doi:10.3969/j.issn.1001-4616.2025.04.011

基于证据信息粒化的深度三支 FCM 聚类方法

郭　静[1],蔡超越[1],陆　杨[1],成晓天[1],樊晓雪[1],鞠恒荣[1,2],丁卫平[1]

(1.南通大学人工智能与计算机学院,江苏 南通 226019)
(2.南京大学计算机软件新技术国家重点实验室,江苏 南京 210000)

[**摘要**] 深度聚类由于其在数据挖掘和计算机视觉领域中处理高维数据的显著效果,已经成为一种流行的无监督学习方法. 高维空间中的数据更容易存在模糊性,然而深度聚类无法直接处理数据中的模糊性. 在许多实际问题中,数据之间相似性和关联性通常更集中的表现在局部邻域内,但是传统的深度聚类方法忽略了数据之间的局部关系. 为了解决上述问题,本文提出了一种基于证据信息粒化的深度三支 FCM 聚类方法. 首先,本文提出一种新的对比深度 FCM 聚类网络框架,将数据从复杂的原始数据空间映射到合适的深度特征空间中. 其次,基于三支决策的思想,将第一阶段的聚类结果划分为正域和边界域,以便处理数据中的不确定性. 最后,引入半球邻域粒化方法,为不确定样本构造信息粒. 基于此,本文利用证据理论对信息粒中的信任度进行融合,从而实现对不确定数据的再分配. 本文所提方法更多地关注数据的局部结构,以准确地捕捉数据的内在特征. 实验结果表明,本文所提出的方法有效地提升了聚类效果.

[**关键词**] 证据理论,三支决策,信息粒化,FCM 聚类,深度聚类,对比学习
[**中图分类号**]TH212　[**文献标志码**]A　[**文章编号**]1001-4616(2025)04-0106-12

Deep Three-Way FCM Clustering Method Based on Evidential Information Granulation

Guo Jing[1],Cai Chaoyue[1],Lu Yang[1],Cheng Xiaotian[1],Fan Xiaoxue[1],Ju Hengrong[1,2],Ding Weiping[1]

(1.School of Artificial Intelligence and Computer Science,Nantong University,Nantong 226019,China)
(2.State Key Lab. for Novel Software Technology,Nanjing University,Nanjing 210000,China)

Abstract:Deep clustering has become a popular unsupervised learning method due to its remarkable effect in processing high-dimensional data in the field of data mining and computer vision. Data in high-dimensional space are more likely to have ambiguity, but deep clustering cannot directly handle the ambiguity in data. In many practical problems, the similarities and correlations between data are usually more concentrated in local neighborhoods, but traditional deep clustering methods ignore the local relationships between data. In order to solve the above problems,this paper proposes a deep three-way FCM clustering method based on evidential information granulation. First,this paper proposes a contrastive deep FCM clustering network framework to map data from complex original data space to a suitable deep feature space. Secondly,based on the idea of three-way decision,the clustering results of the first stage are divided into the positive regions and boundary regions in order to deal with the uncertainty in the data. Finally, the semiball neighborhood granulation method is introduced to construct information granules for uncertain samples. Based on this,this paper uses evidential theory to integrate trust in information particles to achieve the redistribution of uncertain data. The method proposed in this paper pays more attention to the local structure of the data to accurately capture the intrinsic characteristics of the data. Experimental results show that the method proposed in this paper effectively improves the clustering effect.

Key words:evidential theory,three-way decision,information granulation,FCM clustering,deep clustering,contrastive learning

收稿日期:2024-09-24.
基金项目:国家自然科学基金资助项目(62006128、U2433216、62066049)、国家重点研发计划资助项目(2024YFE0202700)、江苏高校'青蓝工程'资助、南京大学计算机软件新技术国家重点实验室资助项目(KFKT2024B30)、江苏省自然科学基金资助项目(BK20231337)、江苏省高等学校自然科学研究重大资助项目(21KJA510004)、江苏省研究生科研与实践创新计划资助项目(SJCX24_2016、SJCX25_2001)、南通市自然科学基金青年基金资助项目(JC2024044).
通讯作者:鞠恒荣,博士,副教授,主要研究方向:粒计算,粗糙集,机器学习和知识发现. E-mail:juhengrong@ ntu.edu.cn

聚类是无监督学习中的一种基本技术[1-2],在许多领域得到了广泛的应用,包括机器学习[3-4]、模式识别[5]、图像分割[6]、推荐系统[7]和其他社会科学[8]等. 聚类的目标是将数据集中的样本分成不同的组或簇,使得每个簇内的样本彼此相似,而不同簇之间的样本尽可能不同. 传统聚类方法如 K-means 算法[9]通常采用欧氏距离作为相似性度量,根据每个样本与聚类中心的距离来确定该样本所属的簇. 尽管这些传统方法取得了一定的成果,但是在处理高维和复杂数据集时表现不佳,难以准确捕捉数据的内在结构,此外,高维数据往往具有大量的冗余特征和噪声,这些因素可能会干扰传统聚类方法的性能,使聚类结果不稳定或不准确.

另一方面,K-means 等聚类算法属于硬聚类范畴,每个样本以 0,1 二进制形式被分配到唯一的簇中,无法很好表达样本间的不确定性. 软聚类通过分配隶属度来表示不确定性,即每个样本被赋予属于每个簇的概率. 模糊 C 均值(fuzzy C-means,FCM)聚类算法是典型的软聚类方法之一,已经引起了相当大的关注. FCM 为每个样本分配属于每个簇的隶属度,它允许样本之间存在一定程度的模糊性. 然而,FCM 算法有几个主要的缺点:首先,FCM 对初始聚类中心的选择非常敏感,不同的初始值可能导致不同的聚类结果,这使得结果不够稳定可靠. 接着,FCM 容易陷入局部最优解,导致聚类结果不理想. 在面对大规模数据时,这些缺点更加显著. 为了弥补 FCM 算法的这一缺陷,Verma 等[10]结合 FCM 算法和粒子群优化算法(particle swarm optimization,PSO)的优点提出了一种利用粒子群优化算法优化的模糊 C 均值算法(fuzzy c-means particle swarm optimization,FCM-PSO). 其次,FCM 对数据中的噪声和异常值非常敏感,即偏离正常样本过多的样本会大大影响 FCM 的聚类过程,最终导致错误的聚类结果. 针对该问题可采用数据加权[11]、空间信息增强[12-13]等方式加以解决. 最后,FCM 无法直接处理数据中的不确定性,即对于那些不明确属于某一类别的样本,FCM 无法提供明确的聚类结果. 在高维数据中,这个缺点可能会更加显著,因为高维空间中的样本更容易存在模糊性,即更多的样本可能处于不明确的类别之间.

传统的聚类算法将聚类定义为具有明确边界的集合,只考虑属于和不属于两种关系. 然而,这种非此即彼的描述无法捕捉数据的不确定性. 为了解决这一问题,Yao 提出了三支决策(three-way decision,3WD)的概念[14-16]. 与传统的二支决策相比,三支决策在属于和不属于的基础上提出了可能属于的概念,将全域划分为正域、负域以及边界域. 这一概念在随后的研究中逐渐受到关注,吸引了越来越多的研究者进行深入研究. 例如,徐天杰等[17]结合人工蜂群算法,提出了一种基于人工蜂群的三支 K-means 聚类算法. 此外,Wang 等[18]指出传统的 K-means 聚类方法可以将数据精确地分配到边界清晰的单个簇中,但其缺点在于未能充分考虑到簇的边界可能并不明确,对此他们提出了一种基于三支决策的三支 K-means 聚类算法,以更好地解决这一问题.

随着深度学习技术的发展,深度聚类逐渐受到关注,现有深度聚类方法针对图像和文本等高维数据利用深度学习网络(如卷积神经网络[19]、自编码器[20]等)自动学习数据更抽象和更具表达性的特征. 深度自编码器(deep autoencoder,DAE)[21]作为深度聚类方法提取特征的重要工具,通过最小化重构损失来学习低维嵌入特征空间. 基于 DAE 的深度聚类方法主要分为两种,一种方法将数据嵌入到低维特征中,然后使用传统的聚类方法对嵌入特征进行聚类[22-23]. 另一种方法是联合优化特征表示和聚类[24-25]. 受 t-SNE 算法思想的启发,Xie 等[26]提出了深度嵌入式聚类(deep embedded clustering,DEC)的联合框架来优化特征学习和聚类,DEC 首先通过重构损失学习从数据空间到低维特征空间的映射,然后通过 KL 散度迭代优化聚类损失. 基于 DEC,研究人员提出多种变体. 例如,改进的深度嵌入聚类(improved deep embedded clustering,IDEC)通过联合优化加权聚类损失和自编码器的重构损失来保证微调阶段的局部结构. 但 IDEC 仍属于硬聚类范畴,深度模糊 K-means 聚类(deep fuzzy K-means,DFKM)[27]则使用了软聚类策略,该方法通过为每个数据样本分配对于聚类中心的隶属度值,表示其对多个簇的归属可能性,从而更好地捕捉数据的模糊边界和复杂结构.

当前,对比学习在无监督学习领域展示了卓越的性能. 其通过在特征空间中最大化正对的相似度并最小化负对的相似度,提高了模型的特征表示能力. 基于这一概念,研究者们提出了多种方法. 例如,SimCLR[28]开发了一种非线性头部架构. 该框架利用数据增强手段,如裁剪、旋转和高斯噪声等,来设计正对和负对,从而通过对比学习来学习视觉表示. 对比聚类(contrastive clustering,CC)[29]通过数据增强构造正实例对和负实例对,然后将其投影到特征空间中进行处理. 该方法在实例级和聚类级上分别对行空间

和列空间进行对比学习来优化模型性能. 许洁等[30]为了减小聚类嵌入被分离所带来的影响,提出了一种结构关系一致的对比聚类方法. 然而,这些方法依赖于合理的数据增强来构造正负样本对. 由于每个数据集都具有其独特的特性,不同的数据集需要采用不同的数据增强方法,这大大限制了这些模型的通用性. 此外,数据增强的引入不仅增加了模型的复杂性,还可能导致计算成本大幅提升. 为了解决这一问题,本文将对比损失集成到自编码器中,以增强聚类性能.

本文针对深度聚类无法直接处理数据中的不确定性以及传统的深度聚类方法忽略了样本之间的局部关系的问题,提出了基于证据信息粒化的深度三支 FCM 聚类方法. 首先提出了一种新的对比深度 FCM 聚类网络框架,将数据从复杂的原始数据空间映射到一个更加合适的深度特征空间中. 其次基于三支决策的思想,将深度 FCM 聚类结果划分为正域和边界域,有助于处理数据中的不确定性. 最后使用半球邻域粒化方法,为不确定样本构造信息粒,并利用证据理论对信息粒的信任度进行融合,以聚焦样本的局部结构,可以准确地捕捉数据的内在特征. 本文的主要贡献总结如下:

(1)提出了一种新颖的对比深度 FCM 聚类网络框架,将数据从复杂的原始数据空间映射到一个更加合适的深度特征空间中. 相对于原始输入数据,对比深度 FCM 聚类网络所学习到的特征更适用于聚类任务. 同时设计了一种联合加权损失函数,该函数由网络重构损失、对比损失和聚类损失组成,不仅提升了对复杂样本的判别能力,也增强了分配样本的确定性.

(2)基于三支决策的思想,将第一阶段的聚类结果划分为正域和边界域,使用半球邻域粒化方法,为不确定样本构造信息粒,接着利用证据理论对信息粒中的信任度进行融合,有效提高样本隶属分配的准确率.

(3)将本文提出的算法与现有的聚类方法在多个类型数据集上进行实验比较. 由于对比深度 FCM 聚类网络框架、三支决策划分方法和信息粒化与证据融合的优势,有效地提升了聚类效果. 实验结果表明本文所提出方法相比于现有方法更具优势.

1 相关概念介绍

1.1 FCM 聚类

FCM 聚类算法首先初始化聚类中心和隶属度参数,接着计算每个样本对每个聚类中心的隶属度,根据计算得到的隶属度更新聚类中心的位置,这一过程不断迭代,直至满足停止条件. 该算法能够处理数据模糊归属情况. FCM 的目标函数 L 为:

$$L=\sum_{i=1}^{n}\sum_{j=1}^{k}u_{ij}^{m}\parallel \boldsymbol{x}_i-\boldsymbol{c}_j\parallel^2,1<m<\infty \tag{1}$$

式中,$i=1,2,\cdots,n$,$j=1,2,\cdots,k$,n 是样本总数,k 是聚类中心的数目,m 为超参数,u_{ij}是样本 $\boldsymbol{x}_i$ 对聚类中心 $\boldsymbol{c}_j$ 的隶属度,$\boldsymbol{x}_i$ 是第 i 个样本,$\boldsymbol{c}_j$ 是第 j 个聚类中心,$\parallel \boldsymbol{x}_i-\boldsymbol{c}_j\parallel$ 是第 i 个样本到第 j 个聚类中心的欧氏距离.

设 $\boldsymbol{U}=\{u_{ij}\}$ 为 $n\times k$ 的隶属度矩阵,u_{ij}是每个样本属于每个聚类中心的概率大小,隶属度矩阵满足 $\sum_{j=1}^{k}u_{ij}=1,\forall i=1,2,\cdots,n.$ 根据隶属度矩阵,将样本 $\boldsymbol{x}_i$ 划分到第 a 类中,其中 $a=\arg\max_j\{u_{ij},i=1,2,\cdots,n\}$. 目标函数中聚类中心 $\boldsymbol{c}_j$ 和隶属度矩阵 u_{ij}迭代更新公式如下:

$$\boldsymbol{c}_j=\frac{\sum_{i=1}^{n}u_{ij}^{m}\boldsymbol{x}_i}{\sum_{i=1}^{n}u_{ij}^{m}}, \tag{2}$$

$$u_{ij}=\left(\sum_{K=1}^{k}\left(\frac{\parallel \boldsymbol{x}_i-\boldsymbol{c}_j\parallel}{\parallel \boldsymbol{x}_i-\boldsymbol{c}_K\parallel}\right)^{\frac{2}{(m-1)}}\right)^{-1}. \tag{3}$$

根据上述迭代更新公式,对样本的隶属度矩阵和聚类中心进行迭代更新,当目标函数收敛时,得到样本对于聚类中心的隶属度矩阵 $\boldsymbol{U}=\{u_{ij}\}$ 和聚类中心 $\boldsymbol{c}_j$. 根据最终每个样本 $\boldsymbol{x}_i$ 的对各个聚类中心的隶属度大小 u_{ij}划分该不确定样本.

1.2　三支聚类

假设有 n 个样本的数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}$,$\boldsymbol{x}_i$ 是数据集中第 i 个样本,有 k 个聚类中心 $\boldsymbol{C}=\{\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_j,\cdots,\boldsymbol{c}_k\}$,$\boldsymbol{c}_j$ 是数据集中第 j 个聚类中心,数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}$ 中确定属于、确定不属于和不确定三种关系的样本可以分为三部分,分别是正域 $POS(\boldsymbol{c}_j)$、边界域 $BND(\boldsymbol{c}_j)$ 和负域 $NEG(\boldsymbol{c}_j)$,正域 $POS(\boldsymbol{c}_j)$ 是确定属于聚类中心 $\boldsymbol{c}_j$ 的样本的集合域,边界域 $BND(\boldsymbol{c}_j)$ 是可能属于聚类中心 $\boldsymbol{c}_j$ 的样本的集合域,负域 $NEG(\boldsymbol{c}_j)$ 是确定不属于聚类中心 $\boldsymbol{c}_j$ 的样本的集合域,正域 $POS(\boldsymbol{c}_j)$ 边界域 $BND(\boldsymbol{c}_j)$ 和负域 $NEG(\boldsymbol{c}_j)$ 满足以下条件:

(1) $POS(\boldsymbol{c}_j)\neq\varnothing, j=1,2,\cdots,k$,

(2) $POS(\boldsymbol{c}_i)\cap POS(\boldsymbol{c}_j)=\varnothing, i\neq j$,

(3) $\bigcup_{j=1}^{k}(POS(\boldsymbol{c}_j)\cup BND(\boldsymbol{c}_j))=\boldsymbol{X}$,

(4) $NEG(\boldsymbol{c}_j)\cup POS(\boldsymbol{c}_j)\cup BND(\boldsymbol{c}_j)=\boldsymbol{X}$.

条件(1)要求每个簇不能为空,条件(2)要求簇的正域两两不相交,条件(3)要求 $\boldsymbol{X}$ 中的每个元素必须至少属于簇的正域或者边界域的一个,一个元素也有可能属于多个簇,条件(4)要求簇的正域、边界域和负域合并起来为全域.

2　提出的方法

2.1　深度网络框架

对于一个数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_n\}\in R^{n\times d}$,$\boldsymbol{x}_i$ 是数据集中第 i 个样本,n 是样本总数,d 是样本的维度. 本文提出的基于证据信息粒化的深度三支 FCM 聚类包括编码器、解码器和两阶段的 FCM 聚类方法三部分,网络框架如图 1 所示.

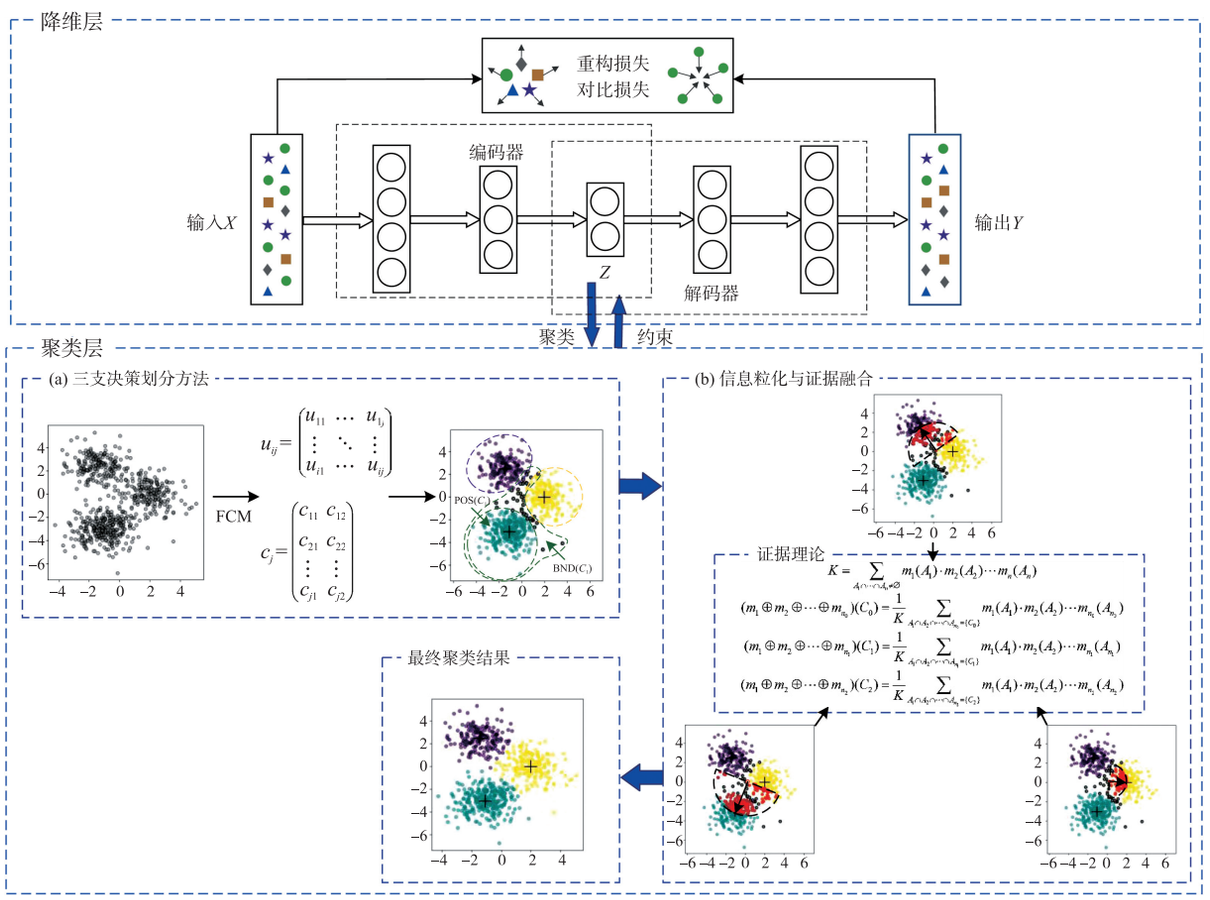


图 1　基于证据信息粒化的深度三支 FCM 聚类方法模型结构

Fig. 1　Deep three-way FCM clustering method based on evidential information granulation model architecture

本文使用(2M+1)层神经网络将原始数据映射到非线性特征空间. 其中 M 个隐藏层作为自编码器的编码器来降低原始数据的维度,另外 M 个隐藏层作为解码器来重构数据,第 m 次映射由第 m 层神经网络和第 m+1 层神经网络来完成. 第 m 层编码器的输出表示如下:

$$\boldsymbol{h}_m = f_m(act(\boldsymbol{h}_{m-1}); \boldsymbol{w}_m, \boldsymbol{b}_m), \tag{4}$$

式中,$m=1,2,\cdots,M$,$f(\cdot)$是编码器第 m 层到第 m+1 层的映射函数,$act(\cdot)$为激活函数,$\boldsymbol{w}_m$ 和 $\boldsymbol{b}_m$ 是该层对应的权重矩阵和偏置. 数据通过自编码器得到降维后的数据 $\boldsymbol{Z}$. 编码器将 d 维输入数据 $\boldsymbol{X}$ 映射到低维特征 $\boldsymbol{Z}$,可以表示为 $\boldsymbol{Z}=F(act(\boldsymbol{X}); \boldsymbol{W}, \boldsymbol{b})$,其中 $F(\cdot)$是编码器映射函数.

解码器部分将 h_m 转换为重构数据,转化公式表示如下:

$$\boldsymbol{y}_m = g_{2M+1-m}(act(\boldsymbol{h}_m)), \tag{5}$$

式中,$g(\cdot)$是编码器第(2M−m+1)层到第(2M−m+2)层的映射函数. 解码器将低维特征 $\boldsymbol{Z}$ 重构为 d 维数据 $\boldsymbol{Y}$,可以表示为 $\boldsymbol{Y}=G(\boldsymbol{Z})$,其中 $G(\cdot)$是解码器映射函数.

在聚类层中,聚类网络与编码器具有相同的结构,初始参数是在对比深度 FCM 聚类网络框架中学习到的 $\boldsymbol{Z}$,首先使用模糊聚类算法得到样本初始隶属度矩阵 $\boldsymbol{U}$,根据 $\boldsymbol{U}$ 和三支聚类策略,将第一阶段的聚类结果划分为正域和边界域,使用半球邻域粒化方法,通过构造信息粒来捕捉不确定样本的局部结构和相互作用,更准确地反映数据的内在特征. 基于半球邻域定义每个样本的局部区域,形成信息粒. 应用证据理论融合这些信息粒中的信任度,计算每个不确定样本相对于不同聚类中心的最终隶属度矩阵 $\boldsymbol{U}'$. 最后根据计算得到的最终隶属度矩阵 $\boldsymbol{U}'$,将样本准确分配到对应的聚类中心,实现更精确的聚类分析.

2.2 损失函数

本节将具体介绍本工作深度模型训练过程中使用的 3 种损失函数,包括重构损失、对比损失和聚类损失.

假设输入数据是 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}\in R^{n\times d}$,它对应的重构输出为 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}\in R^{n\times d}$,本文将输入原始数据和重构数据之间的均方误差作为重构损失. 上述重构损失定义为:

$$L_{\text{rec}} = \| \boldsymbol{Y}-\boldsymbol{X} \|_F^2. \tag{6}$$

对比学习的目标是最大化正对之间的相似性同时最小化负对的相似性,本文应用对比学习来增强从不同聚类中获取判别信息. 在本文提出的框架中,正对由原始数据和其对应的重构数据组成. 对于输入数据 $\boldsymbol{X}$ 和其对应的重构输出 $\boldsymbol{Y}$,其中对于每一个样本 $\boldsymbol{x}_i$,$i=1,2,\cdots,n$,样本的正对$\{\boldsymbol{x}_i,\boldsymbol{y}_i\}$包含 $\boldsymbol{x}_i$ 和它对应的重构样本 $\boldsymbol{y}_i$,样本的 2n−2 个负对包含 $\boldsymbol{x}_i$ 和其它 2n−2 个除 $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_i$ 以外的样本. $\boldsymbol{x}_i$ 和 $\boldsymbol{y}_j$ 的相似性定义为:

$$s(\boldsymbol{x}_i,\boldsymbol{y}_j)=\frac{(\boldsymbol{x}_i^{\mathrm{T}})(\boldsymbol{y}_j)}{\| \boldsymbol{x}_i \| \| \boldsymbol{y}_j \|}, \tag{7}$$

式中,$i=1,2,\cdots,n$,$j=1,2,\cdots,n$. $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 在实例级的对比损失使用以下等式获得:

$$l_i^x = -\log \frac{\exp(s(\boldsymbol{x}_i,\boldsymbol{y}_i)/\tau)}{\sum_{j=1}^{n}[\exp(s(\boldsymbol{x}_i,\boldsymbol{x}_j)/\tau) + \exp(s(\boldsymbol{x}_i,\boldsymbol{y}_j)/\tau)]}, \tag{8}$$

$$l_i^y = -\log \frac{\exp(s(\boldsymbol{y}_i,\boldsymbol{x}_i)/\tau)}{\sum_{j=1}^{n}[\exp(s(\boldsymbol{y}_i,\boldsymbol{y}_j)/\tau) + \exp(s(\boldsymbol{y}_i,\boldsymbol{x}_j)/\tau)]}, \tag{9}$$

式中,τ 是实例级的温度系数,温度系数 τ 越小,对于负样本的影响就越大,这意味着模型更倾向于将不相似的样本彼此分开. 反之,如果 τ 较大,模型将更平等地对待正负样本,可能导致特征空间中的样本分布更加平均. 因此,τ 在对比损失的计算中起到了平衡正负样本影响的作用. 本文取 $\tau=0.5$. 总对比损失由下式给出:

$$L_{\text{con}} = \frac{1}{2n}\sum_{j=1}^{n}(l_i^x+l_i^y). \tag{10}$$

平方 ℓ_2 范数,通常表示为 $\| \boldsymbol{m} \|_2^2$,它对大的异常值敏感,但是当异常值较小时,往往能取得比较好的效果. ℓ_1 范数,通常表示为 $\| \boldsymbol{m} \|_1$,它对小的异常值具有鲁棒性,但对大的异常值敏感. $\ell_{2,1}$范数,通常表示为 $\| \boldsymbol{m} \|_{2,1}$,它对小的异常值敏感,对大的异常值具有鲁棒性,Frobenius 范数,通常表示为 $\| \boldsymbol{m} \|_F^2$,它对大

的异常值敏感. 为了充分利用 $\ell_{2,1}$范数和 Frobenius 范数的优势,根据文献[31-32],提出了自适应损失函数,表示如下:

$$\| \boldsymbol{M} \|_{\sigma} = \sum_{i} \frac{(1+\sigma) \| \boldsymbol{m}^{i} \|_{2}^{2}}{\| \boldsymbol{m}^{i} \|_{2} + \sigma}, \tag{11}$$

式中,σ 是一个权衡参数来控制各种类型异常值的鲁棒性. $\| \boldsymbol{M} \|_{\sigma}$ 的性质总结如下. $\| \boldsymbol{M} \|_{\sigma}$ 是二次可微、凸函数且非负的,因此适合作为损失函数. 对于任意 i,如若 $\| \boldsymbol{M}_{i} \| \leqslant \sigma$,则 $\| \boldsymbol{M} \|_{\sigma} \rightarrow \frac{1+\sigma}{\sigma} \| \boldsymbol{M} \|_{F}^{2}$,对于任意 i,如若 $\| M_{i} \| \geqslant \sigma$,则 $\| \boldsymbol{M} \|_{\sigma} \rightarrow (1+\sigma) \| \boldsymbol{M} \|_{2,1}$,当 $\sigma \rightarrow 0$ 时,$\| \boldsymbol{M} \|_{\sigma} \rightarrow \| \boldsymbol{M} \|_{2,1}$,当 $\sigma \rightarrow \infty$ 时,$\| \boldsymbol{M} \|_{\sigma} \rightarrow \| \boldsymbol{M} \|_{F}^{2}$.

熵越大,无序性越高,根据文献[27]和公式(11),对于有 N 个数据的数据矩阵 $\boldsymbol{X}=[\boldsymbol{x}_{1},\boldsymbol{x}_{2},\cdots,\boldsymbol{x}_{N}]$,具有自适应损失函数和熵正则化的 FCM 聚类目标函数表示如下:

$$L_{\mathrm{clu}} = \sum_{i=1}^{N} \sum_{j=1}^{k} u_{ij} \frac{(1+\sigma) \| \boldsymbol{x}_{i} - \boldsymbol{c}_{j} \|_{2}^{2}}{\| \boldsymbol{x}_{i} - \boldsymbol{c}_{j} \|_{2} + \sigma} + \gamma u_{ij} \log u_{ij}, \tag{12}$$

式中,γ 是控制 u_{ij}的权衡参数. 公式(12)可以改写为:

$$L_{\mathrm{clu}} = \sum_{i=1}^{N} \sum_{j=1}^{k} u_{ij} \| \boldsymbol{x}_{i} - \boldsymbol{c}_{j} \|_{\sigma} + \gamma u_{ij} \log u_{ij}, \tag{13}$$

式中,$\sum_{j=1}^{k} u_{ij} = 1, 0 < u_{ij} < 1$.

本文所提出的深度聚类框架的融合损失函数表示如下:

$$L = L_{\mathrm{rec}} + \lambda_{1} L_{\mathrm{con}} + \lambda_{2} L_{\mathrm{clu}} = \| \boldsymbol{Y} - \boldsymbol{X} \|_{F}^{2} + \lambda_{1} \frac{1}{2n} \sum_{j=1}^{n} (l_{i}^{x} + l_{i}^{y}) + \lambda_{2} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} u_{ij} \| \boldsymbol{x}_{i} - \boldsymbol{c}_{j} \|_{\sigma} + \gamma u_{ij} \log u_{ij} \right) \tag{14}$$

式中,λ_{1} 和 λ_{2} 是权衡参数,L_{rec}、L_{con}和 L_{clu}分别是重构损失、对比损失和聚类损失.

2.3　三支决策划分方法

假设在一次训练过程中会得到降维后的特征样本集 $\boldsymbol{Z}=\{\boldsymbol{z}_{1},\boldsymbol{z}_{2},\cdots,\boldsymbol{z}_{i},\cdots,\boldsymbol{z}_{n}\}$,$\boldsymbol{z}_{i}$ 是样本集中第 i 个样本,n 是样本的个数,聚类中心 $\boldsymbol{C}=\{\boldsymbol{c}_{1},\boldsymbol{c}_{2},\cdots,\boldsymbol{c}_{j},\cdots,\boldsymbol{c}_{k}\}$,$\boldsymbol{c}_{j}$ 是样本集中第 j 个聚类中心,k 是聚类中心总数,隶属度矩阵 $\boldsymbol{U}=\{u_{ij}\}$,$i=1,2,\cdots,n$,$j=1,2,\cdots,k$,$u_{ij}$表示第 i 个样本对第 j 个聚类中心的隶属度. 借鉴三支聚类的思想,根据隶属度矩阵 $\boldsymbol{U}=\{u_{ij}\}$,将样本根据对聚类中心 $\boldsymbol{c}_{j}$ 的隶属度划分为正域、边界域和负域,正域中的样本为确定属于聚类中心 $\boldsymbol{c}_{j}$ 的样本,边界域中的样本为不确定属于聚类中心 $\boldsymbol{c}_{j}$ 的样本,负域中的样本为确定不属于聚类中心 $\boldsymbol{c}_{j}$ 的样本. 确定样本和不确定样本的划分过程如下:

$$\mathrm{mask}(\boldsymbol{z}_{i}) = \begin{cases} 1, & \max\{u_{i1},u_{i2},\cdots,u_{ik}\} \geqslant \varepsilon \\ 0, & \max\{u_{i1},u_{i2},\cdots,u_{ik}\} < \varepsilon \end{cases}, \tag{15}$$

式中,k 是聚类中心的数量,ε 是划分阈值,mask 为 1 的样本为确定样本,mask 为 0 的样本为不确定样本. 样本对聚类中心的隶属度越大,那么该样本越可能属于聚类中心所属聚类. 如果 $\boldsymbol{z}_{i}$ 对于聚类中心 $\boldsymbol{c}_{j}$ 的隶属度大于 ε,则 $\boldsymbol{z}_{i}$ 分配给该聚类中心 $\boldsymbol{c}_{j}$ 的正域,属于确定分配给聚类中心 $\boldsymbol{c}_{j}$ 的样本,如果 $\boldsymbol{z}_{i}$ 对于每一个聚类中心的隶属度都小于等于 ε,则 $\boldsymbol{x}_{i}$ 这个样本划分为该聚类中心边界域,属于不确定样本,需要进行再分配. 聚类中心 $\boldsymbol{c}_{j}$ 的正域 $POS(\boldsymbol{c}_{j})$、边界域 $BND(\boldsymbol{c}_{j})$和负域 $NEG(\boldsymbol{c}_{j})$表示如下:

$$POS(\boldsymbol{c}_{j}) = \{\boldsymbol{z}_{i} \in POS(\boldsymbol{c}_{j}) \mid \mathrm{mask}(\boldsymbol{z}_{j}) = 1, j = \underset{i}{\operatorname{argmax}} u_{ij}\}, \tag{16}$$

$$BND(\boldsymbol{c}_{j}) = \{\boldsymbol{z}_{i} \in BND(\boldsymbol{c}_{j}) \mid \mathrm{mask}(\boldsymbol{z}_{j}) = 0\}, \tag{17}$$

$$NEG(\boldsymbol{c}_{j}) = \{\boldsymbol{z}_{i} \in NEG(\boldsymbol{c}_{j}) \mid \mathrm{mask}(\boldsymbol{z}_{j}) = 1, j \neq \underset{i}{\operatorname{argmax}} u_{ij}\}. \tag{18}$$

2.4　信息粒化与证据融合

针对边界域中的不确定性样本,本文基于粒计算思想,提出基于余弦定律的半球信息粒化方法,并根据证据理论实现第二阶段分配.

假设不确定样本 $\boldsymbol{z}_{i}$、聚类中心 $\boldsymbol{c}_{j}$ 和 cen 个样本的确定样本集合 $\boldsymbol{Z}_{\mathrm{certain}}=\{\boldsymbol{z}_{1},\boldsymbol{z}_{2},\cdots,\boldsymbol{z}_{a},\cdots,\boldsymbol{z}_{cen}\}$,$\boldsymbol{z}_{a}$ 是确定样本集合中第 a 个样本,以 $\boldsymbol{z}_{i}$ 在空间的位置为圆心,$\boldsymbol{z}_{i}$ 到 $\boldsymbol{c}_{j}$ 的欧式距离为半径,面向聚类中心 $\boldsymbol{c}_{j}$ 构造半

球,半球构造如图 2 所示. 半球内样本集 $\boldsymbol{Z}_{\text{semiball}}$ 的选取方法如下:

$$\theta=\arccos\frac{\boldsymbol{a}\cdot\boldsymbol{b}}{\|\boldsymbol{a}\|\ \|\boldsymbol{b}\|},\tag{19}$$

$$\boldsymbol{Z}_{\text{semiball}}=\{z_a\in\boldsymbol{Z}_{\text{semiball}}\mid\theta\leqslant 90°,d_1\leqslant d_2\},\tag{20}$$

式中,$\boldsymbol{a}=z_a-z_i$ 是 z_i 指向 z_a 的向量,$\boldsymbol{b}=\boldsymbol{c}_j-z_i$ 是 z_i 指向 $\boldsymbol{c}_j$ 的向量,θ 是向量 $\boldsymbol{a}$ 与向量 $\boldsymbol{b}$ 的夹角,$d_1=\|z_i-z_a\|_2$ 是不确定样本 z_i 到确定样本 z_a 的距离,$d_2=\|z_i-\boldsymbol{c}_j\|_2$ 是不确定样本 z_i 到聚类中心 $\boldsymbol{c}_j$ 的距离.

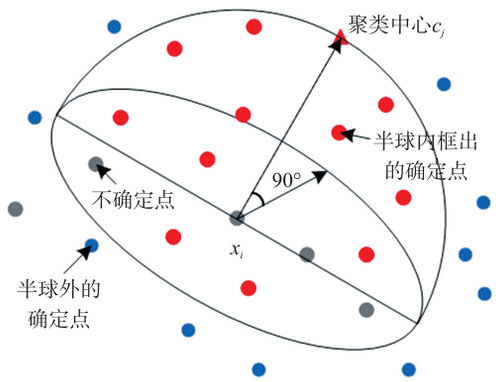


图 2　半球构造

Fig. 2　Semi-ball structure

计算得到 $\boldsymbol{Z}_{\text{semiball}}=\{z_{N_1},z_{N_2},\cdots,z_{N_i},\cdots,z_{N_l}\}$,$l$ 是半球内框出的邻居样本的个数,z_{N_i} 是确定样本集合中第 i 个样本. 不确定样本 z_i 与聚类中心 $\boldsymbol{c}_j$ 构造出的半球内取出所有确定样本的隶属度矩阵 $\boldsymbol{U}_{(i,j)}=\{u_{N_ij}\}$ 如下:

$$\boldsymbol{U}_{(i,j)}=\begin{pmatrix}u_{N_11} & u_{N_12} & \cdots & u_{N_1k}\\ u_{N_21} & u_{N_22} & \cdots & u_{N_2k}\\ \vdots & \vdots & \ddots & \vdots\\ u_{N_l1} & u_{N_l2} & \cdots & u_{N_lk}\end{pmatrix}_{l\times k},\tag{21}$$

式中,u_{N_ij} 是邻居样本 z_{N_i} 相对聚类中心 $\boldsymbol{c}_j$ 的隶属度.

给定 z_{N_i},$\boldsymbol{c}_j$ 和 $\boldsymbol{U}_{(i,j)}=\{u_{N_ij}\}$,采用半球邻域粒化方法构建不确定样本的信息粒,不确定样本 z_{N_i} 与聚类中心 $\boldsymbol{c}_j$ 构造出的半球信息粒可表示为 $[z_{N_i}]_{\boldsymbol{c}_j}=\dfrac{u_{N_1j}}{z_{N_1}}+\dfrac{u_{N_2j}}{z_{N_2}}+\cdots+\dfrac{u_{N_lj}}{z_{N_l}}$.

设 $\boldsymbol{\Theta}$ 是一个识别框架. 在识别框架上的基本概率分配(basic probability assignment,BPA)是一个 $2^{\boldsymbol{\Theta}}\to[0,1]$ 的函数 m,称为 mass 函数. 并且满足 $m(\varnothing)=0$ 且 $\sum\limits_{A\subseteq\boldsymbol{\Theta}}m(A)=1$. n 个 mass 函数的 Dempster 合成规则的定义是,对于 $\forall A\subseteq\boldsymbol{\Theta}$,$\boldsymbol{\Theta}$ 上的有限的 mass 函数 $m_1,m_2,\cdots,m_n$ 的 Dempster 合成规则为:

$$(m_1\oplus m_2\oplus\cdots\oplus m_n)(A)=\frac{1}{K}\sum_{A_1\cap A_2\cap\cdots\cap A_n=A}m_1(A_1)\cdot m_2(A_2)\cdots m_n(A_n),\tag{22}$$

式中,

$$K=\sum_{A_1\cap A_2\cap\cdots\cap A_n\neq\varnothing}m_1(A_1)\cdot m_2(A_2)\cdots m_n(A_n)=1-\sum_{A_1\cap A_2\cap\cdots\cap A_n=\varnothing}m_1(A_1)\cdot m_2(A_2)\cdots m_n(A_n).\tag{23}$$

在本文中,对有限个 mass 函数 $m_{N_1},m_{N_2},\cdots,m_{N_l}$ 运用 Dempster 合成规则进行合成,使用合成规则得到不确定样本 z_i 相对聚类中心 $\boldsymbol{c}_j$ 的隶属度 u_{ij},公式如下:

$$u_{ij}=\frac{1}{K}\sqrt[l]{\prod_{i=1}^{l}m_{N_i}},\tag{24}$$

式中,K 是归一化常数,$K=\sum\limits_{j=1}^{k}\sqrt[l]{\prod\limits_{i=1}^{l}m_{N_i}}$,$k$ 是聚类中心总数,l 是面向聚类中心 $\boldsymbol{c}_j$ 构造半球内框出的邻居样本的个数.

mass 函数 m_{N_i} 为样本 z_{N_i} 相对聚类中心 $\boldsymbol{c}_j$ 的隶属度 u_{N_ij},公式(24)可改写为:

$$u_{ij}=\frac{1}{K}\sqrt[l]{\prod_{i=1}^{l}u_{N_ij}},\tag{25}$$

式中,$K=\sum\limits_{j=1}^{k}\sqrt[l]{\prod\limits_{i=1}^{l}u_{N_ij}}$,$u_{N_ij}$ 是不确定样本 z_i 的邻居相对聚类中心 $\boldsymbol{c}_j$ 的隶属度.

将新得出的不确定样本隶属度替换掉所有原来的隶属度,得到样本 z_i 对于聚类中心 $\boldsymbol{c}_j$ 的最终隶属度 $\boldsymbol{U}'=\{u_{ij}\}$,$i=1,2,\cdots,n$,$j=1,2,\cdots,k$. 根据最终得到的隶属度划分样本到各个聚类中心.

基于此,本文所提的半球信息粒化与证据融合方法算法伪代码如算法 1 所示.

算法 1　信息粒化与证据融合方法

输入:不确定样本集 $\boldsymbol{Z}_{\text{uncertain}}$;
输出:最终隶属度矩阵 $\boldsymbol{U}'$.
$\boldsymbol{M}_j$=[ ];//初始化
for sample in $\boldsymbol{Z}_{\text{uncertain}}$ do://对不确定样本进行再分配
　$\boldsymbol{M}_j$=Construct_semiball_information_granules(sample,$\boldsymbol{c}_j$);//使用半球邻域粒化方法
　if $\boldsymbol{Z}_{\text{certain}}$ in $\boldsymbol{M}_j$://如果确定样本在半球内
　　$\boldsymbol{U}_{\text{evidence}}$=Evidence($\boldsymbol{U}_{M_j}$);//根据邻域内确定样本的隶属度计算得出不确定样本的最终隶属度
end for;
$\boldsymbol{U}'=\boldsymbol{U}_{\text{certain}}+\boldsymbol{U}_{\text{evidence}}$;
Return $\boldsymbol{U}'$.

3　实验设置

3.1　数据集

为了评估提出的方法,本文在 5 个数据集上进行了实验,数据集包括手写数字[33]、人脸[34-35]和物体[36],表 1 对每个数据集进行了简要描述.

表 1　数据集信息

Table 1　Dataset information

名称	样本个数	维度	类别
USPS	2 000	256	10
ORL	400	1 024	40
JAFFE	213	1 024	10
COIL-20	1 440	1 024	20
COIL-100	7 200	1 024	100

3.2　基准方法与评价指标

本文选取了 6 种当前流行并使用广泛的聚类方法与提出的方法进行比较,来评估本文所提出方法的效果. 其简介如下:

(1)K-means[9]是一种硬聚类划分方法,通过迭代优化,使得簇内距离最小化,簇间距离最大化.

(2)FKM[37]为每个样本分配属于各个簇的概率性隶属度,通过迭代优化使得簇内相似度最大化.

(3)DEC[26]是一种使用深度神经网络同时学习特征表示和聚类的方法. 它学习从高维空间到低维特征空间的映射,并迭代优化聚类目标.

(4)DEKM[38]提出通过一个包含 K-means 的类内散点矩阵的特征向量的标准正交变换矩阵将嵌入空间转化为揭示聚类结构信息的新空间,增加了集群结构的信息.

(5)DFKM[27]提出了一种具有自适应损失函数和熵正则化的新型深度模糊 K-means 模型(DFKM). DFKM 同时进行深度特征提取和模糊聚类,以生成更合适的非线性特征图.

(6)CC[29]对于给定的数据集,通过数据增强构造正负对,然后将其投影到特征空间中. 其中,通过最大化正对的相似性同时最小化负对的相似性,分别在行和列空间中进行实例级和聚类级对比学习.

本文使用 3 个常用的评价指标,分别是准确率(Accuracy,ACC)、标准化互信息(Normalized Mutual Information,NMI)和 Purity,3 个都是聚类模型评估指标的外部方法.

4　实验结果

4.1　对比结果

本文在 5 个不同类型的数据集上进行 10 次实验,并记录结果的平均值,聚类性能如表 2 所示,最好的结果加粗表示. 根据实验结果,可得出以下结论.

与 KM、FKM 等传统聚类方法相比,本文所提出的方法在 3 个指标上均取得了显著更好的结果. 例如,在 USPS 手写数字识别数据集上,本文所提出的方法的平均 ACC、NMI 和 Purity 分别为 76.54%、72.64%和 81.35%,比最好的传统聚类方法分别高出 11.87%、9.97%和 9.23%. 这种性能提升主要归因于通过重构损失和对比损失训练的自编码器生成了更有利于聚类的特征空间,从而增强聚类模型的鲁棒性并提高性能.

与 DEC 和 DFKM 等深度聚类方法相比,本文所提出的方法在大多数数据集上取得了最好的结果. 这可能是因为这些方法使用仅有重构损失的自动编码器,而本文所提出的方法采用重构损失与对比损失相

结合的策略来训练自编码器. 通过增加对比学习任务,模型在训练过程中不仅仅关注重构任务,还要优化对比学习任务. 这种多任务学习的方式可以有效缓解过拟合问题,提升模型的泛化能力. 虽然 DEKM 在大多数数据集上表现良好,但其训练过程中需要同时优化表示学习和聚类两个目标,这增加了模型的复杂性,导致较长的训练时间. 与仅基于对比损失的 CC 方法相比,本文所提出的方法在 ACC 和 NMI 两个指标上均优于 CC,这是由于 CC 依赖数据增强策略,如果数据集不适合数据增强或增强策略不恰当,可能会导致性能下降.

表 2 聚类结果

Table 2 Clustering results

模型	评价指标	USPS	ORL	JAFFE	COIL-20	COIL-100
KM	ACC	64.67	59.30	86.01	65.78	49.23
	NMI	62.67	77.67	89.27	78.09	76.67
	Purity	72.12	63.85	87.56	69.29	54.51
FKM	ACC	61.47	56.35	73.05	58.04	48.08
	NMI	57.37	76.97	82.08	75.89	76.93
	Purity	66.61	60.43	75.68	62.30	53.00
DEC	ACC	71.59	39.15	57.09	53.94	32.08
	NMI	68.20	63.84	66.46	73.70	67.57
	Purity	78.82	42.78	60.09	55.76	34.64
DEKM	ACC	75.06	43.68	75.77	65.85	50.90
	NMI	**75.30**	68.27	81.78	78.01	72.50
	Purity	81.33	48.10	78.36	68.06	54.81
DFKM	ACC	71.27	62.65	90.00	64.70	51.33
	NMI	70.72	81.17	92.94	78.47	77.82
	Purity	77.41	66.85	90.56	68.11	56.28
CC	ACC	60.49	47.73	91.31	65.44	49.27
	NMI	58.75	71.89	92.68	77.39	78.43
	Purity	62.16	58.30	**93.71**	**75.26**	**66.19**
Ours	ACC	**76.54**	**65.05**	**92.91**	**67.67**	**51.71**
	NMI	72.64	**81.97**	**93.24**	**78.49**	**78.52**
	Purity	**81.35**	**68.50**	92.91	70.45	56.37

4.2 消融实验

为了分析重构损失、对比损失、聚类损失和两阶段的 FCM 聚类方法在网络中的作用,本文在 USP 和 ORL 这两个数据集进行了消融实验,探讨所提出的方法中每个损失项和两阶段的 FCM 聚类方法对聚类性能的影响. 具体来说,本文首先通过删除相应的损失项构建了 3 个消融模型,然后将两阶段的 FCM 聚类方法改换为单阶段的 FCM 聚类方法构建了第四个消融模型,并在 USPS 和 ORL 数据集上进行了 10 次实验,并记录结果的平均值. 消融实验的结果如表 3 所示,从表 3 中本文可以得出如下判断.

表 3 USPS 与 ORL 数据集的消融实验结果

Table 3 The results of ablation experiments on USPS and ORL dataset

模型	USPS			ORL		
	ACC	NMI	Purity	ACC	NMI	Purity
w/oL_{rec}	60.75	60.46	65.93	42.75	66.76	46.65
w/oL_{con}	69.71	69.69	77.53	60.00	79.53	64.35
w/oL_{cluster}	71.79	70.69	78.32	60.43	80.00	64.78
w/o two-stage	76.12	72.19	80.96	63.13	80.96	66.90
Ours	**76.54**	**72.64**	**81.35**	**65.05**	**81.97**	**68.50**

去掉重构损失后,USPS 数据集的 3 个指标均显著降低,ORL 数据集的 3 个指标也明显下降,表明重构损失对模型在聚类性能上有非常大的作用. 去掉对比损失后,USPS 和 ORL 数据集的性能都有显著下降,表明对比损失在提高聚类效果方面起着关键作用. 去掉聚类损失后,USPS 和 ORL 数据集的性能有所下降. 这表明聚类损失对最终的聚类结果有一定的影响. 两阶段的 FCM 聚类方法更加关注数据样本的局部结构,可以准确地捕捉数据的内在特征. 使用单阶段 FCM 聚类方法替代两阶段的 FCM 聚类方法后,性能

有所下降,这表明两阶段 FCM 聚类方法在提高聚类性能方面具有优势. 完整方法的所有性能指标均为最高. 这说明各个损失项和两阶段 FCM 聚类方法的组合能够最大化地提高模型的聚类性能.

4.3　不同参数的影响

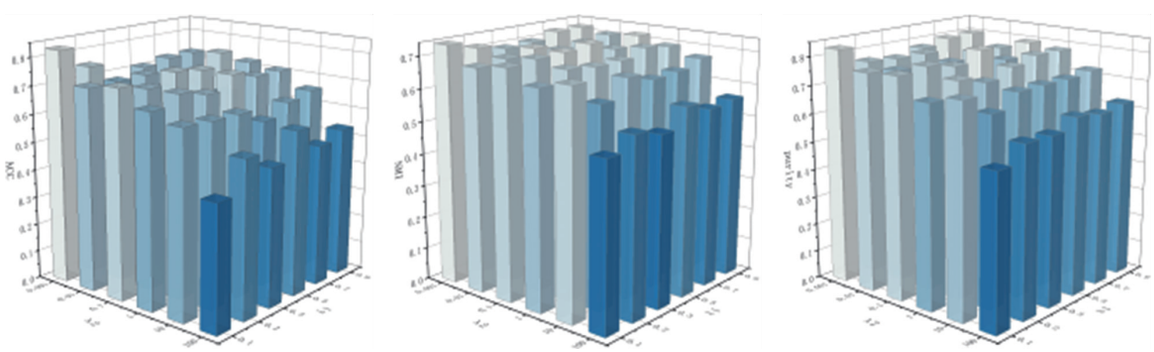
为了在不同的初始化条件下得到更稳定的训练结果,本文对自编码器进行了预训练. 具体来说,对于 COIL-20,本文先训练 1024-500-1024 的子网,然后训练 500-300-500 子网,将每一层子网分别训练 10 个周期,并将这些学习到的参数加载到对比深度 FCM 聚类网络中,对其进行初始化来加快训练速度. 本文使用 Adam[39] 优化器,batch 大小设置为 512,学习率为 0.001. 不同数据集的实验设置如表 4 所示.

表 4　超参数设置

Table 4　Hyperparameter settings

名称	Epochs	λ_1	λ_2	γ	σ	ε	Struct
USPS	30	0.1	0.001	0.1	0.1	0.7	256-128-80-128-256
ORL	900	0.01	0.1	0.1	0.005	0.7	1024-512-80-512-1024
JAFFE	250	0.3	1.5	0.1	0.1	0.7	1024-512-300-512-1024
COIL-20	50	0.5	0.01	0.1	0.1	0.7	1024-500-300-500-1024
COIL-100	250	0.3	0.001	0.1	0.01	0.5	1024-500-300-500-1024

本文所提出的方法有 7 个可以调整的超参数,即 Epochs、λ_1、λ_2、γ、σ、ε 和 Struct. 其中 Epochs 是模型正式训练过程中数据集完整遍历次数,λ_1 和 λ_2 是本文所提出的框架的最终损失函数的权衡参数,γ 是聚类损失中控制 u_{ij} 的权衡参数,σ 是聚类损失中一个来控制各种类型异常值的鲁棒性的权衡参数,ε 是三支决策划分确定样本和不确定样本的阈值,Struct 表示神经网络中每层神经元的数量. 对于大多数数据集,在[0.1,0.2,0.3,0.5,0.7,0.9]的网格中搜索 λ_1 的值,[0.001,0.01,0.1,1,10,100]的网格中搜索 λ_2 的值,在[0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]的网格中搜索 ε 的值,与一般深度模型类似,所有超参数都会很大程度上影响聚类结果. 不同 λ_1 和 λ_2 在 USPS 数据集上对聚类结果的影响如图 3 所示.



(a) ACC　　(b) NMI　　(c) Purity

图 3　不同 λ_1 和 λ_2 对聚类效果的影响

Fig. 3　The impact of different λ_1 and λ_2 on the clustering effect

首先,λ_1 取值适中时,对比损失能够有效地优化特征空间,使相似样本聚集在一起,不相似样本分散开来,使样本在特征空间中的分布更加符合预期的聚类结构,增加了样本的可信度,从而提高聚类性能. 其次,λ_2 太高时会对聚类效果产生负面影响,ACC、NMI 和 Purity 3 个指标的值显著降低. 对比损失的权衡参数 λ_1 会影响特征空间的分布,聚类损失的权衡参数 λ_2 会影响聚类目标的优化. 总之 λ_1 和 λ_2 都需要在合适的范围内取值,才能在重构损失、对比损失和聚类损失之间取得平衡,从而实现最佳的聚类性能.

4.4　定性分析

本文所提出的方法在 USPS 和 JAFFE 数据集上的混淆矩阵如图 4 所示. 对角线上的元素表示正确聚类的样本比例即预测标签与实际标签一致的比例. 颜色越深,表示正确分类的样本数越多,聚类性能越好. 从图中可以看两个混淆矩阵的对角线颜色最深,大多数样本被正确分配到相应的类. 以 USPS 的混淆矩阵为例,有一些有趣的观察结果. 一些数字在手写时形状相似,容易混淆. 例如,3 和 5 在手写时可能形状相似,导致类别 3 被错误分类为 5 的比例较高. 类似的,4 和 9 也可能形状相似,导致类别 4 被错误分类为 9 的比例显著. 这与人类的思维是一致的,因为这些东西在现实世界的场景中有时会被混淆.

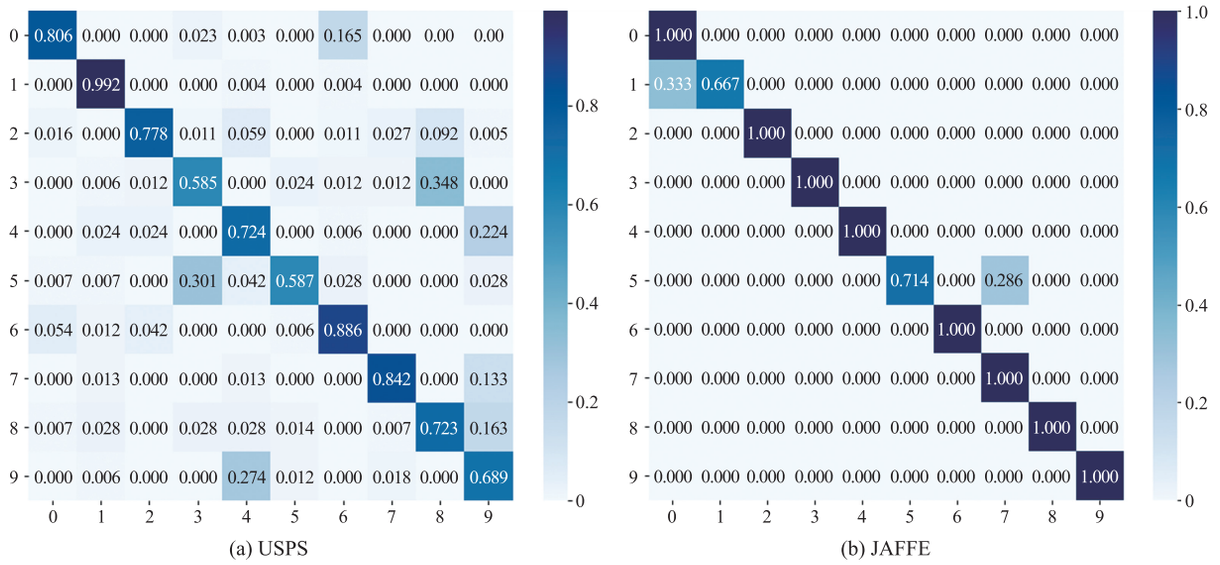


图 4　混淆矩阵

Fig. 4　Confusion matrix

5　结论

在本文中,提出了一种基于证据信息粒化的深度三支 FCM 聚类,首先提出一种对比深度 FCM 聚类网络框架,将数据从复杂的原始数据空间映射到一个更加适合聚类任务的深度特征空间中. 接着基于三支决策的思想,将第一阶段的聚类结果划分为正域和边界域,以处理数据中的不确定性. 最后引入半球邻域粒化方法,为不确定样本构造信息粒,利用证据理论对信息粒的信任度进行融合,使数据样本更加关注其局部结构,可以准确地捕捉数据的内在特征. 在多个公共数据集上的实验结果表明,本文所提出的方法优于其他方法. 本文工作的主要限制在于自编码器中使用全连接网络,未来希望可以通过更复杂的网络结构来增强自编码器.

[参考文献]

[1] TORTORA C,PALUMBO F. Clustering mixed-type data using a probabilistic distance algorithm[J]. Applied soft computing,2022,130:109704.

[2] JU H,LU Y,DING W,et al. Three-way evidence theory-based density peak clustering with the principle of justifiable granularity[J]. Applied soft computing,2024,152:111217.

[3] CHOUDHARY A,BADHOLIA A,SHARMA A,et al. A dynamic K-means-based clustering algorithm using fuzzy logic for CH selection and data transmission based on machine learning[J]. Soft computing,2023,27(10):6135-6149.

[4] RAJPATHAK D G,SINGH S. An ontology-based text mining method to develop D-matrix from unstructured text[J]. IEEE transactions on systems,man,and cybernetics:systems,2013,44(7):966-977.

[5] ZHANG Q,YANG L T,CHEN Z,et al. PPHOPCM:Privacy-preserving high-order possibilistic c-means algorithm for big data clustering with cloud computing[J]. IEEE transactions on big data,2017,8(1):25-34.

[6] KUMAR P,AGRAWAL R K,KUMAR D. Fast and robust spatial fuzzy bounded k-plane clustering method for human brain MRI image segmentation[J]. Applied soft computing,2023,133:109939.

[7] ZHAO Q,WANG C,WANG P,et al. A novel method on information recommendation via hybrid similarity[J]. IEEE transactions on systems,man,and cybernetics:systems,2016,48(3):448-459.

[8] CELEBI M E. Partitional clustering algorithms[M]. Berlin:Springer,2014.

[9] LLOYD S. Least squares quantization in PCM[J]. IEEE transactions on information theory,1982,28(2):129-137.

[10] VERMA H,VERMA D,TIWARI P K. A population based hybrid FCM-PSO algorithm for clustering analysis and segmentation of brain image[J]. Expert systems with applications,2021,167:114121.

[11] LESKI J M. Fuzzy c-ordered-means clustering[J]. Fuzzy sets and systems,2016,286:114-133.

[12] WANG Z, SONG Q, SOH Y C, et al. An adaptive spatial information-theoretic fuzzy clustering algorithm for image segmentation[J]. Computer vision and image understanding, 2013, 117(10): 1412-1420.
[13] ZHANG D Q, CHEN S C, PAN Z S, et al. Kernel-based fuzzy clustering incorporating spatial constraints for image segmentation[C]//Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693). Piscataway: IEEE, 2003, 4: 2189-2192.
[14] YAO Y. Three-way decisions with probabilistic rough sets[J]. Information sciences, 2010, 180(3): 341-353.
[15] YAO Y. The superiority of three-way decisions in probabilistic rough set models[J]. Information sciences, 2011, 181(6): 1080-1096.
[16] YAO Y. Three-way decisions and cognitive computing[J]. Cognitive computation, 2016, 8(4): 543-554.
[17] 徐天杰,王平心,杨习贝. 基于人工蜂群的三支 k-means 聚类算法[J]. 计算机科学,2023,50(06):116-121.
[18] WANG P, SHI H, YANG X, et al. Three-way k-means: Integrating k-means and three-way decision[J]. International journal of machine learning and cybernetics, 2019, 10: 2767-2777.
[19] DUNDAR A, JIN J, CULURCIELLO E. Convolutional clustering for unsupervised learning[J/OL]. arXiv Preprint arXiv: 1511.06241, 2015.
[20] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.
[21] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
[22] PENG X, XIAO S, FENG J, et al. Deep subspace clustering with sparsity prior[C]//International Joint Conference on Artifical Intelligen. San Francisco, 2016: 1925-1931.
[23] JI P, ZHANG T, LI H, et al. Deep subspace clustering networks[J]. Advances in neural information processing systems, 2017, 30: 1-10.
[24] LI F, QIAO H, ZHANG B. Discriminatively boosted image clustering with fully convolutional auto-encoders[J]. Pattern recognition, 2018, 83: 161-173.
[25] AFFELDT S, LABIOD L, NADIF M. Spectral clustering via ensemble deep autoencoder learning (SC-EDAE)[J]. Pattern recognition, 2020, 108: 107522.
[26] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]//International Conference on Machine Learning. New York, 2016: 478-487.
[27] ZHANG R, LI X, ZHANG H, et al. Deep fuzzy k-means with adaptive loss and entropy regularization[J]. IEEE transactions on fuzzy systems, 2019, 28(11): 2814-2824.
[28] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[C]// International Conference on Machine Learning. PMLR, 2020: 1597-1607.
[29] LI Y, HU P, LIU Z, et al. Contrastive clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Polo Alto, 2021, 35(10): 8547-8555.
[30] 许洁,王立松. 一种结构关系一致的对比聚类方法[J]. 计算机科学,2023,50(09):123-129.
[31] NIE F, WANG H, HUANG H, et al. Adaptive loss minimization for semi-supervised elastic embedding[C]//Twenty-Third International Joint Conference on Artificial Intelligence. San Francisco, 2013.
[32] DING C. A new robust function that smoothly interpolates between l1 and l2 error functions[J]. Univerisity of Texas at Arlington technical report, 2013.
[33] HULL J J. A database for handwritten text recognition research[J]. IEEE transactions on pattern analysis and machine intelligence, 1994, 16(5): 550-554.
[34] CAI D, ZHANG C, HE X. Unsupervised feature selection for multi-cluster data[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, 2010: 333-342.
[35] LYONS M J, BUDYNEK J, AKAMATSU S. Automatic classification of single facial images[J]. IEEE transactions on pattern analysis and machine intelligence, 1999, 21(12): 1357-1362.
[36] NENE S A, NAYAR S K, MURASE H. Columbia object image library (coil-20)[R]. Technical Report CUCS-006-93. 1996: 1-4.
[37] RUSPINI E H. A new approach to clustering[J]. Information and control, 1969, 15(1): 22-32.
[38] GUO W, LIN K, YE W. Deep embedded k-means clustering[C]//2021 International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE, 2021: 686-694.
[39] KINGMA D P, BA J. Adam: A method for stochastic optimization[J/OL]. arXiv Preprint arXiv: 1412.6980, 2014.

[责任编辑:陆炳新]