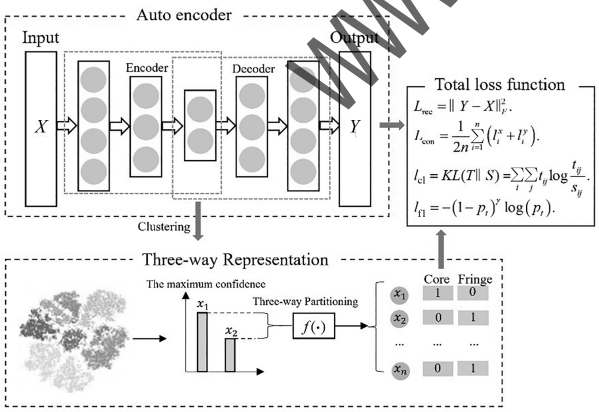


图 1 DE3WC 的框架

Fig. 1 Framework of DE3WC

1 相关工作(Related work)

1.1 深度嵌入聚类

深度嵌入聚类(DEC)方法利用深度自编码器提取数据的低维嵌入特征，随后通过软聚类的方式实现数据的聚类。具体而言，DEC 首先通过预训练自编码器学习数据的初始低维特征表示，随后利用学生 t 分布计算数据点到簇中心的软分配概率，定义如式(1)所示：

$$q_{ij}=\frac{(1+z_i-c_j^2)^{-1}}{\sum_j(1+z_i-c_j^2)^{-1}} \tag{1}$$

进一步，为了增强对高置信度数据点的区分能力，DEC 定义了目标分布以引导聚类：

$$p_{ij}=\frac{q_{ij}^2/\sum_i q_{ij}}{\sum_j q_{ij}^2/\sum_i q_{ij}} \tag{2}$$

然后 DEC 通过最小化目标分布与软分配概率之间的 KL(Kullback-Leibler Divergence)散度来优化聚类效果：

$$KL(P\parallel Q)=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{3}$$

假设 f_W 为编码器的映射函数，x_i 表示来自数据集 X 的第 i 个输入样本。完成预训练阶段后，可通过 f_W 得到所有样本的低维嵌入表示 $Z=\{z_1,z_2,\cdots,z_i,\cdots,z_n\}$，其中 z_i 表示第 i 个样本的嵌入特征。编码器映射函数 f_W 定义如式(4)所示：

$$z_i=f_W(x_i) \tag{4}$$

随后，在嵌入表示集 Z 上应用 K-means 聚类算法，以获取初始聚类中心集合 $C=\{c_1,c_2,\cdots,c_j,\cdots,c_k\}$，其中 c_j 表示第 j 个类别的中心点。每个样本 x_i 最终的聚类标签由 q_{ij} 的最大值决定，即选择其对应概率最大的类别作为预测标签。

在模型训练过程中，通过反向传播可以分别计算损失函数关于样本嵌入表示 z_i 的梯度 $\partial L/\partial z_i$ 和关于聚类中心 c_j 的梯度 $\partial L/\partial c_j$。其中，前者用于更新编码器映射函数 f_W，而后者则用于优化聚类中心 c_j：

$$c_j=c_j-\lambda\frac{\partial L}{\partial c_j} \tag{5}$$

在 DEC 基础上，改进的深度嵌入聚类进一步提出了联合优化重构损失与聚类损失的策略，显著提升了聚类性能与特征表示的质量。该方法通过在聚类任务中加入重构损失，保持了数据原始特征空间的局部结构，极大地增强了模型在高维数据下的聚类能力。

1.2 三支聚类

三支聚类是一种源自三支决策理论的聚类方法，特别适用于处理聚类过程中样本的不确定性和模糊性问题。在三支聚类框架中，数据集 X 中的样本可以依据它对聚类中心的归属程度，被划分为 3 类区域：核心域 $\mathrm{POS}(c_j)$ 由明确归属于聚类中心 c_j 的样本组成；边界域 $\mathrm{BND}(c_j)$ 包含不确定是否属于 c_j 的模糊样本；琐碎域 $\mathrm{NEG}(c_j)$ 由明确不属于聚类中心 c_j 的样本组成。这 3 类区域之间满足一定的判别特性和数学性质，具体如式(6)～式(9)所示：

$$\mathrm{POS}(c_j)\neq\varnothing,j=1,2,\cdots,k \tag{6}$$

$$\mathrm{POS}(c_i)\cap\mathrm{POS}(c_j)=\varnothing,i\neq j \tag{7}$$

$$U_{j=1}^{k}(\mathrm{POS}(c_j)\cup\mathrm{BND}(c_j))=X \tag{8}$$

$$\mathrm{POS}(c_j)\cup\mathrm{BND}(c_j)\cup\mathrm{NEG}(c_j)=X \tag{9}$$

其中：式(6)规定每个簇必须至少包含一个成员；式(7)表明各簇的核心域彼此不相交；式(8)指出数据集中每一个元素必须至少属于某个簇的核心域或边界域，且可能同时属于多个簇；式(9)规定每个簇的核心域、边界域与琐碎域三者联合构成完整的簇域。

1.3 焦点损失

焦点损失[15]最初用于解决分类任务中的类别不平衡问题，它的核心思想是动态调整不同类别样本的损失权重，降低易分类样本的贡献，并增加难分类样本的关注力度，定义公式

www.rjgcz.com

如式(10)所示:

$$FL(p_t)=-(1-p_t)^{\gamma}\log(p_t) \tag{10}$$

其中:p_t 表示真实类别的预测概率;参数 γ 用于调控对困难样本的关注程度。近年来,焦点损失在目标检测与图像分类等领域被广泛采用。Lin 等[16]将焦点损失引入密集型目标检测器 RetinaNet 的训练过程中,有效提升了检测准确率,取得了优异的实验结果。Shu 等[17]在神经网络中引入焦点损失,缓解了类别不平衡问题,并在乳腺癌分类任务中实现了更高的分类性能。

2　深度网络框架(Deep network framework)

设输入数据集为 $X\in\mathbb{R}^{d\times n}$,其中 n 表示样本数,d 表示每个数据点的特征维度。本文方法的目标是在不依赖真实标签的前提下,尽可能准确地将每个样本 $x_i\in X$ 分配到其对应的聚类中。为此,本文通过对比学习获取更具判别性的特征。

深度聚类中的特征学习基本思想是利用自编码器对原始数据进行嵌入表示的提取。典型的自编码器由编码器和解码器两部分组成。编码器通过非线性变换将原始数据 X 投影到一个维度为 d' 的嵌入空间,得到嵌入表示 $H\in\mathbb{R}^{d'\times n}$,即嵌入特征。

由于嵌入特征 $h_i\in H$ 与原始数据点 $x_i\in X$ 一一对应,编码器的映射过程可以表示为

$$h_i=f(x_i|\boldsymbol{W}_e,b_e)=f(\boldsymbol{W}_e x_i+b_e) \tag{11}$$

其中:$f(\cdot)$表示编码器中的激活函数;$\boldsymbol{W}_e\in\mathbb{R}^{d'\times d}$ 和 $b_e\in\mathbb{R}^{d'}$ 分别表示编码器网络的权重矩阵和偏置项。类似地,解码器的目标是将嵌入特征 H 重新映射回原始数据空间,得到重构后的数据 $X'\in\mathbb{R}^{d\times n}$,其过程可表示为

$$x_i'=g(h_i|\boldsymbol{W}_d,b_d)=g(\boldsymbol{W}_d h_i+b_d) \tag{12}$$

其中:g(·)表示解码器中的激活函数;$\boldsymbol{W}_d\in\mathbb{R}^{d\times d'}$ 和 $b_d\in\mathbb{R}^{d}$ 分别为解码器网络的权重矩阵和偏置项,其定义方式与编码器类似。因此,自编码器的优化目标是学习一个映射函数,使得输入数据在经过编码和解码后能够尽可能地还原为原始数据,从而最小化重构误差。该目标函数可表示为

$$\min_{\Omega}\|X-X'\|_F^2 \tag{13}$$

2.1　损失函数

(1)重构损失

采用输入的原始数据与它的重构数据之间的均方误差作为重构损失函数。该重构损失用于衡量自编码器在还原输入数据方面的能力,其定义如式(14)所示:

$$L_{\text{rec}}=\|Y-X\|_F^2 \tag{14}$$

(2)对比损失

为了更有效地学习不同类别之间具有判别性的特征信息,本文引入了对比学习机制。该机制的核心思想是最大化正样本对之间的一致性,同时最小化负样本对之间的相似性。在本文方法中,原始数据与其对应的重构数据被视为正样本对。

设输入的原始数据 $X^{\mathrm{T}}=\{x_1,x_2,\cdots,x_n\}\in\mathbb{R}^{d\times n}$,重构后的输出数据 $Y^{\mathrm{T}}=\{y_1,y_2,\cdots,y_n\}\in\mathbb{R}^{d\times n}$。对于每个样本 x_i,构建一个正样本对(x_i,y_j),即原始样本与其重构结果组成的配对。同时,将除 x_i 与 y_j 以外的其余 $2n-2$ 个样本构成负样本对。这样可以形成一个包含 1 个正对和 $2n-2$ 个负对的对比学习任务。

样本 x_i 与 y_j 之间的相似性定义如式(15)所示:

$$s(x_i,y_j)=\frac{(x_i^{\mathrm{T}})(y_j)}{x_i y_j} \tag{15}$$

其中:$i,j\in[1,n]$。若将 x_i 与 y_j 的顺序对调,即可得到 y_j 与 x_i 之间的相似性,二者是对称的。为了在增强正样本对相似性的同时,降低负样本对之间的相似性,为每个样本 x_i 和其对应的重构样本 y_j 分别定义对比损失函数如式(16)和式(17)所示:

$$l_i^x=-\log\left(\frac{\exp(s(x_i,y_i)/\tau)}{\sum_{j=1}^{n}(2*\exp(s(x_i,x_j)/\tau))}\right. \tag{16}$$

$$l_i^y=-\log\frac{\exp(s(y_i,x_i)/\tau)}{\sum_{j=1}^{n}(2*\exp(s(y_i,y_j)/\tau))} \tag{17}$$

其中:τ 表示温度参数,用于调节对比损失在处理困难负样本时的关注程度。合适的温度系数能够帮助对比学习获得更优的特征表示,它直接影响了对比损失在不同负样本间的权重分配。当温度系数取值过小时,模型的注意力将集中于距离正样本最近的少数几个负样本;而当温度系数过大时,模型会对所有负样本赋予几乎相同的权重,导致判别能力减弱。根据 Cai 等[15]研究,将温度参数设置为 $\tau=0.5$。

最终,对比损失在整个数据集 X 和其重构结果 Y 上进行计算,综合考虑每个样本的正负对,定义如式(18)所示:

$$L_{\text{con}}=\frac{1}{2n}\sum_{i=1}^{n}(l_i^x+l_i^y) \tag{18}$$

(3)聚类损失

聚类的目标是将给定的数据集准确地划分为指定数量的类别。在聚类层中,通过最小化软标签分布 S 与目标分布 T 之间的 KL 散度来定义聚类损失函数,其表达式如式(19)所示:

$$l_{\text{cl}}=KL(T\|S)=\sum_i\sum_j t_{ij}\log\frac{t_{ij}}{s_{ij}} \tag{19}$$

用于计算嵌入层中表示 h_i 与聚类中心 μ_j 之间相似度的软标签分布 S,通过学生分布进行定义,具体公式如式(20)所示:

$$s_{ij}=\frac{(1+h_i-\mu_j{}^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_j(1+h_i-\mu_j{}^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{20}$$

其中:α 参数控制学生 t 分布的自由度。在本文设置中,α 被固定为 1。因此,s_{ij} 表示将样本 i 分配给聚类中心 j 的概率,该概率被称为软标签分配。

为了进一步增强高置信度样本对聚类中心的影响,引入了目标分布 t_{ij},它的定义如式(21)所示:

$$t_{ij}=\frac{s_{ij}^2/\sum_i s_{ij}}{\sum_j s_{ij}^2/\sum_i s_{ij}} \tag{21}$$

其中:t_{ij} 由软分配概率 s_{ij} 计算得到,它的设计目的是使优化过程中软标签分布 S 更加接近目标分布 T。因此,该优化过程本质上可以被视为一种自训练。

如果聚类层仅依赖 KL 散度进行优化,仍然会存在一些难以分类或被误分类的样本。为了解决这一问题,本文在聚类损失中引入焦点损失,以进一步改进标签分配机制。

由于本文的聚类任务通常属于多类别情形,因此需要进行适当修改,以适应多类别聚类的需求。具体而言,重新定义 p_t 如式(22)所示:

$$p_t=y_{\text{pred}}*y_{\text{true}} \tag{22}$$

在多类别情况下,y_{pred} 和 y_{true} 分别表示预测标签和真实标签。通常在这种情形下需要真实标签,但这将转变为一个监督学习问题,这与本文的聚类任务是不一致的。

本文的聚类模块中生成的软标签分布 S 和目标分布 T 为此问题提供了直接的解决方案。具体而言,预测标签由 S 生成,真实标签由 T 生成,因此可以在聚类层的自训练过程中直接引入焦点损失,它的优化目标可以形式化为

www.rjgcz.com

$$l_{\mathrm{fl}}=-(1-p_t)^{\gamma}\log(p_t) \tag{23}$$

根据焦点损失 l_{fl} 的定义可知：对于一个易分类样本，它的对应的概率$(1-p_t)$较小，p_t 接近于 1；而对于一个难分类样本，$(1-p_t)$较大，p_t 接近 0。这意味着焦点损失能够在损失函数中增强难分类样本的贡献，同时减小易分类样本的影响。

因此，本文聚类模块的最终目标函数 L_{c} 可定义为式(24)形式：

$$L_{\mathrm{c}}=l_{\mathrm{cl}}+l_{\mathrm{fl}} \tag{24}$$

其中：l_{cl} 使网络能够以自训练的方式实现聚类；而 l_{fl} 则用于解决仅使用 l_{cl} 时难以挖掘部分难分类样本的问题。

(4)整体损失函数

结合上述损失函数的讨论，提出了一种新的融合损失函数，它的形式定义如式(25)所示：

$$L=L_{\mathrm{rec}}+\beta L_{\mathrm{con}}+\gamma L_{\mathrm{c}} \tag{25}$$

2.2 三支划分

在应用于高维数据集的深度聚类中，传统的聚类方法倾向于采用一种严格的方法，明确地将每个样本分配给单个聚类。这种方式默认每个样本都完全隶属于某一聚类中心，忽略了样本可能存在的模糊性或不确定性。然而，在实际的数据中，样本常常具有复杂性，其中不乏一些难以明确分类的样本。这类样本可能处于不同簇之间的模糊区域，或位于数据分布的边缘。

如果忽视这类样本的不确定性，可能会导致聚类结果的不准确，甚至遗漏潜在的有价值信息。为解决这一问题，三支聚类提供了一种具有优势的解决方案。该方法将样本划分为核心域、边界域和琐碎域，从而为聚类过程引入了灵活性。

通过这种方式，模型不仅能够有效识别具有明确聚类归属的样本，还能够合理处理存在模糊性或不确定性的样本，以及不属于当前聚类的样本。这种更加细致的划分机制能够提升聚类的鲁棒性和表达力，更加贴近现实数据的复杂结构。

设有一个高维数据集 X，利用前述的深度网络对它进行降维处理，获得特征表示集合 $Z=\{z_1,z_2,\cdots,z_i,\cdots,z_n\}$。随后，利用学生 t 分布计算每个样本对簇中心的软标签分布，再计算每个样本的目标分布，定义目标分布矩阵 $\boldsymbol{P}=\{p_{ij}\}$。其中每个样本对应的最大置信度记为

$$p_{\max}(x_i)=\max_j p_{ij} \tag{26}$$

在此基础上，引入三支聚类的思想，计算所有样本最大置信度的均值，如式(27)所示：

$$\bar{p}=\frac{1}{n}\sum_{i=1}^{n}p_{\max}(x_i) \tag{27}$$

基于目标分布矩阵，利用阈值参数 ρ 对数据样本进行核心域、边界域和琐碎域的划分。需要判断每个样本是否可以被明确划分到某个聚类中，或者它的归属仍存在模糊性。该判断过程可通过式(28)实现：

$$\mathrm{mask}(x_i)=\begin{cases}1,\max\{p_{i1},p_{i2},\cdots,p_{ij}\}\geqslant\rho\cdot\bar{p}\\0,\max\{p_{i1},p_{i2},\cdots,p_{ij}\}<\rho\cdot\bar{p}\end{cases} \tag{28}$$

其中：$\bar{p}$ 为所有样本最大置信度的均值；ρ 是划分核心域与边界域的调节参数。当样本对所有聚类中心的最大置信度超过阈值 $\rho\cdot\bar{p}$ 时，该样本被标记为确定样本；否则被视为不确定样本。$\rho\cdot\bar{p}$ 值越大，确定样本的数量越少，不确定样本的数量越多。较高的 p_{ij} 值意味着样本 x_i 更有可能属于对应的聚类中心 c_j。值为 1 的样本被归类为特定样本，而值为 0 的样本被视为不确定样本。当某个样本对于某个类的 mask 值为 1 时被归类为确定样本放入该类核心域中，并放入其他类的琐碎域中。若样本对于所有类的值都为 0 则放入边界域中。对于边界域中的样本，根据 p_{ij} 大小，归入对样本 p_{ij} 最大的类簇中。本文方法的全面概述如算法 1 所示。

算法 1 DE3WC

输入：训练数据集 X，簇数 k，学习率 lr，核心划分参数 ρ，聚类损失权重 γ，对比损失权重 β；

输出：训练后的聚类模型参数。

```
初始化自编码器 AE，构建 IDEC 模型
使用重建损失预训练 AE
使用 K-means 在 AE 编码器输出的隐空间特征上进行初始聚类，得到初始簇中心
将初始簇中心赋值给 IDEC 模型的聚类层参数
for epoch = 1 to max_epoch do
    if epoch mod update_interval == 0 then
        全量推理所有样本，得到软分配概率 q
        根据 q 计算目标分布 P
        计算每个样本在 P 中最大概率 P_max
        计算所有 P_max 的均值 P_mean
        根据阈值 ρ×P_mean 将样本划分为核心域样本与边界域样本
        基于 q 重新预测聚类标签，计算 ACC，NMI，ARI
        计算聚类标签变化率 Δ_label
        if Δ_label < tol then
            终止训练
        end if
    end if
    for each batch (x, id x) ∈ DataLoader(X) do
        使用 AE 得到重建样本 x̄ 和软分配 q
        计算总损失
        反向传播并更新模型参数 D
    end for D
end for
```

3 实验与结果分析(Experiment and result analysis)

3.1 实验设置

本实验选用 5 个具有代表性的数据集，涵盖手写数字、人脸图像与物体图像，以全面评估本文方法在不同场景下的聚类性能。

USPS 数据集广泛用于手写数字识别任务，包含约 1854 张 0～9 数字的灰度图像。每张图像尺寸为 16 像素×16 像素，原始数据来源于美国邮政局。

ORL 人脸数据集由英国剑桥的 Olivetti 研究实验室于 1992—1994 年间采集，包含 400 张灰度人脸图像，分别来自 40 位不同的个体。每位受试者提供 10 张图像，涵盖不同的表情和光照条件。

COIL-20 数据集用于物体识别和三维物体重建研究，包含 1440 张灰度图像，涵盖 20 个不同的物体。每个物体在 5 度间隔下从多个视角拍摄，以模拟姿态变化。

CIFAR-10 是一个广泛应用于自然场景下物体识别的标准数据集，包含 60 000 张 32×32 分辨率的彩色图像，均匀分布在 10 个类别中，包括动物和交通工具等。

COIL-100 数据集是 COIL-20 的扩展版本，规模更大，包含 7200 张彩色图像，涵盖 100 种不同的物体。每个物体从 72 个不同角度拍摄，以提供更丰富的姿态变化特性。

这些数据集被广泛应用于评估机器学习算法在应对不同视觉变化时的鲁棒性和泛化能力。

www.rjgczz.com

数据汇总描述见表1，数据集图像如图2所示。

表1　数据集信息

Tab.1 Dataset information

数据集	样本数	维度	类别数
USPS	2000	256	10
ORL	400	1024	40
COIL-20	1440	1024	20
CIFAR-10	60 000	300	10
COIL-100	7200	1024	100

图2　数据集图像

Fig.2 Dataset images

3.2　比较基线

为了验证本文方法的有效性，选取了6种具有代表性的先进聚类方法作为对比基线，具体如下：

(1)K-means是一种广泛应用的硬聚类方法，它将数据集划分为k个簇。该迭代算法通过不断优化簇中心的位置，最小化簇内方差并最大化簇间距离，从而实现相似样本的聚合与簇间的清晰分离。

(2)FKM为每个样本分配概率型的隶属度值。通过迭代更新隶属度和簇中心位置，该方法最大化簇内相似性，相比传统K-means具有更高的聚类适应性。

(3)DEC方法结合深度神经网络，能够同时进行特征表示学习与聚类任务。它通过编码器将数据从高维空间映射到低维特征空间，并基于聚类目标函数进行迭代优化。

(4)DEKM通过标准正交变换矩阵对嵌入空间进行变换，以揭示数据的聚类结构信息。该变换矩阵由k类内散度矩阵的特征向量推导而成，从而有效丰富了聚类结构的表达能力。

(5)DFKM提出了一种新的深度模糊K-means模型，结合了自适应损失函数和熵正则化机制。该方法能够同时进行深度特征提取与模糊聚类，生成更适合聚类任务的非线性特征表示。

(6)D3WC方法引入了基于三支决策理论的深度聚类框架，通过划分正域、边界域与负域来细化聚类决策。该方法在深度嵌入空间中联合考虑了样本的类别隶属关系和决策不确定性，从而提高了聚类的准确性与鲁棒性。

这些对比方法覆盖了传统聚类、模糊聚类、深度聚类与基于决策理论的聚类策略，有助于从不同角度全面评估本文方法的有效性与优势。

3.3　评价指标

为了全面评估聚类模型的性能，本文采用了两种常用的外部评估指标，即聚类准确率和归一化信息。

聚类准确率(ACCuracy，ACC)：ACC[18]用于衡量聚类结果与真实标签的一致性。它通过计算聚类结果与真实标签完全对应的数据点占总数据量的比例来表示聚类效果。通常而言，ACC值越高，说明聚类性能越好。ACC的计算公式如式(29)所示：

$$ACC=\frac{1}{N}\sum_{i=1}^{N}1\{y_i=\pi(c_i)\} \tag{29}$$

其中：N表示数据集中样本的总数；y_i表示第i个样本的真实标签；c_i表示聚类算法为第i个样本分配的聚类标签；π表示标签到真实标签的最优映射函数；$1\{\cdot\}$为指示函数，当括号内条件成立时取值为1，否则为0。

归一化互信息(Normalized Mutual Information，NMI)：NMI[19]基于信息论概念，通过计算聚类结果与真实类别之间的互信息来评估聚类性能。NMI值越高，表明聚类结果与真实标签之间的一致性越强。它的计算公式如式(30)所示：

$$NMI(U,V)=\frac{2\times I(U;V)}{H(U)+H(V)} \tag{30}$$

其中：$I(U;V)$表示聚类结果U与真实标签V之间的互信息，$H(U)$和$H(V)$分别表示U和V的熵值。

纯度(Purity)：Purity用于评估聚类中数据点属于单一类别的程度。通过计算每个聚类中占主导地位的类别相对于聚类大小的比例来衡量聚类质量。较高的Purity值表示聚类中的大多数数据点属于相同的真实类别，反映了更好的聚类一致性。计算公式如式(31)所示：

$$\text{Purity}=\frac{1}{N}\sum_{k}\max|w_k\cap c_j| \tag{31}$$

其中：w_k表示第k个聚类；c_j表示第j个真实类别；$\max|w_k\cap c_j|$表示属于聚类w_k中最常见真实类别c_j的数据点数量。

上述评估指标能够从不同角度定量反映聚类模型的性能，广泛应用于聚类算法的客观性能比较中。

3.4　实验结果

将本文提出的DE3WC方法与几种流行的聚类方法在5个公开的数据集上进行了比较，结果见表2，从中可以得出以下结论：

表2　聚类方法对比

Tab.2 Comparison of clustering methed　　%

方法	评价指标	USPS	ORL	COIL-20	CIFAR-10	COIL-100
K-means	*ACC*	64.67±0.85	59.30±3.03	65.78±2.57	20.61±0.08	49.23±4.57
	NMI	62.67±0.16	77.67±1.08	78.09±0.02	7.93±0.08	76.67±2.89
	Purity	72.12±0.20	63.85±2.12	69.29±0.97	22.11±0.07	54.51±3.27
FKM	*ACC*	61.47±0.04	56.35±1.71	58.04±1.20	17.51±0.80	48.08±2.00
	NMI	57.37±0.28	76.97±1.12	75.89±0.01	4.27±0.68	76.93±2.11
	Purity	66.61±0.44	60.43±2.44	62.30±1.20	17.78±0.80	53.00±2.29
DEC	*ACC*	68.81±3.21	39.15±8.58	47.50±2.86	16.12±0.91	32.08±1.45
	NMI	64.38±2.57	63.84±7.63	68.09±2.33	5.93±0.60	67.57±1.49
	Purity	75.60±3.13	42.78±9.08	51.32±2.86	17.45±0.40	34.64±1.24
DEKM	*ACC*	75.06±0.32	43.68±5.48	65.85±1.43	/	50.90±1.07
	NMI	75.30±0.25	68.27±6.24	78.01±0.45		72.50±1.24
	Purity	81.33±0.34	48.10±3.56	68.06±1.56		54.81±1.32
DFKM	*ACC*	71.27±2.03	62.65±1.64	64.70±0.84	22.03±0.61	51.33±1.73
	NMI	70.72±0.55	81.17±0.85	78.47±0.53	8.63±0.40	77.82±1.61
	Purity	77.41±0.55	66.85±1.08	68.11±0.86	23.79±0.60	56.28±1.68
D3WC	*ACC*	77.00±1.75	65.03±2.33	**67.78±0.36**	/	**51.88±1.97**
	NMI	72.89±1.39	82.04±1.58	78.56±0.85		**78.71±1.22**
	Purity	81.40±1.80	68.60±2.24	70.63±0.48		56.61±1.01
DE3WC	*ACC*	**79.80±1.38**	**74.25±2.57**	66.60±0.91	**23.48±0.54**	51.78±1.25
	NMI	**82.34±1.21**	**88.39±1.93**	**80.13±0.66**	**10.72±0.71**	78.39±1.16
	Purity	**86.96±1.44**	**75.00±2.45**	**71.67±0.84**	**24.52±0.61**	**56.81±1.23**

DE3WC模型在USPS、ORL、COIL-20、CIFAR-10以及COIL-100数据集上均展现出优异的聚类效果。在USPS数据集上，DE3WC实现了79.80%的聚类准确率和82.34%的归一

www.rjgcz.com

化互信息和86.96%的纯度，相较于传统 *K*-means 方法，分别提升了15.13、19.67 和 14.84 个百分点。在其他数据集上，如ORL 和 COIL-20，DE3WC 同样在 *ACC* 与 *NMI* 指标上超越了DEC、DEKM、DFKM、D3WC 等现有深度聚类方法，体现了良好的普适性与鲁棒性。即使在较具挑战性的 CIFAR-10 数据集上，DE3WC 依然取得了优于传统方法的聚类性能。由此看出 D3WC 能够动态筛选可靠样本，聚焦于难以分类的边界样本，从而进一步提高了聚类性能与泛化能力。在表 2 中用粗体标注的结果代表了最佳的聚类性能。“/”表示这些方法的源代码和分数在相应的数据库中不可用。

3.5 实验可视化

为了进一步验证本文方法的聚类效果，采用 *t*-SNE 对USPS 数据集的聚类结果进行了二维可视化，结果如下：

USPS 原始数据的 *t*-SNE 降维可视化结果如图 3 所示。可以观察到，尽管部分类别的样本在低维空间中呈现出一定的分离趋势，但在图像中央区域仍存在明显的类别混杂现象，尤其是不同类别之间的边界不够清晰，聚类结构较为松散。

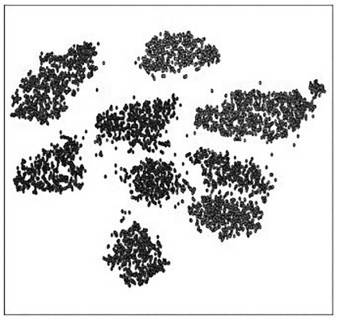


图 3 USPS 原始数据的 *t*-SNE 降维可视化结果

Fig. 3 *t*-SNE dimensionality reduction visualization results of raw USPS data

经过 D3WC 方法训练后的 *t*-SNE 可视化结果如图 4 所示。与原始数据相比，经过深度聚类后的样本在二维空间中表现出更为明显的类别分布特征，大多数类别形成了相对集中的簇状结构，类别间的边界也更加清晰。然而，仍可见部分样本存在分散现象，个别类别之间仍存在一定程度的重叠。

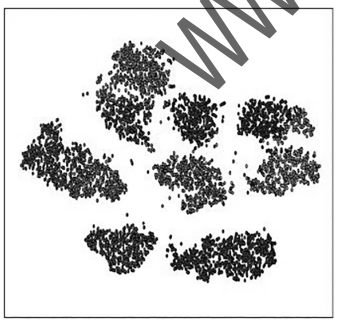


图 4 D3WC 方法训练后的可视化结果

Fig. 4 Visualization results of the D3WC method after training

DE3WC 方法训练后的 *t*-SNE 可视化结果如图 5 所示。可以明显看出，相较于 D3WC 方法，DE3WC 方法在特征空间中形成了更加紧凑且分界清晰的簇结构。各类别样本在低维空间中的分布更加集中，簇内样本密集，簇间距离增大，类别边界更加锐利，极大地减少了类别混杂现象。这表明 DE3WC 在特征学习和聚类任务中具有更优的判别能力和特征表达能力。

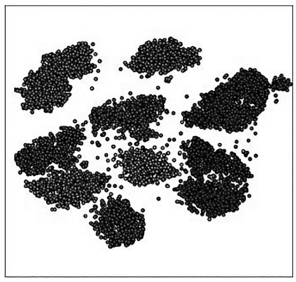


图 5 DE3WC 方法训练后的可视化结果

Fig. 5 Visualization results of the DE3WC method after training

总体而言，通过 *t*-SNE 可视化分析可以直观验证，DE3WC 能够在特征空间中有效区分不同类别的样本，提升聚类结果的可分性和纯净度，进一步佐证了本文方法在深度聚类任务中的有效性与优越性。

3.6 消融实验

为了验证本文方法中引入的对比损失、焦点损失和三支划分的有效性，在 3 个数据集（USPS、ORL 和 COIL-20）上进行了消融实验。具体来说，构建了 4 种降级模型，分别是：使用对比损失、使用焦点损失、使用三支划分以及不使用任何损失项。实验结果见表 3。

表3 消融实验

Tab. 3 Ablation experiments %

对比损失	焦点损失	三支划分	USPS			ORL			COIL-20		
			ACC	*NMI*	Purity	*ACC*	*NMI*	Purity	*ACC*	*NMI*	Purity
×	×	×	68.18	64.38	75.60	39.15	63.84	42.78	47.50	68.09	51.32
×	×	√	71.59	68.20	78.82	50.25	71.90	52.92	53.94	73.70	55.76
√	×	√	73.60	71.44	80.94	50.50	74.56	53.00	51.81	71.45	53.75
×	√	√	73.04	72.08	80.73	64.25	83.66	66.50	51.04	72.55	54.65
√	√	√	**79.80**	**82.34**	**86.96**	**74.25**	**88.39**	**75.00**	**66.60**	**80.13**	**71.67**

可以看到，添加对比损失或焦点损失在聚类任务上都带来了明显的提升，尤其在 ORL 和 COIL-20 数据集上，性能提升尤为显著。在 USPS 数据集上，添加对比惩罚项的效果虽然较为温和，但依然有小幅度的提升。应用焦点损失在 ORL 和COIL-20数据集上的聚类性能显著优于仅使用对比损失，而在USPS 数据集上，焦点损失的效果不如对比损失。这表明在不同的数据集上，这两种损失项各自具有不同的优势。三支划分则在所有数据集上都显著改善了聚类效果。当同时采用对比损失、焦点损失和三支划分时，聚类性能进一步提升，表明这三者相辅相成。

在统计显著性方面，p 值为 0.005 37，表明这些改进在统计上是显著的。这进一步验证了本文方法中对比损失、焦点损失和三支划分的有效性，验证它们在聚类任务中相互增强，共同推动了性能的提升。

3.7 参数灵敏度

本节介绍了 DE3WC 对参数的敏感度，分别针对对比损失权重 β、聚类损失权重 γ 以及三支聚类划分参数 ρ 进行了灵敏度分析，实验结果如图 6 和图 7 所示。

图 6 展示了在 USPS 数据集上，不同 β 与 γ 的组合下，模型在聚类准确率和归一化互信息指标上的变化情况。由 3D 柱状图可以观察到，随着对比损失权重 β 和聚类损失权重 γ 的调整，模型的聚类性能存在一定波动，但整体变化趋势较为平稳。其中，当 γ 取 0.05 至 0.1、β 取 0.3～0.5 的组合时，模型在 *ACC* 和 *NMI* 指标上均达到了较高水平，表明合理平衡聚类目标与特征保持的训练损失，有助于提升模型的聚类性能。γ 值过小时，聚类目标不足，容易导致聚类性能下降；γ 值过大时，则可

www.rjgczz.com

能牺牲重构质量，导致特征表达能力受损。

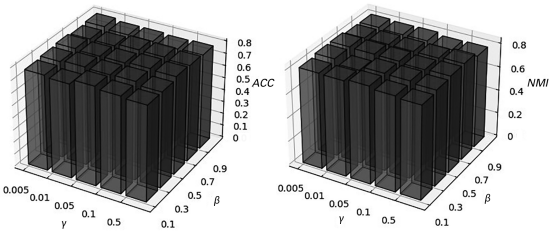
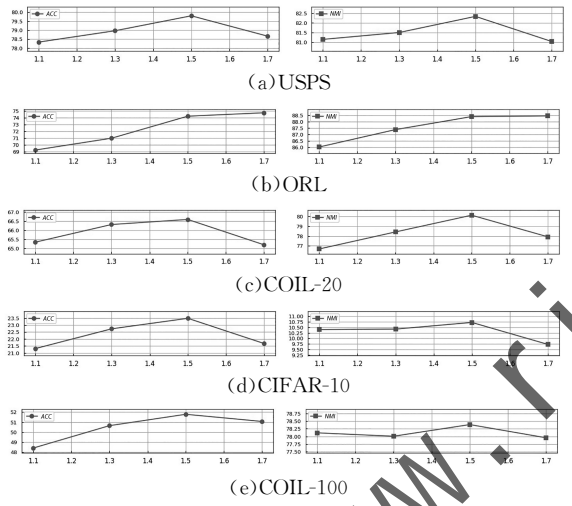


图 6　DE3WC 不同的 γ 和 β 对 USPS 数据集的影响

Fig. 6 Impact of different γ and β of DE3WC on USPS datase

USPS、ORL、COIL-20、CIFAR-10、COIL-100 上，三支聚类参数 ρ 对 *ACC* 和 *NMI* 的影响趋势分析如图 7 所示。整体来看，ρ 的选取对模型聚类性能具有明显影响。当 ρ 值较低时，核心样本筛选过于宽松，导致边界样本混入训练过程，从而影响聚类效果；而当 ρ 值适中时，能够较好地区分区域样本，提升聚类准确率和一致性。进一步观察可知，USPS、ORL、COIL-20 等数据集在 ρ 取 1.5 时，*ACC* 和 *NMI* 指标均达到峰值，说明合理的核心-边界域划分策略对于提升模型的判别能力至关重要。而在 CIFAR-10 和 COIL-100 等复杂数据集上，尽管整体提升幅度相对较小，但仍呈现出相似的变化趋势，进一步验证了三支聚类机制在不同数据集上的适用性。



(a)USPS

(b)ORL

(c)COIL-20

(d)CIFAR-10

(e)COIL-100

图 7　不同 ρ 对五个数据集的影响

Fig. 7 Effects of different ρ values on the five datasets

综合上述实验结果，本文最终确定 $\gamma=0.1$、$\beta=0.3$、$\rho=1.5$ 为最佳超参数组合。在该参数配置下，DE3WC 方法在各个数据集上均能取得优异的聚类性能，实现了准确率与归一化互信息的同步提升。

4　总结(Conclusion)

本文提出了融合深度嵌入特征学习、对比学习、焦点损失和三支决策理论的深度聚类方法 DE3WC，以解决高维数据聚类中的特征判别性不足与样本模糊性问题。

方法上，基于 IDEC 框架，结合对比损失强化特征区分能力，引入焦点损失提升对难分类样本的学习效果，并通过三支划分机制动态区分核心、边界与琐碎样本，实现差异化优化。在 USPS、ORL、COIL-20、CIFAR-10 与 COIL-1005 个数据集上，DE3WC 在聚类准确率与归一化互信息指标上均优于传统及现有深度聚类方法，验证了它的有效性与鲁棒性。总体而言，DE3WC 为复杂高维数据的聚类任务提供了更准确、稳健的解决方案，具有较高的理论意义和应用潜力。

参考文献(References)

[1] Hartigan J A, Wong M A. Algorithm AS 136: a *k*-means clustering algorithm[J]. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1979, 28(1): 100-108.

[2] Dasgupta S, Long P M. Performance guarantees for hierarchical clustering[J]. Journal of Computer and System Sciences, 2005, 70(4): 555-569.

[3] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17: 395-416.

[4] Zhang D Q, Chen S C. A novel kernelized fuzzy c-means algorithm with application in medical image segmentation[J]. Artificial Intelligence in Medicine, 2004, 32(1): 37-50.

[5] Rao L, Jia N X, Hu J, et al. ATPdock: a template-based method for ATP-specific protein-ligand docking[J]. Bioinformatics, 2022, 38(2): 556-558.

[6] Yao Y. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3): 341-353.

[7] Yu H, Wang X, Wang G, et al. An active three-way clustering method via low-rank matrices for multi-view data[J]. Information Sciences, 2020, 507: 823-839.

[8] Wang P, Yao Y. CE3: a three-way clustering method based on mathematical morphology[J]. Knowledge-based Systems, 2018, 155: 54-65.

[9] Yao Y. The geometry of three-way decision[J]. Applied Intelligence, 2021, 51(9): 6298-6325.

[10] Wen H, Xia F, Tang H. Three-way decision for incomplete real-valued data[J]. Journal of Intelligent & Fuzzy Systems, 2020, 39(5): 7843-7862.

[11] Hao C, Li J, Fan M, et al. Optimal scale selection in dynamic multi-scale decision tables based on sequential three-way decisions[J]. Information Sciences, 2017, 415: 213-232.

[12] Ju H, Guo J, Ding W, et al. D3WC: deep three-way clustering with granular evidence fusion[J]. Information Fusion, 2025, 114: 102699.

[13] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis [C/OL]// ICML, 2016: 478-487. https://proceedings.mlr.press/v48/xie16.html.

[14] Guo X, Gao L, Liu X, et al. Improved deep embedded clustering with local structure preservation [C/OL]// IJCAI, 2017: 1753-1759. https: // doi. org/10. 24963/ijcai. 2017/243.

[15] Cai J, Wang S, Xu C, et al. Unsupervised deep clustering via contractive feature representation and focal loss[J]. Pattern Recognition, 2022, 123: 108386.

[16] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C/OL]// ICCV, 2017: 2980-2988. https: // ieeexplore.ieee.org/document/8237586.

[17] Shu W, Wang S, Chen Q, et al. Pathological image classification of breast cancer based on residual network and focal loss [C/OL]// CSAI, 2019: 211-214. https: // doi. org/10.1145/3374587.3374628.

[18] Wu M, Schölkopf B. A local learning approach for clustering [C/OL]// NIPS, 2006: 1529-1536. https: // proceedings. neurips. cc/paper/2006/hash/097e26b2ffb0339458b55da-17425a71f-Abstract.html.

[19] McDaid A F, Greene D, Hurley N. Normalized mutual information to evaluate overlapping community finding algorithms [PP/OL]. V1. arXiv (2011-10-12) [2025-05-03]. https: // arxiv.org/abs/1110.2515.

作者简介：

吴昊哲(2001-)，男(汉族)，南京，硕士生。研究领域：深度学习、三支决策。

王平心(1980-)，男(汉族)，南京，教授，博士。研究领域：粗糙集、粒计算。本文通信作者。

www.rjgcz.com