ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 5,Sep. 2022,pp. 1036－1048
DOI:10. 16337/j. 1004-9037. 2022. 05. 009

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

改进的自步深度不完备多视图聚类

崔金荣[1,2]，黄　诚[1]

(1.华南农业大学数学与信息学院,广州 510642; 2.广州市智慧农业重点实验室,广州 510642)

摘　要:随着数据量的增大,多视图聚类中出现带有缺失视图数据的情况愈发常见,此问题被称为不完备多视图聚类,而引入深度模型进行聚类通常可以获得比浅层模型更为出色的表现。本文提出一种新颖的深度不完备多视图聚类模型,称为改进的自步深度不完备多视图聚类。在该模型中,充分考虑多视图数据之间的互补性,利用基于多视图特性的最近邻填充方案将缺失视图补全。使用多个自编码器分别获取多个视图数据的低维潜在特征,同时引入图嵌入策略保持潜在特征之间的几何结构。运用一致性原则将来自不同的视图潜在特征融合以获得一致潜在特征,在此基础上运用自步学习的方法来增强聚类效果。实验结果表明,对比现有的不完备多视图聚类模型,本文模型可以更加灵活且高效地应对各种不完备多视图聚类情况,提升了不完备多视图聚类的鲁棒性与表现效果。

关键词:聚类;深度聚类;多视图聚类;缺失多视图;图嵌入

中图分类号:TP391　　**文献标志码**:A

Improved Self-paced Deep Incomplete Multi-view Clustering

CUI Jinrong[1,2], HUANG Cheng[1]

(1. College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China; 2. Guangzhou Key Laboratory of Intelligent Agriculture, Guangzhou 510642, China)

Abstract: With the increase of the volume of data, multi-view clustering with missing view data is becoming progressively common, which is regarded as the incomplete multi-view clustering. Powered by the development of deep learning models, clustering models introduced deep learning can normally get more outstanding performance than shallow models. A novel deep incomplete multi-view clustering model is proposed, which is called improved self-paced deep incomplete multi-view clustering. In this model, the complementarity of multi-view data is fully considered, and the missing views are completed by the nearest neighbor imputation scheme based on multi-view data characteristics. Multiple encoders are exerted to obtain the low-dimensional potential features of multiple views. Meanwhile, the graph embedding strategy is introduced to maintain the geometric structure among the potential features. The consistency principle is exerted to fuse the potential features from different views to obtain consistent potential features. Experimental results indicate that, compared with the existing incomplete multi-view clustering models, our model can deal with various incomplete multi-view clustering more flexibly and efficiently, thus improving the robustness and performance of incomplete multi-view clustering.

Key words: clustering; deep clustering; multi-view clustering; incomplete multi-view; graph embedding

---

基金项目:国家自然科学基金(61976097);广州市智慧农业重点实验室(201902010081)。
收稿日期:2021-10-24;**修订日期**:2022-01-26

引 言

在近些年提出的聚类方法中,对于同样的一个数据样本可以使用不同的视图来观测和描述。例如,一段电影片段可以用画面和音频两个视图来描述;相同的一张图片可以使用RGB图像和灰度图像来描述。使用多个视图数据之间的互补信息来提升聚类性能,称为多视图聚类。但是在实际情况中,由于人为因素或机器故障,在收集到的数据中可能会出现缺失视图的情况。在带有缺失视图的多视图数据上进行聚类任务,研究人员称之为"不完备多视图聚类"[1]。显然,传统的聚类方法无法进行不完备多视图数据的处理。此外,从缺失的多视图数据中得出不同视图之间的互补信息和一致信息较为困难,这使得不完备多视图聚类成为一项非常具有挑战性的任务。

为了应对不完备多视图聚类的种种问题,在过去的10年间研究人员已经提出了数种不完备多视图聚类模型。这些模型大致可以分为浅层模型与深层模型两类。在浅层模型中,一部分不完备多视图聚类模型设计了基于缺失视图对齐的多矩阵分解模型,以此来学习所有视图的一致特征,并将一致特征用于聚类[2]。但是,这类模型仅适用于两个视图之间的缺失情况。对于具有两个以上视图的任意缺失视图数据,研究人员提出了一系列基于加权多矩阵分解的模型[3]。这些模型大部分是在所有视图的矩阵分解项上添加一些预先构造好的特定于视图的对角矩阵,运用特定的对角矩阵避免缺失视图数据的负面影响。除了这些基于矩阵分解的不完备多视图聚类模型之外,基于多核的聚类模型[4-5]和基于图的模型[6]也开发出了缺失视图的模型。多核不完备多视图聚类模型通常对每个视图相对应的核的缺失部分进行预测和填充,然后从这些填充后的核中寻找所有视图的一致特征。基于图的不完备多视图模型旨在从缺失的多视图数据中建立相似图矩阵并得到多个相似图矩阵之间的一致图特征。相较于基于矩阵分解的模型,基于多核和图的模型可以捕获样本之间的几何结构。此外,有研究人员基于张量奇异值分解并结合多秩张量的高阶相关性提出了一系列不完备多视图聚类模型[7-8]。

以上模型都是基于浅层方法的模型,近年来,由于深度学习模型在提取深层特征中的表现非常出色,因此受到了许多研究人员的青睐。为了进一步提升不完备多视图聚类的表现,有研究人员将深度模型运用于不完备多视图聚类上。例如,有研究人员将自编码器[9]和生成对抗网络(GAN)[10]结合在一起以解决缺失数据带来的负面影响。该模型将两种深度学习模型集成在一起,一方面,自编码器提取深层特征来提升生成对抗网络的训练效果;另一方面,生成对抗网络对缺失数据进行预测与填充来促进自编码器提取深层特征的能力。此外,有研究人员从信息论的角度出发,结合对比学习提出了一种一致表示学习和跨视图数据恢复的理论框架[11]。也有研究人员将单视图的深度聚类模型迁移到了多视图数据中并以此来解决缺失多视图数据情况下的聚类,将"自步学习"的概念引入不完备多视图聚类中,与自编码器相结合,提出了认知不完备多视图聚类网络[12](Cognitive deep incomplete multi-view clustering network,CDIMC-net),该聚类模型提供了一种基于人类认知的"由易到难"学习策略,增强了深度聚类模型的鲁棒性和聚类表现。然而在CDIMC-net中,作者将缺失视图样本做了"丢弃"处理,将缺失部分以"0"值填充,这样一来该模型在训练时仅利用了未缺失的数据,不仅造成样本数量的减少,同时在特征提取方面仅利用存在的视图,不利于聚类网络的训练。

在CDIMC-net中,作者考虑到了数据之间的几何关系,在训练过程中将样本与样本之间的邻居关系作为约束项加入损失函数,目的是在深度模型在训练中获得更具有区分度的深层特征。受此启发,本文将样本与样本之间的邻居关系用于处理缺失视图的填充,现有的缺失视图处理方法中,通常使用"0"值填充或者均值填充的方法。"0"值填充相当于放弃了那些缺失数据,造成训练时样本量减少。均值填充的方法虽然补全了残缺视图,但是由所有样本计算得到的均值与残缺视图样本值差距较大,在训练时无法保证得到有效的信息。受到邻居关系的启发,在考虑到样本之间的邻居图结构以及多视图

数据之间的互补性与一致性，在此基础上，本文使用基于多视图特性的最近邻填充方法对缺失视图进行填充处理，对CDIMC-net进行改进，提出了改进的自步深度不完备多视图聚类(Improved self-paced deep incomplete multi-view clustering, ISPDIMC)。本模型将基于多视图特性的最近邻填充方法、多视图的自编码器、多视图数据的一致性潜在特征、图嵌入策略和基于人类认知的“自步学习”K-means聚类网络这些策略和方法集成到一个深度模型中。合理运用以上策略提升不完备多视图数据的聚类表现。

本文主要贡献如下：

(1) 提出了一种新颖深度不完备多视图聚类模型，可以灵活处理各种数据缺失的情况，提升了不完备多视图聚类的鲁棒性与表现效果。

(2) 提出的深度不完备多视图聚类模型中使用基于多视图特性的最近邻填充方法，充分利用了各个视图潜在的互补信息，让缺失视图的填充值与原样本值更加相近，并用实验证明了这种基于多视图特性的最近邻填充方案的合理性。

1 相关工作

1.1 K-means聚类算法

K-means[13]是聚类算法中的经典算法，其算法的主要思想如下：对于给定的数据集$\{\boldsymbol{X}\}\in\mathbf{R}^{m\times n}$，其中包含$n$个样本，每个样本的特征数为$m$，以欧式距离为度量比较样本之间的近似程度，依照样本之间的近似程度将数据集划分以K个聚类簇中心$\{\boldsymbol{C}\}\in\mathbf{R}^{m\times k}$为基点的$K$个簇，让属于同一个簇的样本尽量紧密的围绕在簇心周围，而不同簇之间的距离应尽量的大，表示为

$$\min_{C,S}\left\|\boldsymbol{X}-\boldsymbol{CS}\right\|_{\mathrm{F}}^{2}\quad \text{s.t.}\ \boldsymbol{S}\in\{0,1\}^{k\times n},\boldsymbol{S}^{\mathrm{T}}\boldsymbol{1}=1 \tag{1}$$

式中：$\boldsymbol{S}$表示指示矩阵，其中$s_{i,j}=1$表示编号为i的样本属于第j个簇；$\boldsymbol{1}$表示一个由全1元素组成的向量。

1.2 自编码器

自编码器是一种经典的神经网络模型，通常运用在自监督学习中。其基本思想是通过神经网络不断进行自监督训练让输出的重构样本接近输入的样本。本文中使用的是自编码器是欠完备自编码器，它的内部包含3个部分：编码器$f_{\mathrm{EN}}()$、隐藏层z和解码器$f_{\mathrm{DE}}()$。其中编码器$z=f_{\mathrm{EN}}(x)$对输入的样本进行处理，之后将处理后样本输入隐藏层z，隐藏层的维度通常比输入样本的维度要小，之后将隐藏层的特征通过解码器$\bar{x}=f_{\mathrm{DE}}(z)$进行输出得到重构样本，通过这样先降维后升维的训练方法，迫使自编码器的隐藏层中要学习到输入样本最具有区分性的特征，其过程可以表述为

$$L=\left\|\boldsymbol{x}-\boldsymbol{f}_{\mathrm{DE}}(\boldsymbol{f}_{\mathrm{EN}}(x))\right\|_{2}^{2} \tag{2}$$

1.3 自步学习

“自步学习”[14]用于模拟人类学习新知识的过程：从简单到困难。当给出一些新任务时，人类往往首先选择最简单的任务。当对任务的认知有所提升后，处理难度更高一层的任务，在不断加强任务难度的同时获得更多的知识。在这个不断学习的策略下，可能会获得关于这个任务的所有知识。这种基于人类认知的学习策略被认为是一种有效的训练方式。因为该策略的优异表现，“自步学习”已经被广泛应用于机器学习任务中，如人重识别[15]、分类[16]以及人脸识别[17]。本文在聚类阶段使用“自步学习”策略进行训练，将那些分布于簇中心周围的样本被认为是“简单任务”，而将那些分布于簇边界的样本被认为是“困难任务”。

2 本文方法

本文提出的ISPDIMC模型中首先利用多视图特性选择有效的数据对缺失数据进行填充,之后进行预训练和微调两个阶段,其中预训练阶段为多视图自编码器的自监督训练,并在其中加入了一致潜在特征融合层,目的是寻找到有效的隐藏层特征。而微调阶段则是运用包含“自步学习”的聚类策略,对隐藏层特征进行微调和聚类,使得到的隐藏层特征更适用于不完备多视图聚类任务。

2.1 相关定义

给定一个包含 l 个视角的数据集 $\{\boldsymbol{X}\}^{(v)}\in\mathbf{R}^{m_v\times n}$,其中 m_v 表示各个视图的特征维度,n 表示样本个数。在该数据集中,缺失的样本统一用“NaN(Not a number)”来表示。为了表示各个视图上样本的缺失情况,在每一个视图上设置一个对角矩阵 $W^{(v)}$。其中 $W_{i,i}^{(v)}=1$ 表示该视图上的样本是存在的。与多视图聚类的目的一致,不完备多视图聚类的目的是将 n 个样本划分为 c 个簇。

2.2 预训练阶段

基于多视图特性的最近邻填充。根据多视图的一致性与互补性原则。如果两个样本在某一视图上是邻居关系,那么它们在另一视图上应仍保持邻居关系。基于这一假设,在每个缺失视图上都构造出未缺失样本之间的近邻图,如果在视图 i 上的样本是缺失的,那么将按照邻居的顺序,从第一邻居开始,依次搜索未缺失视图相同序号的邻居,使用邻居顺序和视图顺序最靠前且在视图 i 上未缺失的样本对缺失样本进行填充。

多视图自编码器与潜在特征融合。对于每一个视图的数据,分别使用自编码器进行潜在特征的学习。经过编码器后,可以得到各个视图的潜在特征 $\{h^1,h^2,h^3,\cdots,h^l\}$。在本文中,对各个视图的潜在特征进行融合获取不同视图之间的一致潜在特征表示

$$\boldsymbol{h}^*=\sum_{v=1}^{l}\boldsymbol{h}^{(v)}/l \tag{3}$$

式中:$\boldsymbol{h}^*$ 表示一致潜在特征;l 表示多视图数据集的视图数量。在得到了一致潜在特征之后,将一致潜在特征输入解码器中。利用解码器输出的重构样本和编码器的输入样本进行自监督学习。

图嵌入策略。受到来自子空间学习领域的假设启发[18]:如果高维空间中的一对样本距离很接近,那么从中提取的低维潜在特征应该在低维空间中也很接近。为此,通过如下策略建立邻居关系图

$$\min\frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{h}_i^*-\boldsymbol{h}_j^*\right\|_2^2\boldsymbol{G}_{i,j} \tag{4}$$

式中 $\boldsymbol{G}\in\mathbf{R}^{n\times n}$ 表示最近邻图关系的矩阵。图矩阵定义如下

$$\boldsymbol{G}_{i,j}=\begin{cases}1 & x_i\in\Delta(x_j)\text{或}x_i\in\Delta(x_j)\\ 0 & \text{其他}\end{cases} \tag{5}$$

式中 $\Delta(x_i)$ 表示欧氏距离度量下样本 x_i 的最近邻。在后续的实验中,将此图嵌入策略应用于预训练阶段和微调阶段。

目标函数。在预训练阶段,ISPDIMC将自编码器的重构样本误差和图嵌入策略相结合,构成预训练阶段的目标函数

$$L=\min\frac{1}{n}\sum_{v=1}^{l}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i^v-\boldsymbol{f}_{\mathrm{DE}}^v(h^*)\right\|_2^2+\alpha\frac{1}{2nl}\sum_{v=1}^{l}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\|\boldsymbol{h}_i^*-\boldsymbol{h}_j^*\right\|_2^2\boldsymbol{G}_{i,j}^{(v)} \tag{6}$$

式中:$\boldsymbol{h}^*$ 由式(3)计算得到;α 为图嵌入策略的超参数。

图1展示了预训练阶段的网络结构。

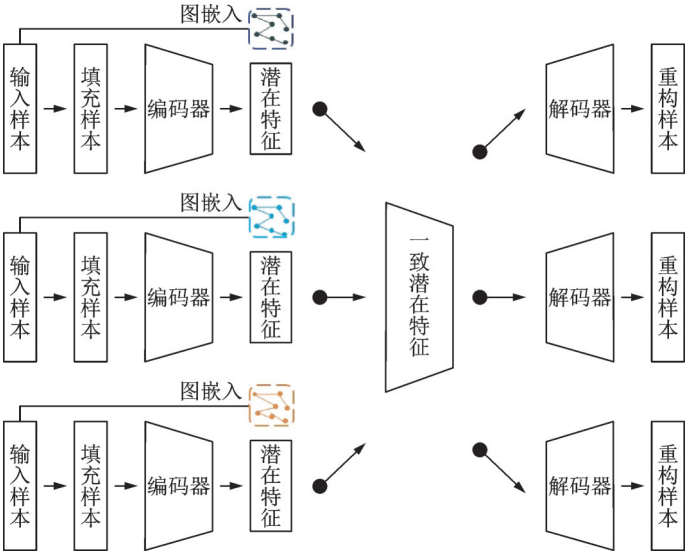

图1 预训练阶段示意图

Fig.1 Illustration of pre-training phase

2.3 微调与聚类阶段

通过预训练阶段后，自编码的隐藏层可以获得有效的潜在特征，但是这些潜在特征可能并不适用于不完备多视图聚类任务。对此，受到相关深度聚类算法的启发，在微调阶段将预训练阶段训练好的自编码器的解码器部分放弃，将编码器部分与聚类层相连，利用一致潜在特征进行聚类任务，而聚类的结果又将反馈回指导编码器的训练。

对一致潜在特征进行微调，使其能更好地适应不完备多视图聚类任务。本文在微调阶段采用“自步学习”K-means策略，对隐藏特征进行优化与聚类。

自步学习K-means策略。受到人类认知假设的启发，遵循“由简单到困难”的原则，在聚类时一开始不使用全部样本进行聚类，而是选择靠近聚类簇中心的一部分样本进行聚类，在训练时不断增加样本数量，通过这样的策略，尽可能减少那些分布在聚类簇边缘的样本的负面影响，自步学习K-means策略的目标函数如下[19]

$$\begin{cases}\min\limits_{S,r,\lambda}\dfrac{1}{nk}\sum\limits_{i=1}^{n}\left(r_i\left\|\boldsymbol{h}_i^{*}-\boldsymbol{C}\boldsymbol{S}_{:,i}\right\|_2^2-\lambda r_i\right)\\ \text{s.t. } r_i\in\{0,1\},\boldsymbol{S}\in\{0,1\}^{k\times n},\boldsymbol{S}^{\mathrm{T}}\boldsymbol{1}=1\end{cases}\tag{7}$$

式中：$\boldsymbol{C}$表示聚类簇中心矩阵；$\boldsymbol{S}$表示聚类结果指示矩阵。与传统的K-means算法不同，自步学习K-means策略在优化时固定了聚类簇中心，这一做法主要时为了避免在训练时所有样本收敛到一起。r_i表示各个样本的权重$r_i\in\{r_1,r_2,r_3,\cdots,r_n\}\in\mathbf{R}^n$，$\lambda$表示从所有样本选取样本所占的比重。

一般来说，为了达到逐渐增加训练样本进入训练的目的，λ应伴随训练次数的增加而增大，但不同训练集，样本数量的不同，较难使用统一的数值来确定λ的成长速度。在自步学习K-means策略中，采用如下方法来自适应调整与增加λ的值

$$\lambda=\mu(L^t)+t\sigma(L^t)/T\tag{8}$$

式中：L^t表示在第t此迭代中的损失；T表示总迭代次数；$\mu(L^t)$和$\sigma(L^t)$分别表示损失的平均值和标准差。经过式(8)的定义，λ可以灵活地适用于不同数据集与任务。

目标函数。在微调阶段，ISPDIMC将自步学习K-means策略误差和图嵌入策略相结合，构成微调阶段的目标函数

$$\begin{cases} L=\min\limits_{S,r,\lambda}\dfrac{1}{nk}\sum\limits_{i=1}^{n}\left(r_i\left\|\boldsymbol{h}_i^*-\boldsymbol{C}\boldsymbol{S}_{:,i}\right\|_2^2-\lambda r_i\right)+\alpha\dfrac{1}{2nl}\sum\limits_{v=1}^{l}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\left\|\boldsymbol{h}_i^*-\boldsymbol{h}_j^*\right\|_2^2\boldsymbol{G}_{i,j}^{(v)} \\ \text{s.t. } r_i\in\{0,1\},\boldsymbol{S}\in\{0,1\}^{k\times n},\boldsymbol{S}^{\mathrm{T}}\boldsymbol{1}=1 \end{cases} \tag{9}$$

2.4 参数更新与优化

与大多数自编码器的训练方法一样,预训练阶段可以直接用随机梯度下降法(SGD)和反向传播进行优化。本节中重点关注微调阶段的参数更新与优化,交替优化式(9)中的聚类指示矩阵$\boldsymbol{S}$,各样本权重r以及编码器的参数。

固定编码器参数调整$\boldsymbol{S}$。当编码器参数固定时,输入样本经由编码器得到的隐藏层一致特征$\boldsymbol{h}^*$是确定的。在聚类簇中心固定的情况下,得到聚类结果$\boldsymbol{S}$为

$$S_{i,j}=\begin{cases}1 & j=\arg\min\limits_{c}\left\|\boldsymbol{h}_i^*-\boldsymbol{C}_{:,c}\right\|_2^2 \\ 0 & \text{其他}\end{cases} \tag{10}$$

固定编码器参数和$\boldsymbol{S}$调整r。当编码器参数和聚类结果$\boldsymbol{S}$都固定时,根据每一个样本与其聚类中心的距离大小来调整样本的权重

$$r_i=\begin{cases}1 & \left\|\boldsymbol{h}_i^*-\boldsymbol{C}\boldsymbol{S}_{:,i}\right\|_2^2\leqslant\lambda \\ 0 & \text{其他}\end{cases} \tag{11}$$

式中λ由式(8)计算得到。

固定$\boldsymbol{S}$和r调整编码器参数。聚类结果$\boldsymbol{S}$和权重参数r固定时,对于编码器来说就变成了有监督训练,可以依靠SGD和反向传播算法对各个视图的编码器参数进行更新。

停止训练。为了更好地判断训练是否已经收敛,规定当两次迭代之间的聚类结果$\boldsymbol{S}$变化小于阈值δ停止训练。计算标准如下

$$1-\frac{1}{n}\sum_{i,j}S^tS^{t-1}<\delta \tag{12}$$

式中:S^t和S^{t-1}是在第t次迭代和第$t-1$次迭代中预测的聚类结果,在本文实验中将δ设置为1e−7。

图2展示了微调阶段的网络结构。

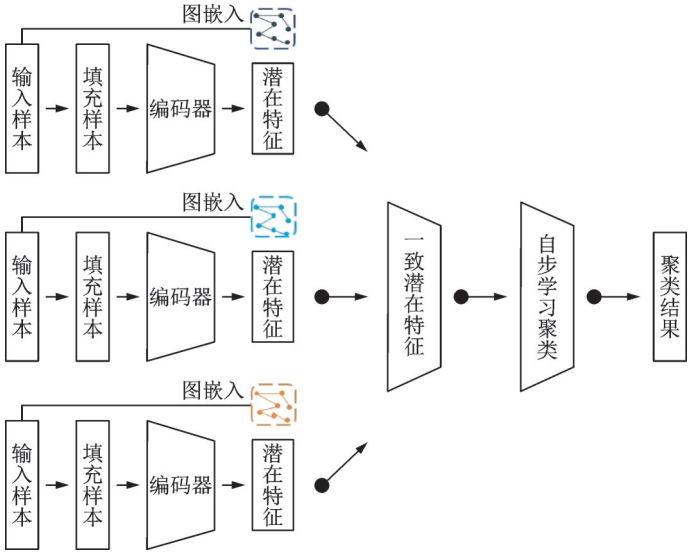

图2 微调与聚类阶段示意图

Fig.2 Illustration of fine-tunning and clustering phase

2.5 模型预处理

一般来说,深度学习模型在训练时对数据进行分批处理,本文提出的方法中在训练时不仅要加载数据,而且还要加载由数据计算出的最近邻图矩阵,这样一来,传统的批处理方法将会破坏最近邻图矩阵中携带的数据几何结构信息。为了将批处理中的每个最近邻矩阵的子矩阵能够携带尽可能多的几何结构信息,对数据集进行如下预处理:

(1)基于多视图特性的最近邻填充。通过每个视图的对角矩阵 $\boldsymbol{W}^{(v)}$ 得到每个视图的缺失样本信息。对每个视图的未缺失样本分别计算它到全体未缺失样本的欧式距离,得到距离矩阵后将其排序,这样就能得到每个未缺失样本的邻居关系。如果在视图 i 上的样本是缺失的,那么将按邻居的顺序,从第一邻居、第二邻居,……,依次搜索未缺失视图上相同序号的邻居,将邻居顺序以及视图顺序最靠前且在视图 i 上未缺失的样本对缺失样本进行填充。

(2)数据从新排列与最近邻计算。来自同一聚类簇的样本更有可能是邻居。出于这一假设,先对所有样本进行一次K-means算法,利用K-means计算出的预测标签对样本进行重新排序,在排序后的样本上计算最近邻图。这样一来,在批处理时每一个最近邻子图上将会携带更多的样本间几何关系信息。其步骤为:①将所有填充后的视图的特征拼接到一个视图中;②在拼接视图上执行K-means算法;③根据聚类结果对数据重新排序,将分到同一聚类簇中的样本放在一起;④根据式(5)从重新排序的数据构造最近邻图。经过最近邻填充之后,被填充后的值在没有图嵌入策略的限制下在它们的低维特征也可以很接近,为了最大化图嵌入策略的作用,在对角矩阵 $\boldsymbol{W}^{(v)}$ 的指导下,选择计算每个视图未缺失样本之间的最近邻图,以此来达到更好的训练效果。

经过重新排列之后的数据集定义为 $\{\boldsymbol{Y}\}^{(v)}\in\mathbf{R}^{m_v\times n}$。同时计算新的缺失指示对角矩阵 $\bar{\boldsymbol{W}}^{(v)}$ 以及最近邻图 $\boldsymbol{G}_{i,j}^{(v)}$。

2.6 模型训练

预训练。在预训练阶段,使用SGD对重新排列后的样本进行预训练处理。预训练阶段的批处理表示如下

$$\min\sum_{v=1}^{l}\frac{1}{m_v b_s}\left\|\boldsymbol{Y}_{\text{batch}}^{(v)}-\bar{\boldsymbol{Y}}_{\text{batch}}^{(v)}\right\|_{\text{F}}^{2}+\alpha\frac{1}{b_s l}\sum_{v=1}^{l}\text{Tr}\left(\boldsymbol{H}_{\text{batch}}^{(v)}\boldsymbol{L}_{N_{\text{batch}}^{(v)}}\boldsymbol{H}_{\text{batch}}^{(v)\text{T}}\right)\tag{13}$$

式中:$\boldsymbol{Y}_{\text{batch}}^{(v)}$ 表示每一次批处理的数据;$\bar{\boldsymbol{Y}}_{\text{batch}}^{(v)}$ 表示编码器输出的重构数据;$\boldsymbol{H}_{\text{batch}}^{(v)}$ 表示隐藏特征;$\boldsymbol{L}_{N_{\text{batch}}^{(v)}}$ 为批处理数据对应的拉普拉斯矩阵;b_s 表示一次批处理样本的数量。

微调。对于微调阶段,将该阶段的详细步骤记录在算法1中。

算法1 ISPDIMC的微调阶段算法

输入:重新排序后数据集 $\{\boldsymbol{Y}\}^{(v)}\in\mathbf{R}^{m_v\times n}$,缺失指示对角矩阵 $\bar{\boldsymbol{W}}^{(v)}$,最近邻图 $\boldsymbol{G}_{i,j}^{(v)}$,参数 α,最大迭代次数 T,内训练最大迭代次数Maxiter,批处理大小 b_s,停止训练阈值 δ;

输出:聚类指示矩阵 $\boldsymbol{S}$。

① 预训练:将 $\{\boldsymbol{Y}\}^{(v)}\in\mathbf{R}^{m_v\times n}$,$\bar{\boldsymbol{W}}^{(v)}$,$\boldsymbol{G}_{i,j}^{(v)}$ 传入预训练网络得到同一隐藏特征 H^{*},在统一潜在特征上进行K-means算法得到聚类中心矩阵 $\boldsymbol{C}$ 和初始聚类分布矩阵 $\boldsymbol{S}$,将所有样本的权重向量 $\boldsymbol{r}$ 置为1。

② for t in $\{1,2,\cdots,T\}$ do

③ for j in $\{1,2,\cdots,\text{Maxiter}\}$ do

④ 批处理更新编码器参数

⑤ end for

⑥ 如式(10)、(11)和(8)更新参数

⑦ if 满足式(12) then
⑧ 停止训练
⑨ end if
⑩ end for
⑪ return $\boldsymbol{S}$

3 实验验证

3.1 数据集

本文提出的ISPDIMC方法将在表1列出的3个数据集上进行评估:

(1) Handwritten[20]。该数据集为手写数字图片数据集,其中包含10个数字(即0~9)的2 000个样本,其中包含5个视图特征,分别是傅里叶系数特征、轮廓特征、Karhunen-Love系数特征、Zernike矩特征和像素均值特征。

(2) BDGP[21]。该数据集为伯克利果蝇基因组计划基因表达模式数据库,该数组集包含5个类别,其中每个类别有500个样本,共2 500个果蝇胚胎样本。每个样本包含4个视图特征,即纹理特征和从侧面、背部和腹部图像中提取的3种视觉特征。

(3) MNIST[22]。该数据集为知名的手写数字数据集。其中包含了手写数字0~9共4 000个样本。每个样本包含2个视图特征,分别是像素特征和边缘特征。

表1 数据集信息

Table 1 Information of datasets

数据集	样本数	类别数	视图数	特征数
Handwritten	2 000	10	5	76/217/64/240/47
BDGP	2 500	5	4	79/1 000/500/250
MNIST	4 000	10	2	784/784

3.2 不完备多视图数据集构造

由于以上3个数据集都是完整数据集,为了方便进行实验与对比,采用以下方法来构造不完备多视图数据集:对于拥有两个以上视图的数据集,在所有样本至少保留一个视图的条件下,从每个视图中分别随机删除10%、30%、50%的样本。对于MNIST数据集,保留10%、30%、50%的样本作为视图完整的成对样本,余下样本作为单视图样本处理,即余下样本中的一半样本只有第一视图,另一半样本只有第二视图。

3.3 对比方法与参数设置

用来对比的方法包括2类。第1类为浅层模型中3个具有代表性算法:双对齐不完备多视图聚类(DAIMC)[23]、单趟不完整多视图聚类(OPIMC)[24]和在线不完整多视图聚类(OMVC)[25]。第2类为深层模型:基于一致性生成对抗网络的缺失视图聚类(PVC-GAN)[10]、认知深度不完备多视图聚类网络(CDIMC-net)以及在预训练网络得到的一致潜在特征上运用K-means算法(K-means on consistency hidden features,KoCHF)。其中OMVC对缺失值使用均值填充,DAIMC、OPIMC和CDIMC-net对缺失值使用"0"值填充。PVC-GAN只能运用在两个视图的数据上,在实验中仅在MNIST数据集上进行对比。在参数设置方面,编码器部分由全连接神经网络构成,每层的神经元数量为[Input, 0.8Input, 0.8Input, 1 500, 10],其中Input为输入样本的维度。解码器部分是编码器部分的镜像,每层神经元个数为[10, 1 500, 0.8Input, 0.8Input, Input]。每一层之间的激活函数都使用非线性激活函数"ReLU"。在预训练阶段,采用"SGD"优化器对网络进行优化。在微调阶段,采用"Adam"优化器对网络进行优化。

3.4 评价指标

本文采用聚类精度(Acc)和归一化互信息(NMI)这两种常用的聚类评价指标对不同方法进行评估[12]。

聚类精度的计算公式为

$$\mathrm{Acc}(t,p)=\frac{\sum_{i=1}^{n}\delta(p,g(t))}{n} \tag{14}$$

式中:t表示样本所表示的真实类标签;p表示经过聚类算法后得到预测类标签;函数g表示真实类标签与预测类标签的映射关系。函数$\delta()$表示二值指示函数,当p等于$g(t)$时函数值为1,p不等于$g(t)$时函数值为0[26]。

归一化互信息度量的计算公式为

$$\mathrm{NMI}=\frac{2\mathrm{MI}(t,p)}{H(t)+H(p)} \tag{15}$$

式中:MI表示互信息;H表示信息熵[27]。

以上两个评价指标的取值皆为区间[0,1],指标越高代表该算法的聚类效果更佳优秀。

3.5 实验结果与分析

表2展示了各算法在3个数据集上不同缺失率和成对率在评价指标Acc和NMI的对比结果,其中粗体数字标志出了最优结果,下划线数字标志出了次优结果。

表2 不同方法在3个数据集的Acc和NMI对比

Table 2 Comparison of Acc and NMI on three datasets

数据集	方法	Acc/%			NMI/%		
		缺失率10%	缺失率30%	缺失率50%	缺失率10%	缺失率30%	缺失率50%
Handwritten	DAIMC	78.05	74.45	72.90	71.29	69.94	61.19
	OPIMC	63.35	68.75	64.25	64.48	58.60	48.06
	OMVC	69.75	60.00	51.31	60.97	50.88	40.02
	CDIMC-net	93.50	**95.45**	82.91	88.73	**90.30**	82.90
	KoCHF	88.75	91.35	88.80	81.78	84.29	79.76
	ISPDIMC	**95.90**	95.00	**94.63**	**92.10**	90.16	**89.49**
BDGP	DAIMC	43.80	39.76	32.12	20.45	19.08	6.51
	OPIMC	69.24	75.40	62.21	63.70	53.47	36.66
	OMVC	54.16	47.00	37.88	29.46	24.12	11.75
	CDIMC-net	87.32	76.82	54.59	76.36	61.26	34.72
	KoCHF	85.72	85.69	64.20	70.21	66.81	39.44
	ISPDIMC	**88.68**	**86.66**	**66.58**	**76.73**	**71.04**	**44.62**
数据集	方法	Acc/%			NMI/%		
		成对率10%	成对率30%	成对率50%	成对率10%	成对率30%	成对率50%
MNIST	DAIMC	44.30	45.46	45.78	32.07	35.48	36.48
	OPIMC	41.80	45.86	46.08	34.99	38.27	40.59
	OMVC	40.90	41.88	46.73	32.62	33.22	35.89
	PVC-GAN	45.17	48.36	52.80	39.33	43.22	49.61
	CDIMC-net	56.32	56.45	54.55	54.43	56.63	57.00
	KoCHF	54.70	60.22	55.37	49.69	52.44	50.16
	ISPDIMC	**60.88**	**62.28**	**60.97**	**58.99**	**59.44**	**59.64**

由表2可知:(1)运用了深度模型的不完备多视图聚类方法在不同的数据集和缺失率上的聚类评价指标结果都是最优或次优,这表明了深度聚类模型对比传统浅层模型在聚类表现上有巨大提升;(2)随着视图缺失率的增加,所有对比方法的聚类评价指标都出现下降的情况,这表明视图的缺失对多视图聚类带来了负面影响;(3)对于使用"0"值或均值填充的方法来说,ISPDIMC使用基于多视图特性的最近邻填充方法,有效减少了样本缺失带来的负面影响,特别是在Handwritten和BDGP数据集上,当缺失值为50%时,对比其他方法,ISPDIMC的聚类指标下降的最少;(4)在多个实验中,ISPDIMC和Ko-CHF得到的实验结果超越了CDIMC-net,这表明了运用基于多视图特性的最近邻填充方案对提升缺失多视图聚类的性能是有帮助的。

3.6 超参数分析

为了探究超参数的设置对聚类结果的影响,在Handwritten和BDGP数据集上,在缺失率为10%的情况下,对微调阶段的图嵌入系数α和学习率进行调整,观察上述2个超参数对聚类结果的影响。具体地,将图嵌入系数α和学习率依次设置为[1e−3, 5e−4, 1e−4, 5e−5, 1e−5],之后将它们的聚类指标Acc和NMI进行对比,结果如图3所示。

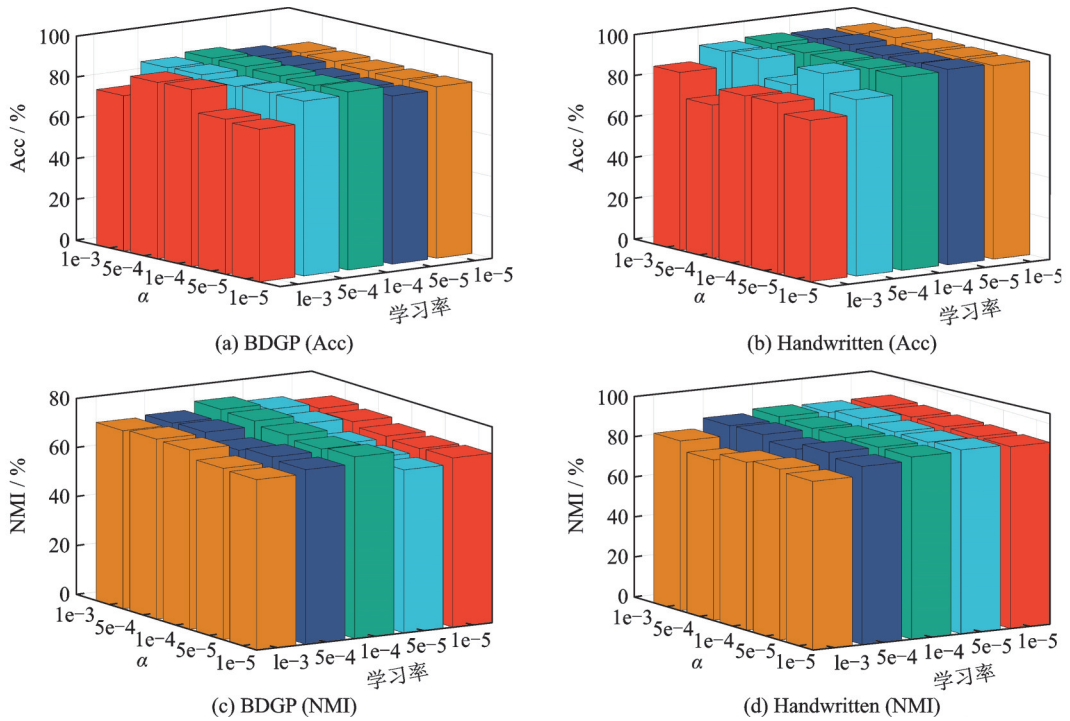

图3 不同图嵌入系数和学习率下的Acc和NMI

Fig.3 Acc and NMI under different graph embedding coefficients and learning rates

可以看出,学习率设置为[1e−3, 5e−4]时,Acc和NMI低于将学习率设置为[1e−4, 5e−5, 1e−5]的表现。同时,将图嵌入系数α设置为[1e−4, 5e−5, 1e−5]时表现更优秀。综合2个数据集上Acc和NMI的比较,建议将图嵌入系数α和学习率设置为[1e−4, 5e−5, 1e−5]。

3.7 消融实验

本文提出的模型中综合运用了多种策略,为了探究每一种策略对聚类效果的影响,将模型退化为

缺少预训练的情况、缺少自步学习的情况以及缺少图嵌入的情况，在Handwritten和BDGP数据集上进行不同样本缺失率下的实验。结果如图4所示。由图4可以看出，本文提出的模型表现最佳，缺少了预训练的模型表现最差，这表明了预训练在本模型中十分重要。

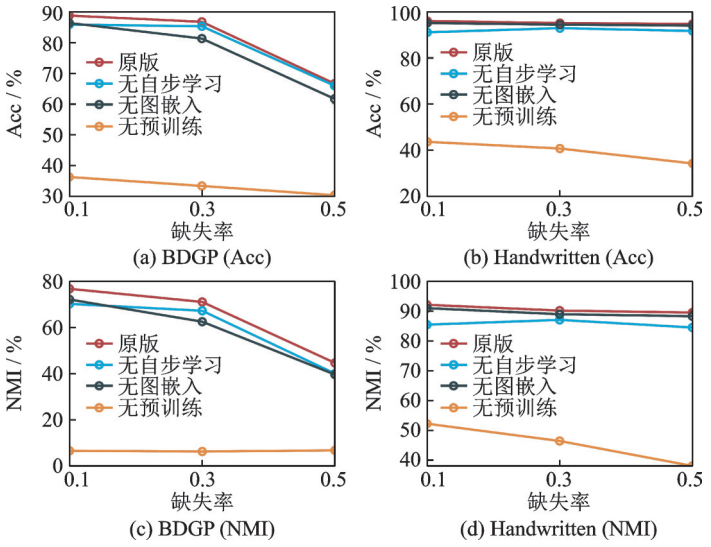

图4 不同策略对聚类效果的影响

Fig.4 Influence of different strategies on clustering

3.8 训练收敛性分析

图5为ISPDIMC在Handwritten和BDGP数据集上不同缺失率的情况下，在微调阶段的损失随迭代次数的变化情况。由图5可看出，损失在最初的几轮迭代训练中快速下降，然后下降速度放缓，最后趋于稳定。

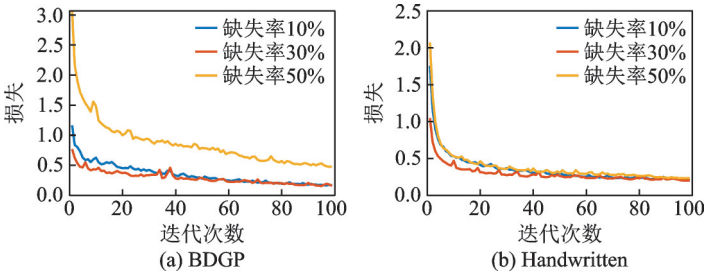

图5 损失随迭代次数的变化情况

Fig.5 Variation of loss with the number of epochs

图6为ISPDIMC在Handwritten和BDGP数据集上缺失率为10%的情况下，评价指标Acc和NMI随迭代次数的变化情况。由图6可看出，总体上Acc和NMI在开始几轮的迭代训练中持续上升，最后逐渐趋于稳定的状态。

由以上的训练收敛性分析可知，ISPDIMC在不同缺失率的情况下都可以在经过训练之后达到一个较为理想的稳定状态。

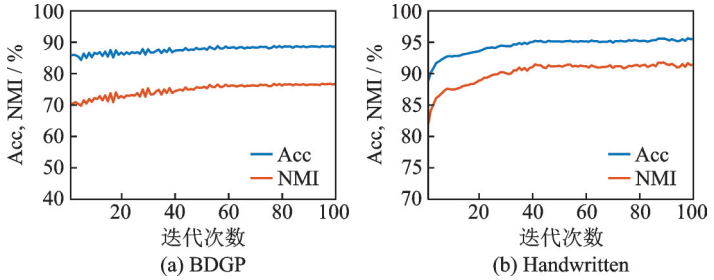

图6 Acc和NMI随迭代次数的变化情况

Fig.6 Variation of Acc and NMI with the number of epochs

4 结束语

本文提出改进的自步深度不完备多视图聚类(ISPDIMC),使用基于多视图特性的最近邻填充策略将缺失多视图数据以合理的数值进行填充之后,使用自步学习策略对多视图数据进行聚类分析。ISPDIMC包含两个阶段:第1阶段通过基于多视图特性的最近邻填充将缺失视图样本进行补全,之后以自监督的方式进预训练,运用合理方案将多视图潜在特征融合后获取一致潜在特征;第2阶段通过自步学习K-means聚类方案对数据进行聚类分析。同时,在上述两个阶段都加入图嵌入策略来保持样本之间的几何关系。在3个公开的数据集上进行实验,表明了ISPDIMC以更为合理的数据填充方式减小了缺失样本带来的负面影响,同时也提升了聚类与泛化性能。消融实验表明预训练阶段直接影响微调与聚类阶段的表现,今后将会探究更有效的缺失值填充方案与特征融合机制,用于减少缺失视图带来的负面影响,挖掘多视图数据之间更深层次的潜在互补信息,以更加鲁棒的一致潜在特征进行聚类,探索更为先进的不完备多视图深度聚类算法。

参考文献:

[1] WEN Jie, ZHANG Zheng, XU Yong, et al. Unified embedding alignment with missing views inferring for incomplete multi-view clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 5393-5400.

[2] LI Shaoyuan, JIANG Yuan, ZHOU Zhihua. Partial multi-view clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2014: 1968-1974.

[3] RAI N, NEGI S, CHAUDHURY S, et al. Partial multi-view clustering using graph regularized NMF[C]//Proceedings of the International Conference on Pattern Recognition. Piscataway, NJ: IEEE, 2016: 2192-2197.

[4] LIU Xinwang, ZHU Xinzhong, LI Miaomiao, et al. Multiple kernel k-means with incomplete kernels[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1191-1204.

[5] LIU Xinwang, LI Miaomiao, TANG Chang, et al. Efficient and effective regularized incomplete multi-view clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(8): 2634-2646.

[6] WEN Jie, XU Yong, LIU Hong. Incomplete multi-view spectral clustering with adaptive graph learning[J]. IEEE Trans on Cybernetics, 2020, 50(4): 1418-1429.

[7] XIE Yuan, TAO Dacheng, ZHANG Wensheng, et al. On unifying multi-view self-representations for clustering by tensor multi-rank minimization[J]. International Journal of Computer Vision, 2018, 126(11): 1157-1179.

[8] 赵博宇, 张长青, 陈蕾, 等. 生成式不完整多视图数据聚类[J]. 自动化学报, 2021, 47(8): 1867-1875.
ZHAO Boyu, ZHANG Changqing, CHEN Lei, et al. Generative model for partial multi-view clustering[J]. Acta Automatica Sinica, 2021, 47(8): 1867-1875.

[9] KUSNER M J, BROOKS P, JOSÉ M H L. Grammar variational autoencoder[C]// Proceedings of the International Conference on Machine Learning. New York: ACM, 2017: 1945-1954.

[10] WANG Qianqian, DING Zhengming, TAO Zhiqiang, et al. Partial multi-view clustering via consistent GAN[C]//Proceedings of the International Conference on Data Mining(ICDM). Piscataway, NJ: IEEE, 2018: 1290-1295.

[11] LIN Yijie, GOU Yuanbiao, LIU Zitao, et al. COMPLETER: Incomplete multi-view clustering via contrastive prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 11174-11183.

[12] WEN Jie, ZHANG Zheng, XU Yong, et al. CDIMC-Net: Cognitive deep incomplete multi-view clustering network[C]//Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2020: 3230-3236.

[13] JAIN A K. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[14] KUMAR M, PACKER B, KOLLER D. Self-paced learning for latent variable models[C]//Proceedings of Neural Information Processing System 23 (NIPS2010). MA: MIT Press,2010: 1189-1197.

[15] ZHOU Sanping, WANG Jinjun, MENG Deyu, et al. Deep self-paced learning for person re-identification[J]. Pattern Recognition, 2018, 76: 739-751.

[16] REN Yazhou, ZHAO Peng, SHENG Yongpan, et al. Robust softmax regression for multi-class classification with self-paced learning[C]//Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2017: 2641-2647.

[17] LIN Liang, WANG Keze, MENG Deyu, et al. Active self-paced learning for cost-effective and progressive face identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(1): 7-19.

[18] ZHAO Kang, PENG Chong, CHENG Qiang. Clustering with adaptive manifold structure learning[C]//Proceedings of the International Conference on Data Engineering. Piscataway, NJ: IEEE, 2017: 79-82.

[19] GUO Xifeng, LIU Xinwang, ZHU En, et al. Adaptive self-paced deep clustering with data augmentation[J]. IEEE Trans on Knowledge and Data Engineering, 2019, 32(9): 1680-1693.

[20] ARTHUR A, DAVID N. UCI machine learning repository[EB/OL]. [2019-12-28]. http://archive.ics.uci.edu/ml.

[21] CAI Xiao, WANG Hua, HUANG Heng, et al. Joint stage recognition and anatomical annotation of drosophila gene expression patterns[J]. Bioinformatics, 2012, 28(12): i16-i24.

[22] LECUN Y. The mnist database of handwritten digits[EB/OL]. [2016-12-29]. http://yann. lecun. com/exdb/mnist/.

[23] HU Menglei, CHEN Songcan. Doubly aligned incomplete multi-view clustering[C]//Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann, 2018: 2262-2268.

[24] HU Menglei, CHEN Songcan. One-pass incomplete multi-view clustering[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI, 2019: 3838-3845.

[25] SHAO Weixiang, HE Lifang, LU Chunta, et al. Online multi-view clustering with incomplete views[C]//Proceedings of the IEEE International Conference on Big Data(ICBD). Piscataway, NJ: IEEE, 2016: 1012-1017.

[26] 文杰. 图嵌入聚类模型研究[D]. 哈尔滨: 哈尔滨工业大学, 2019.
WEN Jie. Research on graph embedded clustering models[D]. Harbin: Harbin Institute of Technology, 2019.

[27] 郭西风. 基于深度神经网络的图像聚类算法研究[D]. 长沙: 国防科技大学, 2020.
GUO Xifeng. A study on image clustering algorithms with deep neural networks[D]. Changsha: National University of Defense Technology, 2020.

作者简介:

崔金荣(1984-),女,博士,讲师,研究方向:模式识别、机器学习、图像处理、深度学习等,E-mail: tweety1028@163.com。

黄诚(1994-),**通信作者**,男,硕士研究生,研究方向:图像处理、深度学习和机器学习,E-mail:2518803 06@qq.com。

(编辑:夏道家)