DOI:10.11721/cqnuj20240305

深度超球面变分聚类网络*

余南骏，吴昊旻，陈飞宇，刘学文
(重庆师范大学 数学科学学院，重庆 401331)

摘要:基于变分自动编码器的深度聚类方法因较好的生成能力和聚类性能而受到广泛关注。然而现有的基于变分自动编码器的聚类算法通常需处理具有挑战性的证据下界(Evidence lower bound，ELBO)来实现聚类。同时，这些算法需要先验知识来假设类别分布，大多数方法还依赖于高斯分布或高斯混合模型。为了克服这些问题，提出了一种基于超球面变分自动编码器的端到端深度聚类网络。该方法采用 2 个后验分布的 KL 散度代替了复杂的 ELBO，并将 von Mises-Fisher 混合分布作为潜在空间嵌入的模型。此外它最大化了潜在表示和预测聚类分配之间的扩展互信息以实现更有区分性和平衡性的分配。通过与最先进的深度聚类技术在基准数据集上进行比较，验证了所提出的深度聚类方法的有效性。所建立的模型为复杂数据特征学习和聚类分析提供了一套新的方法，拓展了贝叶斯推断方法在深度聚类上的应用。

关键词:KL 散度；证据下界；超球面变分自动编码器；von Mises-Fisher 混合模型；扩展互信息
中图分类号:TP391.41　　**文献标志码**:A　　**文章编号**:1672-6693(2024)03-0100-13

传统的聚类方法，包括基于质心、层次和密度的算法[1-3]，都依赖于手工设计的特征提取器或距离度量来量化数据点之间的相似性。此外基于混合模型的流行方法[4-6]通过参数估计和概率分配实现聚类。然而这 2 类方法通常无法捕捉数据中的复杂结构和潜在特征，在实际应用中对于高维和复杂数据集往往效果不佳。

随着深度学习理论的不断完善，基于深度神经网络的模型在表示学习和生成方面展现出了强大的能力，例如变分自动编码器(variational auto-encoders，VAE)[7]和生成对抗网络(generative adversarial network，GAN)[8]。最近，例如变分深度嵌入(variational deep embedding，VaDE)[9]、高斯混合变分自编码(Gaussian mixture variational autoencoders，GMVAE)[10]及其他一些方法[11-12]将传统的 VAE 与高斯混合模型(Gaussian mixture model，GMM)相结合，并在聚类应用中取得了成功。这些方法将 GMM 集成到 VAE 框架中，以学习潜在表示并执行聚类任务。随后超球面变分自动编码器(hyperspherical variational auto-encoders，HVAE)[13]打破了长期以来将高斯分布或 GMM 作为潜在嵌入的传统，采用了 von Mises-Fisher(vMF)分布[14]。相较于高斯分布，vMF 分布的概率密度函数定义在单位超球面上，更适合捕获具有球形潜在结构的数据，并被应用于球形数据的建模[4,6]。此外，最近的研究表明[15-17]，使用球形嵌入可以带来更稳定和更优越的优化，且有助于改善聚类性能。然而当前基于 VAE 的聚类算法通常假设潜在嵌入遵循高斯或 GMM 分布，面临以下挑战：

1) 处理证据下界(evidence lower bound，ELBO)通常需要逼近方法，并且缺乏特征学习和聚类任务之间的协同作用；

2) 在聚类阶段需引入额外的离散分类变量(需要先验信息)到混合分布中，因此，将特征学习和聚类集成到单个神经网络中变得困难。

为解决上述问题，本文提出了一种新型深度变分聚类网络(deep hyperspherical variational clustering network，DHVC)。与大多数基于 VAE 的深度聚类方法不同，首先本文采用 vMF 混合模型(von Mises-Fisher mixture modle，vMFMM)作为潜在嵌入，然后通过最小化给定观察样本的潜在变量的后验分布与聚类标签的后验分布之间的 KL 散度来驱动整个网络，从而搭起表征学习和无监督聚类之间的桥梁；其次 DHVC 是一个端到

* **收稿日期**:2023-12-12　**修回日期**:2024-04-02　**网络出版时间**:2024-06-07T14:26
资助项目:国家自然科学基金青年科学基金项目(No. 12101098)；重庆市自然科学基金面上项目(No. cstc2021jcyj-msxmX1053)；重庆市教育委员会科学技术研究计划重点项目(No. KJQN201800116)
第一作者简介:余南骏，男，研究方向为模式识别与人工智能，E-mail:1138831016@qq.com；**通信作者**:陈飞宇，男，副教授，E-mail:fchen@cqnu.edu.cn
网络出版地址:https://link.cnki.net/urlid/50.1165.n.20240607.1058.010

端的聚类网络，通过全连接的聚类层预测软聚类分配，而无需使用任何离散的分类变量；最后为了提高聚类性能，引入潜在变量和分配之间的扩展互信息（extended mutual information，EMI）[18]作为正则化项，以获得更有区分性和平衡性的分配。这项研究的主要贡献可以简要概括为以下 3 个方面：

1）提出了通过优化可计算的 2 个后验分布之间的 KL 散度代替难以处理的 ELBO 的方法，同时建立了表征学习和聚类任务之间的联系；

2）采用 vMFMM 作为潜在嵌入，并设计了一个端到端的聚类网络，无需额外的离散变量或先验信息，所有网络参数可通过反向传播进行更新；

3）引入了扩展互信息作为正则化项，以实现更有区分性和平衡性的分配。

1　相关工作

近些年来，随着计算技术的进步和数据挖掘的兴起，聚类算法在应用范围上有了明显的拓展，并且研究工作更加深入。自 Song 等人[19-20]的开创性工作开始，自动编码器首次被用于特征提取，并将 K-means 损失应用于约束这些自动编码器的嵌入层特征，标志着基于自动编码器的深度聚类算法开始迅速发展。例如，Xie 等人[21]提出了深度嵌入聚类（deep embedded clustering，DEC），该方法利用自动编码器网络获取原始数据的潜在特征表示，然后用于无监督聚类任务，从而实现了自适应的聚类分配。在 DEC 的基础上，Guo 等人[22]提出了改进的深度嵌入式聚类（improved deep embedded clustering，IDEC），它采用欠完备自动编码器，并使用聚类损失来微调编码器；在微调阶段，损失函数结合了相对熵和重构损失，以确保嵌入空间的代表性，产生了更优越的聚类结果。此外，Guo 等人[23]还提出了 DCEC，它扩展了 DEC 架构并通过整合卷积自动编码器并在特征空间中保持局部数据结构来改善聚类结果。Yan 等人[24]提出了基于卷积神经网络的深度聚类方法（joint unsupervised learning of deep representations and image clusters，JULE）并将它用于特征学习和图像聚类；该方法在前向传递中执行分层图像聚类，在反向传递中学习特征表示，通过循环迭代优化目标。Dizaji 等人[25]提出了深度嵌入正则化聚类（deep embedded regularized dustering，DEPICT）；该模型同样也是基于卷积自编码器的深度聚类模型，其中在编码器上添加了 Softmax 层以进行聚类，使用相对熵最小化定义聚类目标函数，并通过聚类分配频率的先验进行正则化。

基于生成模型的 GAN 也在聚类领域有着大量的研究工作。其中比较著名的是由 Mukherjee 等人[26]提出的 ClusterGAN（C-GAN）。该方法将生成器-逆向网络组合成自编码器，并将随机变量的分布定义为以 one-hot 方式编码的离散变量的混合，同时连续变量遵循高斯分布，通过最小化重构损失来实现随机变量与样本的潜在特征之间的对应关系。另外 Dizaji 等人[27]添加了约束条件，以确保随机变量的各个维度正交，从而捕获类别信息，并引入了一个新的聚类网络，将真实样本映射到与随机变量相同维度的向量。然而，与其他方法不同的是他们选择了对抗策略来实现联合分布的平衡。对抗性自编码（adversarial autoencoders，AAE）是由 Makhzani 等人[28]提出的概率自编码器，结合了 GAN 网络进行联合训练，改进了数据表示学习，不仅有助于数据重建，还增强了聚类任务的性能。最后 Ge 等人[29]提出了双重对抗自编码器（dual adversarial autoencoders for clustering，Dual-AAE）聚类方法，该方法最大化了观察样本和潜在变量子集之间的似然函数和互信息，以扩展 AAE 的能力；同时还设计了新的重构损失，使得聚类准确性可与一些监督的卷积神经网络（convolutional neural network，CNN）方法相媲美。

另一个用于聚类的生成模型是 VAE，其中一个著名的方法是由 Jiang 等人[9]提出的 VaDE。VaDE 是基于 VAE 的深度聚类方法，它在潜在空间中采用了 GMM 作为嵌入，并与 VAE 相结合进行变分推断。相较其他方法，VaDE 实现了更优越的聚类结果，同时保持了生成真实样本的能力。类似地，GMVAE[10]也采用了高斯混合先验，但是使用了不同的变分推断目标进行训练。该研究表明，最小信息约束的启发式方法可以缓解 VAE 经常遇到的过度正则化问题。另一个 VAE 的变体是潜在树变分自编码（latent tree variational autoencoder，LTVAE[11]，它假设潜在变量遵循树形结构的层次关系，其中每个潜在变量定义了聚类的一个方面，并采用梯度下降和通过消息传递的逐步 EM 进行更新。变分信息瓶颈聚类（variational information bottleneck for unsupervised clustering：deep Gaussian mixture embedding，VI-GMM）[12]则是在目标函数上推广了 ELBO，采用变分信息瓶颈方法，并将潜在空间建模为高斯混合。正如前面所提到的，尽管文献[13，30]讨论了球面嵌入，但这些工作主要使用 vMF 作为先验，并非专门为聚类任务而设计。因此所提出的 DHVC 模型在聚类任务中扩

展了 HVAE 和 vMFMM,同时采用了易处理的 KL 散度来优化,而不是繁琐的 ELBO,这也是与 DSVAE[17]的一个关键区别。

2 vMF 分布与超球面变分自编码

2.1 vMF 分布与 vMFMM 分布

vMF 分布通常被视为超球面上的正态高斯分布。类似于高斯分布,它由 d 维的平均方向 $\boldsymbol{\mu}\in\mathbf{R}^d$ 和浓度 $\lambda\geqslant 0\in\mathbf{R}$ 参数化构成。对于 $\lambda=0$ 的特殊情况,vMF 分布则为球面均匀分布 $U(\boldsymbol{z};S^{d-1})=\left(\frac{2(\pi^{m/2})}{\Gamma(m/2)}\right)^{-1}$,这里 S^{d-1} 表示 $d-1$ 维的球面,$\Gamma(\cdot)$表示伽马函数。将随机单位向量 $\boldsymbol{z}\in\mathbf{R}^d$ 的 vMF 分布的概率密度函数定义如下:

$$V(\boldsymbol{z}\mid\boldsymbol{\mu},\lambda)=C_d(\lambda)\exp(\lambda\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{z})。\tag{1}$$

其中:$\|\boldsymbol{\mu}\|^2=1$,$C_d(\lambda)$是归一化常数,$C_d(\lambda)=\lambda^{(d/2)-1}/(2\pi)^{d/2}I_{(d/2)-1}(\lambda)$,$I_v$ 表示阶数为 v 的第一类修正贝塞尔函数。

如果假设 $\boldsymbol{z}\in\mathbf{R}^d$ 是由 k 个 vMF 分布混合,则 vMFMM 可以定义为:

$$p(\boldsymbol{z}\mid\pi,\boldsymbol{\mu},\lambda)=\sum_{j=1}^{k}\pi_j V(\boldsymbol{z}\mid\boldsymbol{\mu}_j,\lambda_j)。\tag{2}$$

其中:$\pi=\{\pi_j\}_{j=1}^k$ 且 $\pi_j\geqslant 0$,$\sum_{j=1}^{k}\pi_j=1$。π_j 表示第 j 个 vMF 分布的权重,$\boldsymbol{\mu}_j$ 和 λ_j 分别表示 $V(\boldsymbol{z}\mid\boldsymbol{\mu}_j,\lambda_j)$中的均值方向和浓度参数。

2.2 HVAE

在 VAE 中,现假设观测变量 $\boldsymbol{x}\in\mathbf{R}^d$,$\boldsymbol{z}\in\mathbf{R}^r$ 是潜变量,$p_\varphi(\boldsymbol{x},\boldsymbol{z})$是通过 φ 参数化的联合分布,VAE 的目标是优化数据的对数似然(log-likelihood)$\int_{\boldsymbol{z}}p_\varphi(\boldsymbol{x},\boldsymbol{z})\mathrm{d}\boldsymbol{z}$,当 $p_\varphi(\boldsymbol{x},\boldsymbol{z})$被神经网络参数化的时候,对潜变量的边际似然积分是难以处理的,因此通过最大化 ELBO 来解决:

$$\log\int_{\boldsymbol{z}}p_\varphi(\boldsymbol{x},\boldsymbol{z})\mathrm{d}\boldsymbol{z}\geqslant E_{q(\boldsymbol{z})}[\log p_\varphi(\boldsymbol{x}\mid\boldsymbol{z})]-\mathrm{KL}(q(\boldsymbol{z})\|p(\boldsymbol{z}))。\tag{3}$$

其中:$E_{q(\boldsymbol{z})}(\cdot)$表示随机变量 $\boldsymbol{z}$ 服从分布 $q(\boldsymbol{z})$的条件下,对括号内的概率分布进行期望计算;$\mathrm{KL}(\cdot\|\cdot)$表示表示 2 个概率分布之间的 KL 散度;$q(\boldsymbol{z})$是近似后验分布。当 $q(\boldsymbol{z})=p(\boldsymbol{z}\mid\boldsymbol{x})$时,$q(\boldsymbol{z})$被优化为近似的真实后验且边界是紧的。$q(\boldsymbol{z})$是针对每个数据点 $\boldsymbol{x}$ 进行的优化,为了能扩展到更大的数据集,引入一个由神经网络参数化的推理网络 $q_\psi(\boldsymbol{z}\mid\boldsymbol{x};\theta)$,该神经网络的输出为每个数据点 $\boldsymbol{x}$ 的概率分布,因此式(3)中 ELBO 变为:

$$\mathcal{L}(\varphi;\psi)=E_{q_\psi(\boldsymbol{z}\mid\boldsymbol{x};\theta)}[\log p_\varphi(\boldsymbol{x}\mid\boldsymbol{z})]-\mathrm{KL}(q_\psi(\boldsymbol{z}\mid\boldsymbol{x};\theta)\|p(\boldsymbol{z}))。\tag{4}$$

在原始的 VAE 中,先验和后验都被定义为高斯分布,而 HVAE 的贡献则是将先验和后验定义为了上述的 vMF 分布,并且推导了基于 vMF 分布的 ELBO。此外还设计了基于 vMF 分布的接受-拒绝采样方法,该方法会被用到后续网络结构的设计上。

在这里引入一个公式推导,为后续的损失函数推导做铺垫。

2 个 vMF 分布 $V(\boldsymbol{z}\mid\boldsymbol{\mu}_1,\lambda_1)=C_d(\lambda_1)\exp(\lambda_1\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{z})$和 $V(\boldsymbol{z}\mid\boldsymbol{\mu}_2,\lambda_2)=C_d(\lambda_2)\exp(\lambda_2\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{z})$之间的 KL 散度具有以下形式:

$$\mathrm{KL}(V(\boldsymbol{z}\mid\boldsymbol{\mu}_1,\lambda_1)\|V(\boldsymbol{z}\mid\boldsymbol{\mu}_2,\lambda_2))=\int_{\boldsymbol{z}}V(\boldsymbol{z}\mid\boldsymbol{\mu}_1,\lambda_1)\log\frac{C_d(\lambda_1)\exp(\lambda_1\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{z})}{C_d(\lambda_2)\exp(\lambda_2\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{z})}\mathrm{d}\boldsymbol{z}=$$

$$\log\frac{C_d(\lambda_1)}{C_d(\lambda_2)}+E_{V(\boldsymbol{z}\mid\boldsymbol{\mu}_1,\lambda_1)}[\lambda_1\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{z}]-E_{V(\boldsymbol{z}\mid\boldsymbol{\mu}_1,\lambda_1)}[\lambda_2\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{z}]=\frac{C_d(\lambda_1)}{C_d(\lambda_2)}+\frac{I_{d/2}(\lambda_1)}{I_{(d/2)-1}(\lambda_1)}[\lambda_1-\lambda_2\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_2]。\tag{5}$$

如果将其中的一个 vMF 分布退化为球面均匀分布 $U(\boldsymbol{z};S^{d-1})$,则两者之间的 KL 散度退化为在 HVAE 中需优化的 KL 散度项:

$$\log C_d(\lambda)+\lambda\frac{I_{d/2}(\lambda)}{I_{(d/2)-1}(\lambda)}+\log\left(\frac{2(\pi^{d/2})}{\Gamma(d/2)}\right)^{-1}。\tag{6}$$

其中:$\Gamma(\cdot)$表示伽马函数。同时,式(6)结果即为式(3)中第二项的最终推导。

2.3 ELBO

DHVC 的训练过程与大多 VAE 类似,主要包括前向传播、损失计算和反向传播。其中值得注意的是

DHVC 与其他方法在损失函数上不同。下面介绍经典的基于 VAE 的深度聚类方法 VaDE 的 ELBO 处理过程，VaDE 的优化策略本质为最大似然法则，即最大化概率 $\log\ p(\boldsymbol{x})$ 为：

$$\log\ p(\boldsymbol{x})=\log\int_{\boldsymbol{z}}\sum_{\boldsymbol{c}}p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{c})\mathrm{d}\boldsymbol{z}=$$

$$\log\int_{\boldsymbol{z}}\sum_{\boldsymbol{c}}q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})\frac{p(\boldsymbol{x},\boldsymbol{z},\boldsymbol{c})}{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}\mathrm{d}\boldsymbol{z}\geqslant E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{z},\boldsymbol{c},\boldsymbol{x})}{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}\right]\triangleq\mathcal{L}_{\mathrm{ELBO}}。\tag{7}$$

另外，基于平均场(mean-field)假设可将 $q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})$ 分解为 $q(\boldsymbol{z}\mid\boldsymbol{x})$ 和 $q(\boldsymbol{c}\mid\boldsymbol{x})$ 的乘积。结合 VaDE 中的图模型，可将联合概率 $p(\boldsymbol{z},\boldsymbol{c},\boldsymbol{x})$ 拆解为 $p(\boldsymbol{z},\boldsymbol{c},\boldsymbol{x})=p(\boldsymbol{x}\mid\boldsymbol{z})p(\boldsymbol{z}\mid\boldsymbol{c})p(\boldsymbol{c})$，则最终 ELBO 可以分解为如下形式：

$$E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{z},\boldsymbol{c},\boldsymbol{x})}{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}\right]=E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}[\log\ p(\boldsymbol{z},\boldsymbol{c},\boldsymbol{x})-\log\ q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})]=$$

$$E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}[\log\ p(\boldsymbol{x}\mid\boldsymbol{z})]+E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}[\log\ p(\boldsymbol{z}\mid\boldsymbol{c})]+E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}[\log\ p(\boldsymbol{c})]-$$

$$E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}[\log\ q(\boldsymbol{z}\mid\boldsymbol{x})]-E_{q(\boldsymbol{z},\boldsymbol{c}\mid\boldsymbol{x})}[\log\ q(\boldsymbol{c}\mid\boldsymbol{x})]。\tag{8}$$

最后，VaDE 对 $q(\boldsymbol{c}\mid\boldsymbol{x})$ 采用一种特殊的计算方式以最大化 ELBO：

$$q(\boldsymbol{c}\mid\boldsymbol{x})=p(\boldsymbol{c}\mid\boldsymbol{z})=\frac{p(\boldsymbol{c})p(\boldsymbol{z}\mid\boldsymbol{c})}{\sum_{\boldsymbol{c}'=1}^{k}p(\boldsymbol{c}')p(\boldsymbol{z}\mid\boldsymbol{c}')}。\tag{9}$$

上述这种特殊近似方法是对 ELBO 进行重新推导以后，结合合理的假设给出的计算方法。值得注意的是，因为 $p(\boldsymbol{z}\mid\boldsymbol{c})$ 捕获了 $\boldsymbol{c}$ 和 $\boldsymbol{z}$ 之间的关系，式(9)的计算方法允许减少由平均场近似引起的信息损失，并且在实践中效果很好。然而这种对 $q(\boldsymbol{c}\mid\boldsymbol{x})$ 的处理方式就数学的严谨性而言存在一定问题，并且对 ELBO 的处理上包含但不限于要计算复杂的期望以及优化目标通常是非凸的等问题，这些都导致 ELBO 难以处理。因此提出了一个新的损失函数代替 ELBO 来优化迭代。

3　DHVC

本节主要讨论基于 VAE 的 DHVC。先将 vMFMM 作为潜在空间的嵌入，提出用 KL 散度代替 ELBO 并推导出 DHVC 模型的主要目标函数，再将 EMI 引入到模型中，最后描述所提出的 DHVC 模型的网络结构。

3.1　DHVC 目标函数

现给定观测样本 $\boldsymbol{x}\in\mathbf{R}^d$ 是通过潜在变量 $\boldsymbol{z}\in\mathbf{R}^r$ 生成，$c\in\{1,\cdots,k\}$ 表示样本 $\boldsymbol{x}$ 的真实聚类标签，k 表示预先给定的聚类数，现假设 $q(\boldsymbol{z}\mid\boldsymbol{x})$ 为 $\boldsymbol{x}$ 的后验分布(记为 $\boldsymbol{x}$-后验)和 $q(\boldsymbol{z}\mid\boldsymbol{c})$ 为 $\boldsymbol{c}$ 的后验分布(记为 $\boldsymbol{c}$-后验)。对于具有多类结构的数据，假设来自第 j 类的变量 $\boldsymbol{z}$ 遵循个体分布 $q(\boldsymbol{z}\mid\boldsymbol{c}=j)$，将整个 $\boldsymbol{c}$-后验分布 $q(\boldsymbol{z}\mid\boldsymbol{c})$ 定义为：

$$q(\boldsymbol{z}\mid\boldsymbol{c})=\sum_{j=1}^{k}p_jq(\boldsymbol{z}\mid\boldsymbol{c}=j)。\tag{10}$$

其中：软分配 $p_j=p(\boldsymbol{c}=j\mid\boldsymbol{z})$ 表示潜变量 $\boldsymbol{z}$ 属于第 j 类的概率，并且有 $\sum_{j=1}^{k}p_j=1$。

提出用 $\boldsymbol{x}$-后验分布和 $\boldsymbol{c}$-后验分布之间的 KL 散度作为目标函数：

$$\mathrm{KL}(q(\boldsymbol{z}\mid\boldsymbol{x})\|q(\boldsymbol{z}\mid\boldsymbol{c}))=\int_{\boldsymbol{z}}q(\boldsymbol{z}\mid\boldsymbol{x})\log\frac{q(\boldsymbol{z}\mid\boldsymbol{x})}{q(\boldsymbol{z}\mid\boldsymbol{c})}\mathrm{d}\boldsymbol{z}。\tag{11}$$

式(11)建立了表征学习与无监督聚类之间的关键联系。表征学习的目标是找到一种映射，能有效地将观察样本中所含的信息保留到潜在变量中，同时无监督聚类任务需要在学习的潜在变量内部获得解耦合的表示。通过最小化式(11)的 KL 散度，潜在变量既适用于表征学习又适用于聚类任务。这意味着聚类任务有助于获取更具意义的表示，而表征学习任务则增强了聚类的性能。

结合式(10)、(11)有：

$$\mathrm{KL}(q(\boldsymbol{z}\mid\boldsymbol{x})\|q(\boldsymbol{z}\mid\boldsymbol{c}))=\int_{\boldsymbol{z}}q(\boldsymbol{z}\mid\boldsymbol{x})\log\frac{q(\boldsymbol{z}\mid\boldsymbol{x})}{\sum_{j=1}^{k}p_jq(\boldsymbol{z}\mid\boldsymbol{c}=j)}\mathrm{d}\boldsymbol{z}=$$

$$\int_{\boldsymbol{z}}q(\boldsymbol{z}\mid\boldsymbol{x})\log\ q(\boldsymbol{z}\mid\boldsymbol{x})\mathrm{d}\boldsymbol{z}-\int_{\boldsymbol{z}}q(\boldsymbol{z}\mid\boldsymbol{x})\log\sum_{j=1}^{k}p_jq(\boldsymbol{z}\mid\boldsymbol{c}=j)\mathrm{d}\boldsymbol{z}\leqslant$$

$$\int_{\boldsymbol{z}}q(\boldsymbol{z}\mid\boldsymbol{x})\log\ q(\boldsymbol{z}\mid\boldsymbol{x})\mathrm{d}\boldsymbol{z}-\sum_{j=1}^{k}p_j\int_{\boldsymbol{z}}q(\boldsymbol{z}\mid\boldsymbol{x})\log\ q(\boldsymbol{z}\mid\boldsymbol{c}=j)\mathrm{d}\boldsymbol{z}=\sum_{j=1}^{k}p_j\mathrm{KL}(q(\boldsymbol{z}\mid\boldsymbol{x})\|q(\boldsymbol{z}\mid\boldsymbol{c}=j))。\tag{12}$$

假设 $\boldsymbol{x}$-后验分布和 $\boldsymbol{c}$-后验分布分别遵循式(1)的 vMF 分布和式(2)的 vMFMM 分布:

$$q(\boldsymbol{z}\mid\boldsymbol{x})=V(\boldsymbol{z}\mid\boldsymbol{\mu}_x,\lambda_x),q(\boldsymbol{z}\mid\boldsymbol{c})=\sum_{j=1}^{k}\pi_j V(\boldsymbol{z}\mid\boldsymbol{\theta}_j,\lambda)。$$

其中:平均方向 $\boldsymbol{\mu}_x$ 和浓度 λ_x 由多层感知器训练得到,$\boldsymbol{\theta}_j$ 代表第 j 类的聚类中心,λ 为对应第 j 类的浓度,这里把浓度设为固定的相同值,下面会阐述这样设置的原因。p_j 表示潜变量 $\boldsymbol{z}$ 属于第 j 类的概率,从生成角度思考,π_j 为从类别中选取第 j 类生成潜在嵌入 $\boldsymbol{z}$ 的概率。因此可以假设 π_j 为 p_j。结合式(5)、(12)可得 2 个后验分布的 KL 散度解为:

$$\sum_{j=1}^{k}p_j\mathrm{KL}(q(\boldsymbol{z}\mid\boldsymbol{x})\|q(\boldsymbol{z}\mid\boldsymbol{c}=j))=\sum_{j=1}^{k}p_j\left[\log\frac{C_d(\lambda_x)}{C_d(\lambda)}+\frac{I_{d/2}(\lambda_x)}{I_{(d/2)-1}(\lambda_x)}(\lambda_x-\lambda\boldsymbol{\theta}_j^{\mathrm{T}}\boldsymbol{\mu}_x)\right]=$$

$$\log\frac{C_d(\lambda_x)}{C_d(\lambda)}+\frac{I_{d/2}(\lambda_x)}{I_{(d/2)-1}(\lambda_x)}\left[\lambda_x-\lambda\boldsymbol{\mu}_x^{\mathrm{T}}\sum_{j=1}^{k}p_j\boldsymbol{\theta}_j\right]。\tag{13}$$

式(13)即为 DHVC 的主要目标函数。需要注意的是,λ 原本表示的是第 j 类的浓度。然而在这里将每个类的浓度设置为一个固定值,主要是因为:首先在球形潜在空间中固定每个类的浓度可以确保每个类占据固定的区域,从而保证更稳定和可控的聚类过程;此外将浓度设为固定值可以保持每一类有相似的形状,避免不必要的扭曲或拉伸;最后在数据集较小或噪声水平较高的情况下,固定 λ 可增强聚类算法的稳定性。

3.2 EMI 正则化

预测的聚类标签 $\boldsymbol{c}$ 应尽可能包含潜在变量 $\boldsymbol{z}$ 的基本信息,因此提出最大化潜在变量 $\boldsymbol{z}$ 和 $\boldsymbol{c}$ 之间的 EMI,定义如下:

$$I_E(\boldsymbol{c};\boldsymbol{z})=-\alpha H(\boldsymbol{c}\mid\boldsymbol{z})+H(\boldsymbol{c})=\alpha\sum_{\boldsymbol{c}}p(\boldsymbol{c}\mid\boldsymbol{z})\log\ p(\boldsymbol{c}\mid\boldsymbol{z})-\sum_{\boldsymbol{y}}\tilde{p}(\boldsymbol{c})\log\ \tilde{p}(\boldsymbol{c})=$$

$$\frac{\alpha}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}p(\boldsymbol{c}_j\mid\boldsymbol{z}_i)\log\ p(\boldsymbol{c}_j\mid\boldsymbol{z}_i)-\sum_{j=1}^{k}\tilde{p}\log\ \tilde{p}。\tag{14}$$

其中:$H(\boldsymbol{c}\mid\boldsymbol{z})$表示条件熵,$H(\boldsymbol{c})$表示 $\boldsymbol{c}$ 的熵,$\tilde{p}(\boldsymbol{c})$是 $\boldsymbol{c}$ 的边缘分布,$\tilde{p}(\boldsymbol{c})$的每个分量可以通过 $\tilde{p}(\boldsymbol{c})=\sum_{i=1}^{n}p(\boldsymbol{c}=j\mid\boldsymbol{z}_i)$ 计算,并且 α 是一个平衡参数用于适应不同的数据集分布。

根据熵和条件熵的定义,最大化 $H(\boldsymbol{c})$可以鼓励平衡的分配,因为当 $\tilde{p}(\boldsymbol{c})=\frac{1}{k}$对于所有 j 时,$H(\boldsymbol{c})$最大化。另一方面,当软分配 $p(\boldsymbol{c}=j\mid\boldsymbol{z}_i)$大约为 0 或 1 时,$H(\boldsymbol{c})$最小,这也使得 $p(\boldsymbol{c}_j\mid\boldsymbol{z}_j)$变得稳定。此外参数 α 被设置为一个适当的值,用于控制不同数据集的聚类分配的平衡性和稳定性。

最后为保留模型的生成能力引入重构项结合式(12)、(13),因此所提出的 DHVC 的整体目标函数可以定义为:

$$\mathcal{L}=\mathrm{KL}(q(\boldsymbol{z}\mid\boldsymbol{x})\|q(\boldsymbol{z}\mid\boldsymbol{c}))-\alpha_e I_E(\boldsymbol{c};\boldsymbol{z})-E_{\boldsymbol{x}\sim q(\boldsymbol{z}\mid\boldsymbol{x})}[\log\ p(\boldsymbol{x}\mid\boldsymbol{z})]。\tag{15}$$

其中:$p(\boldsymbol{x}\mid\boldsymbol{z})$表示 VAE 中先验概率,可以用 Monte Carlo 期望估计来计算该项;α_e 是控制 EMI 权重的超参数。值得注意的是,由于在 EMI 中同时存在熵和交叉熵,因此还有 2 个相应的超参数分别表示为 α_{e1} 和 α_{e2}。

3.3 DHVC 结构

结合先前的假设和理论推导,所提出的 DHVC 包括推断网络和聚类网络以及生成网路,架构如图 1 所示。

推断网络:$[\boldsymbol{\mu}_x,\lambda_x]=f(\boldsymbol{x};\psi),q(\boldsymbol{z}\mid\boldsymbol{x})=V(\boldsymbol{z}\mid\boldsymbol{\mu}_x,\lambda_x)$。

聚类网络:$p=h(\boldsymbol{z};\theta),\ p_j=p(\boldsymbol{c}=j\mid\boldsymbol{z})=\exp(\boldsymbol{z}^{\mathrm{T}}\theta_j)\Big/\sum_{j'=1}^{k}\exp(\boldsymbol{z}^{\mathrm{T}}\theta_{j'})$。

生成网络:$\overline{\boldsymbol{x}}=g(\boldsymbol{z};\varphi)$。

推断网络与传统的变分自编码器(VAE)网络类似,其中 $f(\boldsymbol{x};\psi)$代表推断网络,以 $\boldsymbol{x}$ 和网络参数 ψ 作为输入,并输出平均方向 $\boldsymbol{\mu}_x$ 和浓度 λ_x。此外潜变量 $\boldsymbol{z}$ 是使用 $q(\boldsymbol{z}\mid\boldsymbol{x})=V(\boldsymbol{z}\mid\boldsymbol{\mu}_x,\lambda_x)$进行采样,采样过程参考 HVAE 的形式进行。$h(\boldsymbol{z};\theta)$则代表聚类网络,它是利用全连接层构建,以 softmax 作为激活函数。它以 $\boldsymbol{z}$ 和网络参数 θ 作为输入,并在前向计算中输出软聚类分配 $p\in\mathbf{R}^k$,参数 θ_j 表示的是第 j 类的聚类中心。显然 p_j 越大,$\boldsymbol{z}$ 和 $\boldsymbol{\theta}_j$ 离得越近。$\overline{\boldsymbol{x}}=g(\boldsymbol{z};\varphi)$是以 $\boldsymbol{z}$ 为输入,φ 为参数的生成网络,网络的输出则是样本$\overline{\boldsymbol{x}}$。

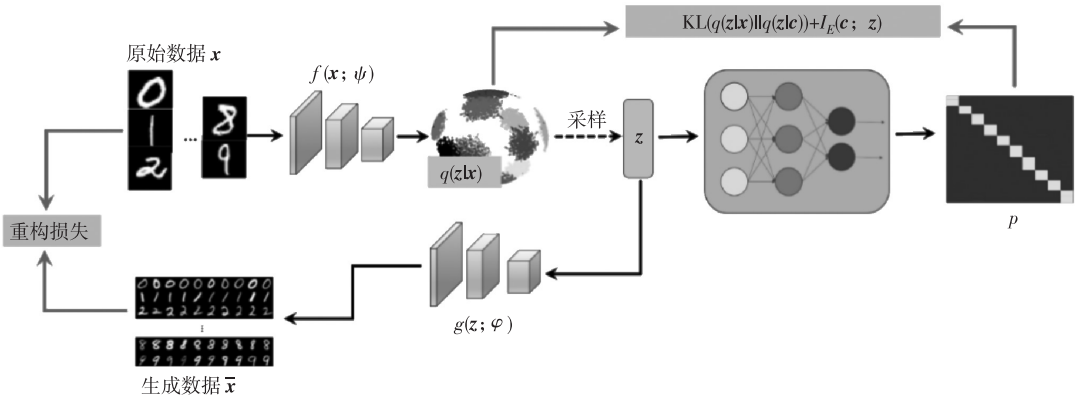


图 1　DHVC 结构

Fig. 1　Deep hyperspherical variational clustering network structure

这里对于每个质心 θ_j 的更新涉及计算它在第 j 类的期望值。为了优化中心 θ，采用了 SGVB 估计器和 Adam 优化器。由于从 $q(\boldsymbol{z}\mid\boldsymbol{x})$ 中采样涉及到离散操作，传统的梯度下降不能直接用于参数更新。因此在这种情况下，通过采用基于接受-拒绝采样的重参数化技巧[31]来解决这个问题，这使得重参数化技巧能够扩展到 vMF 分布，就像 HVAE 中所讨论的那样。

下面的算法 1 大致概括了 DHVC 的流程。

算法 1　DHVC 算法

1）给定一组观测数据 $\boldsymbol{x}$，聚类数 k

2）初始化网络权重 ψ，θ 以及 φ

3）for $t=1:T$ do

4）　推断阶段：计算潜在特征分布 $q(\boldsymbol{z}\mid\boldsymbol{x})=V(\boldsymbol{z}\mid\boldsymbol{\mu}_x,\lambda_x)$，其中$[\boldsymbol{\mu}_x,\lambda_x]=f(\boldsymbol{x};\psi)$

5）　采样通过 HAVE：$\boldsymbol{z}\sim V(\boldsymbol{z}\mid\boldsymbol{\mu}_x,\lambda_x)$

6）　聚类阶段：计算软分配 $p=h(\boldsymbol{z};\theta)$

7）　生成阶段：生成新的样本$\overline{\boldsymbol{x}}=g(\boldsymbol{z};\varphi)$

8）　权重更新：通过反向传播更新网络权重

9）end for

4　实验及分析

4.1　数据集和实验设置

通过使用 Python 中的深度学习框架 PyTorch 实现了提出的 DHVC 模型。此外本研究对聚类方法进行了全面评估，涵盖了 4 个应用广泛的基准数据集：MNIST[32]、FASHION-MNIST(F-MNIST)[33]、REUTERS-10K(REU-10k)[34]和 CIFAR-10[35]。另外对 MNIST 数据集的 2 个版本进行了实验：MNIST-full 包含完整的 MNIST 数据集，以及 MNIST-test 只包含原始 MNIST 数据集的测试集。表 1 显示了数据集详细的信息。

表 1　数据集统计

Tab. 1　Dataset statistics

数据集	样本数	维度	类别数
MNIST-full	70 000	1×28×28	10
MNIST-test	10 000	1×28×28	10
F-MNIST	70 000	1×28×28	10
CIFAR-10	60 000	2 048	10
REUTERS-10K	10 000	2 000	4

在实验中,对于像MNIST的图像数据集,实验中采用了卷积层,并在其中使用诸如池化和批量归一化[36]等技术。对于文本和传感器记录的数据,实验中采用全连接层。具体的网络设置见表2。

表2 数据的网络设置

Tab.2 The network setting for the data

MNIST、F-MNIST		Reuters-10K、CIFAR-10	
编码器	解码器	编码器	解码器
输入 $x\in 1*28*28$	输入 z	输入 x	输入 z
5*5卷积,通道数32,ReLu,批量归一化	(聚类数,2*2*128)全连接	(特征维度,500)全连接批量归一化,ReLu	(聚类数,2 000)全连接批量归一化,ReLu
5*5卷积,通道数64,ReLu,批量归一化	3*3转置卷积,通道64,ReLu,批量归一化	(500,500)全连接批量归一化,ReLu	(2 000,500)全连接批量归一化,ReLu
5*5卷积,通道数128,ReLu,批量归一化	5*5转置卷积,通道32,ReLu,批量归一化	(500,2 000)全连接批量归一化,ReLu	(500,500)全连接批量归一化,ReLu
(2*2*128,聚类数)全连接	5*5转置卷积,通道数1,Sigmoid	(2 000,聚类数)全连接	(500,特征维度)全连接

在所有实验中,详细配置了潜在变量 z 的维度、编码器中近似后验的浓度 λ 以及软分配参数 k 以获得更好的聚类效果。在实践中,潜在空间的维度:MNIST、F-MNIST、CIFAR-10均设置为10,Reuters-10K设置为4,设置原理是潜在空间维度与数据集的类别数匹配。所有数据集实验得到浓度 λ 的最优值在15附近。软分配 k 的值与潜在空间的维度的值相对应,即MNIST、F-MNIST、CIFAR-10均设置为10,Reuters-10K设置为4。与传统的深度神经网络类似,在DHVC的情况下,选择了ReLU[37]作为非线性激活函数,并使用Adam优化器[38]进行优化学习,MNIST和F-MNIST的学习率设定为 10^{-3},CIFAR-10和REUTERS-10K的学习率则是 10^{-4}。

4.2 聚类结果比较

为了评估聚类性能的质量,本文采用了2个最广泛使用的评估指标:聚类准确度(accuracy,ACC)和归一化互信息(normalized metual information,NMI)[39]。将本文提出的DHVC与3种经典的方法进行了比较:K-means[40]、GMM[5]、vMFMM[4];并对14种先进技术进行了比较,包括DEC[21]、IDEC[22]、DCEC[23]、JULE[24]、DEPICT[25]、VaDE[9]、GMVAE[10]、LTVAE[11]、VIB-GMM[12]、C-GAN[26]、InfoGAN[41]、AAE[28]、Dual-AEE[29]、DGG[42]。详细的比较结果见表3,其中各种方法的聚类结果主要来自于对应的文献。

根据表中的数据,可以得出以下结论:首先,与传统的聚类方法相比,DHVC模型在ACC和NMI有明显的提升。这可以归因于深度神经网络学习优秀的潜在特征用于聚类的能力。其次,与最先进的模型相比,DHVC模型也表现出色。值得注意的是,在F-MNIST数据集上,大多数先进的聚类方法的ACC约为60%,而DHVC模型在准确度上超过了70%。这提升可以归因于用可优化的KL散度替换了ELBO,从而明显提高了聚类性能。此外F-MNIST上的NMI也有所改善。

最后还应用了 t-SNE降维技术[43]来可视化训练过程中MNIST-test数据集上潜在嵌入 z 的球面分布。从图2可观察到,随着训练迭代次数增加及聚类准确度提高的同时,类别之间的分离也更明显,类内聚合更紧密。

4.3 稳定性评估

本节进行2组稳定性评估实验。第1组实验基于3.3节中提到的假设:将浓度参数 λ 设置为固定值,为探究 λ 值是否会影响实验结果。因此实验在不同浓度下MNIST-full数据集收敛时的ACC以及收敛速度的比较。第2组实验则关注3.4节中提到的扩展化互信息正则化中不同数值的 α_{e1} 和 α_{e2} 参数在MNIST-test数据集上对实验结果的影响。

图3比较了ACC不同 λ 值下的收敛速度。从收敛速度的角度来看,当 $\lambda=10$ 时,它在大约第40个epoch时接近收敛,这比其他 λ 值更快,这里当 $\lambda=0.01$、0.1、1时,大约在150个epoch时收敛,而 $\lambda=100$ 始终没有收敛的趋势。此外与其他 λ 值相比,$\lambda=10$ 收敛时的ACC和NMI更高。然而 $\lambda=100$ 的收敛曲线却始终处于较低的精度并且训练过程中波动较大,推测是当 λ 值越大时,每一类占的区域就更小,数据点就难以跳出当前所在类去

寻找更优的中心，从而影响聚类效果。另一方面，在收敛稳定性方面，显然 $\lambda=10$ 的曲线比其他值更加平滑，表现出更高的稳定性。根据上述实验发现，$\lambda=10$ 时收敛最快。为进一步体现出模型在经过预训练后，在模型微调阶段具有更快的收敛性，将模型的收敛速度与使用相同训练方案的 3 种方法（DEC[21]、IDEC[22] 和 VaDE[9]）进行了比较。图 4 是不同聚类方法基于 MNIST-test 数据集在 ACC 和 NMI 的收敛效果。不难发现这 3 种方法收敛最快的是 IDEC，大约在 80 个 epoch 时就已接近收敛。而 DEC 和 VaDE 在 150 个 epoch 时还未体现出收敛的趋势，而本文的模型在 60 个 epoch 时已接近收敛，比上述最快的方法快了 20 个 epoch。实验结果表明，本文模型具有较快的收敛速度。

表 3　不同聚类方法在不同数据集上的聚类精度对比

Tab. 3　Comparison of clustering accuracies of different clustering methods on different datasets

聚类方法	MNIST-test		MNIST-full		F-MNIST		REU-10K		CIFAR-10	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
K-means	53.3	49.9	53.2	50.0	47.4	51.2	53.8	40.5	73.3	60.0
GMM	53.7	50.1	54.0	50.9	53.2	40.9	54.7	41.0	55.3	54.5
VMFMM	55.1*	52.8*	56.3*	53.0*	51.7*	41.7*	55.8*	43.0*	—	—
DEC	84.0	83.0	86.3	83.4	59.5	51.0	74.3*	49.7*	—	—
IDEC	88.1	86.7	87.4	85.9	60.9	51.3	75.6	49.8		
DCEC	85.3	83.6	89.0	88.5	—	—	—	—	—	—
AAE	90.5*	—	83.5	59.1	50.0	52.0	69.8	—	—	—
Dual-AAE	97.9	94.2*	97.6*	93.9*	—	—	81.5	58.1*	—	—
VaDE	94.4*	88.5*	94.4	89.1	62.9	61.1	79.8	52.9*	59.7	—
GMVAE	86.0	85.8	88.5	—	59.6*	57.0*	76.1*	50.9*	69.2	—
LTVAE	—	—	86.3	—	61.3	—	81.0	—	69.2	—
VIB-GMM	—	—	96.2	96.2	—	—	80.4	—	—	—
C-GAN	95.1*	89.3*	95.0	89.0	63.0	61.0	79.5*	55.3*	41.2	—
Info-GAN	89.0*	86.0*	87.0	84.0	61.0	59.0	75.3*	50.0*	—	—
JULE	96.1	91.5	96.4	91.3	56.3	60.8	—	—	—	—
DEPICT	96.3	91.5	96.5	91.7	39.2	39.2	—	—	—	—
DGG	97.7*	94.3*	97.6	94.0*	—	—	82.3	59.3*		
DHVC	**98.1**	**94.9**	**98.0**	**94.5**	**71.9**	**65.6**	**86.6**	**64.3**	**83.8**	**70.4**

注：* 表示该数据不是直接从原始论文获取的结果，加黑的数据表示最好的结果。

此外图 5 展示了 EMI 中不同参数 α_{e1} 和 α_{e2} 在 MNIST-test 数据集上对 ACC 和 NMI 的影响。从图中直观地观察到，对应较高 α_{e1} 和 α_{e2} 值的点更红，表示更高的 ACC 和 NMI。将 α_{e2} 设为 0 对 ACC 和 NMI 有明显影响，而相反地将 α_{e1} 设为 0 也会降低 ACC 和 NMI，但降低程度小于前者。

4.4　模型的生成能力与可视化

该实验旨在更加直观地展示模型的生成能力。为此在实验中对比了使用 VaDE 模型生成样本与使用 DHVC 模型生成的样本。通过比较这 2 种生成样本，可以更好地评估和理解该模型的生成能力。图 6 是 VaDE 模型生成的 MNIST 和 F-MNIST 样本数据集，图 7 是 DHVC 模型生成的 MNIST 和 F-MNIST 样本数据集。对比发现，DHVC 模型与 VaDE 模型不论是 MNIST 还是 F-MNIST 都具有相似的清晰度，并且同样具有很好的多样性，值得注意的是 DHVC 模型的多样性要比 VaDE 模型更加丰富，从数字 5 可以明显的看出来。对于 F-MNIST 而言，由于准确率低，VaDE 和 DHVC 在同一行的时尚服饰不能显示同一簇。同时，VaDE 还生成了一些不存在的图像，而 DHVC 具有高度逼真的样本，并且在平滑和多样性方面都有良好的性能。

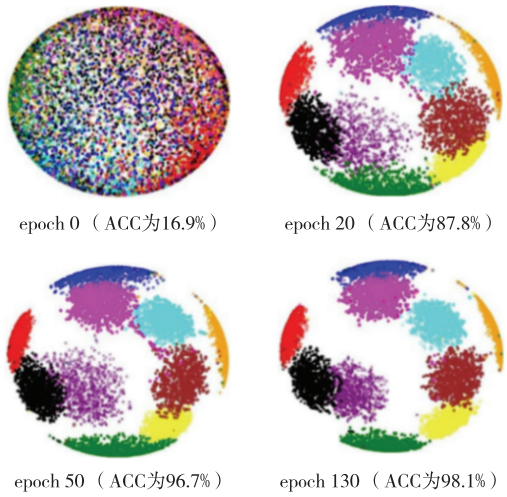


图 2 MNIST-test 数据集在训练过程中潜在特征的可视化

Fig. 2 Visualization of latent features of the MNIST-test dataset during the training process

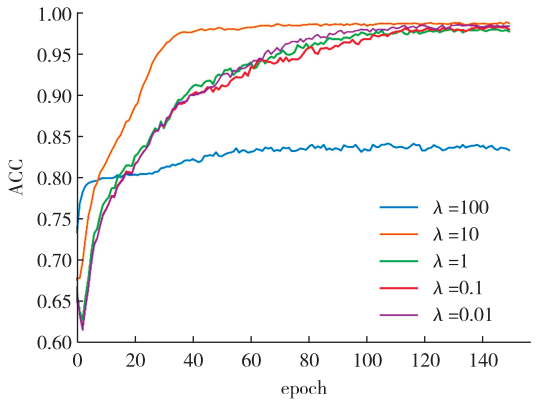
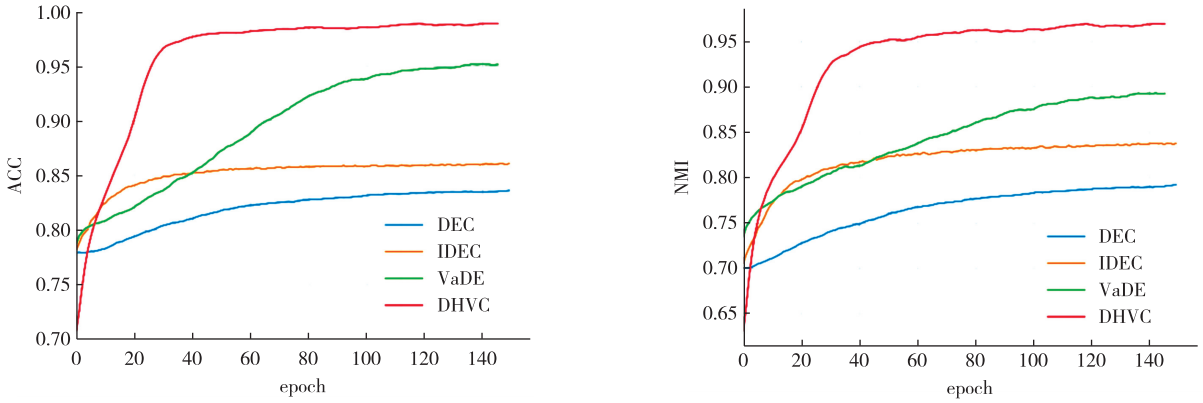


图 3 在 MNIST-test 上不同浓度对 ACC 收敛速度的影响

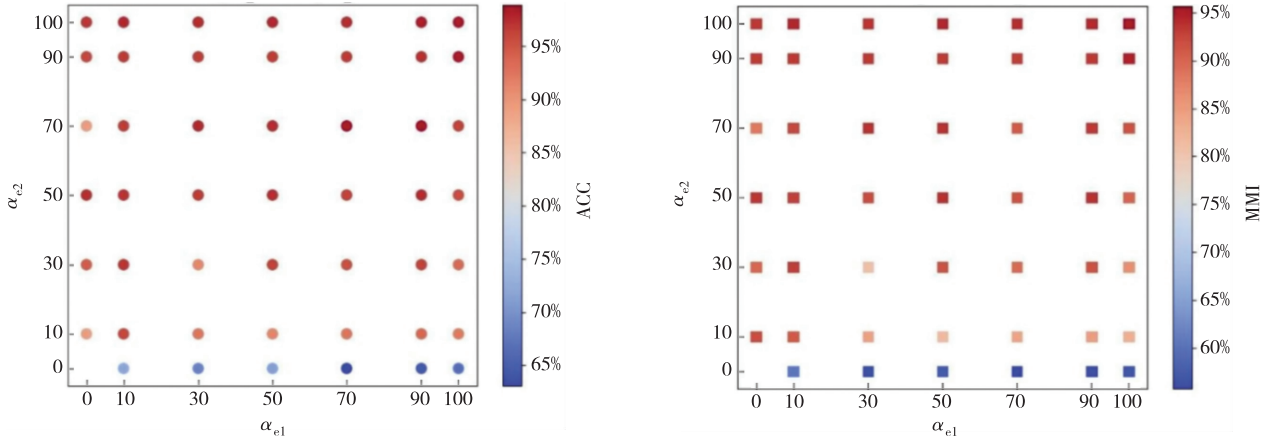
Fig. 3 The impact of different concentrations on the convergence speed of ACC on MNIST-test



a ACC b NMI

图 4 不同聚类方法在 MNIST-test 数据集上的收敛效果

Fig. 4 The impact of different clustering methods on the convergence speed on MNIST-test



a ACC b NMI

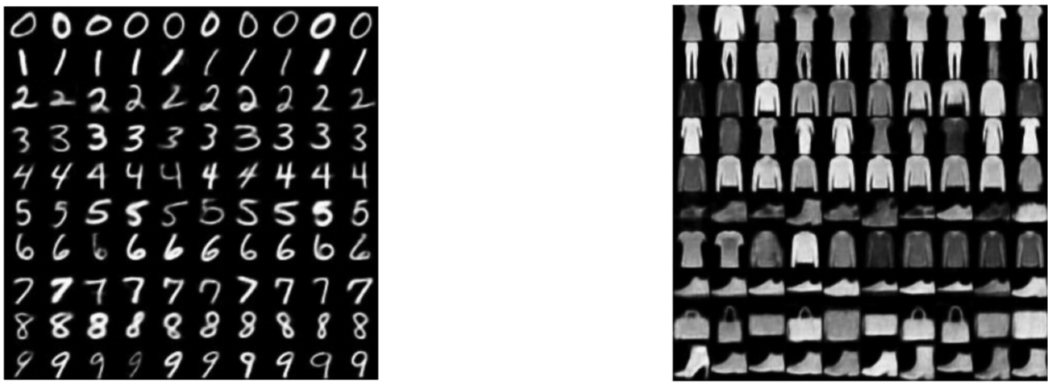
图 5 不同参数 α_{e1} 和 α_{e2} 在 MNIST-test 数据集上的影响

Fig. 5 The impact of different parameters α_{e1} and α_{e2} on the MNIST-test dataset

a　MNIST 数据集　　b　F-MNIST 数据集

图 6　VaDE 在不同数据集上的生成样本

Fig. 6　Generated samples of VaDE on different datasets



a　MNIST 数据集　　b　F-MNIST 数据集

图 7　DHVC 在不同数据集上的生成样本

Fig. 7　Generated samples of DHVC on different datasets

4.5　消融实验

为了评估 DHVC 模型内每个组件的贡献，进行了 3 项消融实验，结果见表 4。在这个分析中，KL 代表后验分布和近似后验分布之间的 KL 散度，MSE 是生成数据与原始数据之间的重建误差项，EMI 表示潜在特征 $\boldsymbol{z}$ 和标签 $\boldsymbol{c}$ 之间的扩展互信息正则化。

表 4　不同数据集在 ACC 和 NMI 上的消融实验

Tab. 4　Ablation experiments of different datasets on ACC and NMI

方法	MNIST-test		MNIST-full		F-MNIST		REU-10K		CIFAR-10	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KL＋MSE	76.57	71.92	75.80	71.16	66.87	61.83	76.26	48.70	73.44	58.12
KL＋EMI	97.15	93.27	96.99	92.44	70.79	63.66	86.52	64.17	83.32	69.81
KL＋EMI＋MSE	98.17	94.94	97.98	94.51	71.87	65.55	86.58	64.28	83.77	70.37

为了对每组实验结果进行更有力的对比，实验中保留了 4 位有效数字。通过上表可观察到：以 KL＋MSE 作为基线，去除 MSE 并添加 EMI 明显提高了性能，ACC 和 NMI 大多增加了超过 10%。当同时使用 EMI 和 MSE 虽未导致 ACC 和 NMI 得到明显改善，但 ACC 和 NMI 也增加了 1%左右，因此在实践中同时保留了 MSE 和 EMI，这使得模型具有生成功能的同时又可以明显提升聚类性能。

5　总结

本文提出了一种基于 VAE 的深度聚类网络。通过对现有方法的理论进行改进以及实验的创新，提出的模

型主要有以下贡献。首先引入了可优化计算KL散度,用于替代具有挑战性ELBO,而无需先验信息。其次采用了vMFMM分布而不是常用的GMM分布用于潜在嵌入。最后通过扩展互信息(EMI)对聚类进行优化来增强和改进模型。在4个广泛使用的数据集上的实验结果证明了所提模型的有效性。

参考文献:

[1] ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//FAYYAD U,PIATETSKY-SHAPIRO G,SMYTH P. Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining,Menlo Park:AAAI,1996:226-231.

[2] JOHNSON S C. Hierarchical clustering schemes[J]. Psychometrika,1967,32(3):241-254.

[3] MAC QUEEN J B. Some methods for classification and analysis of multivariate observations[C]//CAM L M L,NEYMAN J. Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability,Berkeley:University of California Press,1967,1:281-297.

[4] BANERJEE A,DHILLON I S,GHOSH J,et al. Clustering on the unit hypersphere using von Mises-Fisher distributions[J]. Journal of Machine Learning Research,2005,6(9):1345-1382.

[5] MC LACHLAN G J,LEE S X,RATHNAYAKE S I. Finite mixture models[J]. Annual Review of Statistics and Its Application,2019,6:355-378.

[6] TAGHIA J,MA Z,LEIJON A. Bayesian estimation of the von-Mises Fisher mixture model with variational inference[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2014,36(9):1701-1715.

[7] KINGMA D P,WELLING M. Auto-encoding variational bayes[EB/OL]. (2013-12-20)[2023-12-12]. https://arxiv.org/abs/1312.6114v1.

[8] GOODFELLOW I J,POUGET-ABADIE J,MIRZA M,et al. Generative adversarial nets[EB/OL]. (2014-06-10)[2023-12-12]. https://arxiv.org/abs/1406.2661v1.

[9] JIANG Z,ZHENG Y,TAN H,et al. Variational deep embedding:an unsupervised and generative approach to clustering[C]//NEBEL B,SABLJAKOVIC-FRITZ V. Proceedings of The 26th International Joint Conference on Artificial Intelligence,Melbourne:AAAI Press,2017:1965-1972.

[10] DILOKTHANAKUL N,MEDIANO P A M,GARNELO M,et al. Deep unsupervised clustering with gaussian mixture variational autoencoders[EB/OL]. (2017-01-13)[2023-12-12]. https://arxiv.org/abs/1611.02648.

[11] LI X,CHEN Z,POON L K M,et al. Learning latent superstructures in variational autoencoders for deep multidimensional clustering[EB/OL]. (2019-02-22)[2023-12-12]. https://arxiv.org/abs/1803.05206.

[12] UĞUR Y,ARVANITAKIS G,ZAIDI A. Variational information bottleneck for unsupervised clustering:deep gaussian mixture embedding[J]. Entropy,2020,22(2):213.

[13] DAVIDSON T R,FALORSI L,DE CAO N,et al. Hyperspherical variational auto-encoders[C]//GLOBERSON A,SILVA R. Proceedings of The Thirty-Fourth Conference on Uncertainty in Artificial Intelligence,Monterey:AUAI Press,2018:856-865.

[14] MARDIA K V,JUPP P E,MARDIA K V. Directional statistics[M]. New York:Wiley,2000.

[15] FAN W,BOUGUILA N. Spherical data clustering and feature selection through nonparametric Bayesian mixture models with von Mises distributions[J]. Engineering Applications of Artificial Intelligence,2020,94:103781.

[16] FAN W,YANG L,BOUGUILA N,et al. Sequentially spherical data modeling with hidden Markov models and its application to fMRI data analysis[J]. Knowledge-Based Systems,2020,206:106341.

[17] YANG L,FAN W,BOUGUILA N. Deep clustering analysis via dual variational autoencoder with spherical latent embeddings[J]. IEEE Transactions on Neural Networks and Learning Systems,2021,34(9):6303-6312.

[18] PANG Y,CHEN F,HUANG S,et al. Deep fuzzy clustering with weighted intra-class variance and extended mutual information regularization[C]//DI FATTA G,SHENG V S,CUZZOCREA A,et al. 2020 International Conference on Data Mining Workshops. Sorrento:Institute of Electrical and Electronics Engineers,2020:464-471.

[19] SONG C,HUANG Y,LIU F,et al. Deep auto-encoder based clustering[J]. Intelligent Data Analysis,2014,18(6):65-76.

[20] SONG C,LIU F,HUANG Y,et al. Auto-encoder based data clustering[C]//RUIZ-SHULCLOPER J,SANNITI D B G. Progress in Pattern Recognition,Image Analysis,Computer Vision,and Applications:18th Iberoamerican Congress,CIARP 2013,Havana,Cuba,November 20-23,2013,Proceedings,Part I 18. Berlin:Springer,2013:117-124.

[21] XIE J,GIRSHICK R,FARHADI A. Unsupervised deep embedding for clustering analysis[C]//BALCAN M F,WEINBERGER K Q. Proceedings of the 33rd International Conference on International Conference on Machine Learning,New York:JMLR,

2016，48：478-487.

[22] GUO X，GAO L，LIU X，et al. Improved deep embedded clustering with local structure preservation[C]//NEBEL B，SABLJAKOVIC-FRITZ V. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence，Melbourne，Australia：AAAI Press，2017：1753-1759.

[23] GUO X，LIU X，ZHU E，et al. Deep clustering with convolutional autoencoders[C]//LIU D. Neural Information Processing：24 thInternational Conference，ICONIP2017，Guangzhou，China，November14-18，2017，Proceedings，Part II 24. Berlin：Springer，2017：373-382.

[24] YANG J，PARIKH D，BATRA D. Joint unsupervised learning of deep representations and image clusters[C]//BAJCZY R，LI F-F，TUYTELAARS T. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas：Institute of Electrical and Electronics Engineers，2016：5147-5156.

[25] GHASEDI DIZAJI K，HERANDI A，DENG C，et al. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization[C]//IKEUCHI K，MEDIONI G，PELILLO M. 2017 IEEE International Conference on Computer Vision (ICCV). Venice：Institute of Electrical and Electronics Engineers，2017：5747-5756.

[26] MUKHERJEE S，ASNANI H，LIN E，et al. Clustergan：Latent space clustering in generative adversarial networks[C]//VAN HENTENRYCK P，ZHOU Z-H. Proceedings of the AAAI Conference on Artificial Intelligence，Honolulu：AAAI Press，2019，33(1)：4610-4617.

[27] GHASEDI K，WANG X，DENG C，et al. Balanced self-paced learning for generative adversarial clustering network[C]//DAVIS L，TORR P，ZHU S-C. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach：Institute of Electrical and Electronics Engineers，2019：4386-4395.

[28] MAKHZANI A，SHLENS J，JAITLY N，et al. Adversarial autoencoders[EB/OL]. (2016-05-25)[2023-12-12]. https://arxiv.org/abs/1511.05644.

[29] GE P，REN C X，DAI D Q，et al. Dual adversarial autoencoders for clustering[J]. IEEE transactions on Neural Networks and Learning Systems，2019，31(4)：1417-1424.

[30] XU J，DURRETT G. Spherical latent spaces for stable variational autoencoders[EB/OL]. (2018-10-12)[2023-12-12]. https://arxiv.org/abs/1808.10805.

[31] NAESSETH C，RUIZ F，LINDERMAN S，et al. Reparameterization gradients through acceptance-rejection sampling algorithms[C]//SINGH A，ZHU J. Artificial Intelligence and Statistics. Ft Lauderdale：PMLR，2017：489-498.

[32] LE C Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278-2324.

[33] XIAO H，RASUL K，VOLLGRAF R. Fashion-mnist：a novel image dataset for benchmarking machine learning algorithms[EB/OL]. (2017-09-15)[2023-12-12]. https://arxiv.org/abs/1708.07747.

[34] APTÉ C，DAMERAU F，WEISS S M. Automated learning of decision rules for text categorization[J]. ACM Transactions on Information Systems，1994，12(3)：233-251.

[35] KRIZHEVSKY A，HINTON G. Learning multiple layers of features from tiny images[J]. Master's Thesis，University of Tront，2009：32-33.

[36] IOFFE S，SZEGEDY C. Batch normalization：accelerating deep network training by reducing internal covariate shift[C]//BACH F，BLEI D. Proceedings of The 32nd International Conference on International Conference on Machine Learning. Lille：JMLR，2015，37：448-456.

[37] MAAS A L，HANNUN A Y，NG A Y. Rectifier nonlinearities improve neural network acoustic models[C]//KINGSBURY B，SAINATH T N，DENG L，et al. 2013 Deep Learning for Audio，Speech and Language Processing. Atlanta：ICML Workshop，2013：1.

[38] KINGMA D P，BA J. Adam：a method for stochastic optimization[EB/OL]. (2017-01-30)[2023-12-12]. https://arxiv.org/abs/1412.6980v1.

[39] XU W，LIU X，GONG Y H. Document clustering based on non-negative matrix factorization[C]//26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval，July 28-August 1，2003，Toronto，Canada. New York：Association for Computing Machinery，2003：267-273.

[40] LLOYD S. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory，1982，28(2)：129-137.

[41] CHEN X，DUAN Y，HOUTHOOFT R，et al. Infogan：interpretable representation learning by information maximizing generative adversarial nets[C]//LEE D D，von LUXBURG U，GARNETT R，et al. Proceedings of the 30th International

Conference on Neural Information Processing Systems. New York:Curran Associates Inc,2016:2180-2188.

[42] YANG L X,CHEUNG N-M,LI J Y,et al. Deep clustering by gaussian mixture variational autoencoders with graph embedding [C]//O'CONNER L. Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway:IEEE,2019:6440-6449.

[43] van der MAATEN L,HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research,2008,9(11):2579-2605.

Deep Hyperspherical Variational Clustering Network

YU Nanjun, WU Haomin, CHEN Feiyu, LIU Xuewen

(School of Mathematical Sciences, Chongqing Normal University, Chongqing 401331, China)

Abstract: In recent years, the deep clustering approach based on variational autoencoders (VAE) has garnered significant attention due to its remarkable generative capabilities and clustering performance. However, existing VAE-based clustering algorithms typically revolve around dealing with the challenging Evidence Lower Bound (ELBO) to achieve clustering. Moreover, they often require prior knowledge of the class distribution, with most relying on Gaussian distributions or Gaussian mixture models. A deep clustering network based on Hyperspherical Variational Auto-Encoders (HVAE) is proposed, named DHVC (Deep Hyperspherical Variational Clustering Network). This approach substitutes the complex Evidence Lower Bound (ELBO) with the Kullback-Leibler Divergence of two posterior distributions and employs the von Mises-Fisher Mixture Model (vMFMM) distribution as an embedding in the latent space. Additionally, it maximizes the Extended Mutual Information (EMI) between latent representations and predicted cluster assignments for more discriminative and balanced allocations. The effectiveness of the proposed deep clustering method is validated through comparisons with state-of-the-art deep clustering techniques on benchmark datasets. The established model provides a novel approach for complex data feature learning and cluster analysis, extending the application of Bayesian inference methods in deep clustering.

Keywords: KL Divergence; evidence lower bound; hyperspherical variational auto-encoders; von Mises-Fisher mixture model; extended mutual information

(责任编辑 黄 颖)