基于重叠度的聚类数目判断指标

张 娟[1] 李 娜[2]
ZHANG Juan LI Na

摘 要 如今对于聚类问题，很多技术与模型都倾向于人为给定聚类数目。而现有的内部聚类判断指标更多考虑的是类内信息，忽略了类间信息，这导致指标的效果不理想，因此如何利用类间信息选择适当数量的聚类是聚类分析领域中广泛研究的问题。而基于重叠度聚类数目预测算法可以很好地解决这个问题。具体来说，首先使用基于改进过后高斯混合聚类算法对未标记的数据集进行预聚类；接着计算聚类类间的重叠，对重叠度的判断以确定和选择聚类的数量。其指标优势在于算法最大限度地减少了超参数的影响，并能够准确确定正确的聚类数。除此之外，所提出的指标还可以当作评判聚类结果的一种通用指标。经过大量实验表明，效果相对于同类型的方法有明显的提升。

关键词 深度学习；深度聚类；聚类数目

doi：10.3969/j.issn.1672-9528.2024.11.011

0 引言

聚类分析增强了对客观现实的认识。在商业等诸多领域，对于聚类数目 k 的预测是至关重要的，这是概念描述和偏差分析的前提。但随着数据对象越来越复杂，数据量越来越大，对于数据集的聚类数目 k 也越来越不可人为预测。因此，通过何种算法或技术手段对数据集聚类数目 k 的预测变成了一个研究难题。如果输入估计错误，会对聚类结果产生十分恶劣的影响。

传统的聚类算法可以通过很多指标预测聚类数目 k，例如内部指标 AIC 值，BIC 值 [1]，轮廓系数 [2] 等。它们的思想是通过观察不同 k 值下的不同参数值判断最适合当前数据集的聚类数目值。另外求预测值与真实值的误差平方和（SSE）也被看作一种聚类判断指标用于判断聚类数目。它们在一定程度上开创了类内指标预测聚类数目 k 值的先河。但近年来随着深度聚类的兴起，聚类算法可以适应更高维度体量更大的数据。例如文献 [3] 中提出的聚类算法采用基于动态规划的有效算法来交替优化子序列聚类和分割，从而实现高纬度时间序列数据中的子序列聚类分割问题；文献 [4] 中提出的算法通过神经网络完成了对高维数据的提取与文献 [5] 通过对图数据的学习后使用自监督机制聚类。类似的深度聚类算法性能优秀，但都需要预先输入聚类数目 k。而单一内部指标会导致利用信息失衡，得到对聚类数目的错误估计。因此在这种情况下类间指标的应用就显得格外重要。

在本文中提出了一种新颖的基于聚类重叠度的聚类数目判断指标。它首先使用聚类算法对数据集进行预聚类。接下来使用有效性度量来迭代评估聚类结果，从类间重叠度这一角度对聚类结果进行评估，并研究类间重叠数据的划分。最后观察重叠度的变化趋势，最终完成具体聚类个数 k 的确定。这是一种充分利用类间信息的判断指标，它不仅适用于传统算法，也适应于经过神经网络的高维数据进行过特征提取后的深度聚类算法。经过与其他内部指标的对比，本文的基于重叠度的类间信息指标判断方法效果更好。

1 重叠度算法

1.1 重叠度概念

在众多方法中有一个名为重叠度的概念被提出。它的直观含义是簇之间的重叠程度。现在的研究一般认为聚类结果越好，则簇间越分离。重叠度以这一思想为基本准则，通过簇之间的重叠程度这一类间信息来直观地了解聚类结果的重叠情况，判断聚类结果的好坏，同时判断簇的最佳中心数 k。它最早是文献 [6] 提出的。经过研究，又在文献 [7] 和高斯混合（GMM）模型聚类中被应用。本文根据这些研究进一步改进了重叠度的计算，使它可以来判断当前聚类效果的好坏。不需要更大的人为干预，最大限度减少超参数的影响。

1.2 基于高斯混合聚类算法改进后的预聚类

实际计算重叠度过程需要通过 GMM 模型计算。具体利用 GMM 模型投影后样本点不是得到一个确定的分类标记，

1. 太极计算机股份有限公司 北京 100012
2. 江西省生态文明研究院 江西南昌 330036

而是得到每个类的概率的信息计算。GMM 模型的概率密度函数为：

$$p_M(x)=\sum_{k=1}^{K}p(k)p(x|k)=\sum_{k=1}^{K}\alpha_k p(x|\mu_k,\Sigma_k) \tag{1}$$

式中：k 是模型的个数，即簇的个数；α_k 为属于第 k 个高斯的概率，也称为先验分布，其需要满足大于零，且对一个样本点 x 而言，α_k 之和等于 1；$p(x|k)$ 为第 k 个高斯的概率密度，其均值向量为 $\boldsymbol{\mu}_k$；$\boldsymbol{\Sigma}_k$ 为协方差矩阵。

改进后的聚类步骤如下。

（1）设置聚类数目 K 的范围，$K\subset[2,n)$，其中 n=3,4,5,⋯即初始化高斯混合模型的成分个数的取值范围。

（2）计算每个数据点属于每个高斯模型的概率，即计算后验概率。

（3）计算各项参数使得数据点的概率最大化，使用数据点概率的加权来计算这些新的参数，权重就是数据点属于该簇的概率。

（4） 重复迭代（2）和（3）直到收敛。

（5）迭代取值 K，建立不同 K 值下的 GMM 模型。直至 k=n，共得到 n 个 GMM 预聚类模型。

1.3 基于 GMM 模型的聚类重叠度计算

将每个簇的分布视为高斯分布，当研究两个不同的簇时，等价为正在研究不同高斯分布之间的关系。因此，高斯混合模型可以看作是 k 个单高斯模型的组合的模型，即 k 个聚类得到的簇的结合。而这 k 个子模型是混合模型的隐藏变量。通常，混合模型可以使用任何概率分布。这使得高斯混合模型对于聚类的计算具有良好的数学性质。

当使用 GMM 对具有两个高斯混合分布的数据进行聚类或分析时 (即在聚类时)，聚类期望两个高斯分布的距离尽可能远，以便数据是可微的。但是，在许多情况下，两个高斯分布会重叠。这就是重叠度的计算，是划分簇间重叠数据的重要依据，如图 1 所示。

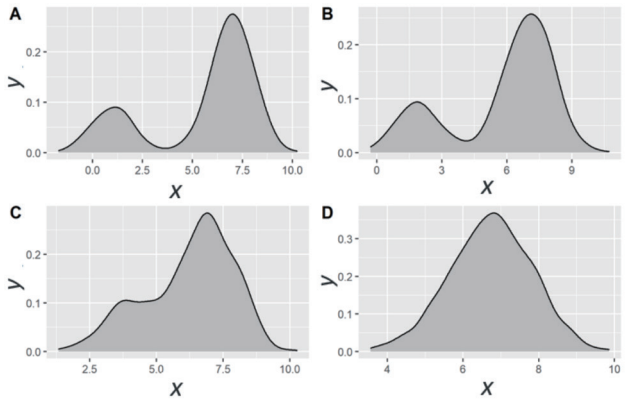

图 1 不同类型高斯混合模型的分布

部分指标可以间接反映两个高斯分布的重叠。例如，可以计算马氏距离，Bhattacharyya 距离或 Kullback-Leibler (KL) 距离，可以测量两个高斯分布的相似性。但是马氏距离的前提是两个分布具有相同的协方差。Bhattacharyya 距离和 KL 距离都考虑了协方差，但没有考虑高斯混合分布的混合系数。此外，KL 距离对于高维正态分布没有解析解，计算起来较为复杂。文献 [7] 提出的重叠相似度的计算是基于两个高斯分布混合形成的脊线理论。该方法考虑了高斯混合分布中的所有参数，包括均值、协方差和混合系数。d 维空间中 K 个分量的高斯混合模型的概率密度函数（PDF）的数学公式为：

$$p(x)=\sum_{i=1}^{k}\lambda_i G_i(X,\mu,\sum_i),X\in\mathbb{R}^d \tag{2}$$

式中：λ_i 满足 λ_i> 0，是$\sum_{i=1}^{k}\lambda_i$=1 的混合因子，其中样本点向量为 $\boldsymbol{X}$，均值向量为 $\boldsymbol{\mu}$，协方差均值为 $\boldsymbol{\Sigma}_i$；G_i 是一个 d 维高斯分布，可以表示为：

$$G_i(X,\mu,\sum_i)=\frac{\exp(\frac{1}{2}(X-\mu_i)^T\sum_i^{-1}(X-\mu_i))}{(2\pi)^{d/2}|\sum_i|^{1/2}} \tag{3}$$

因此，给定两个混合高斯分布的分量，用于混合高斯分布的脊曲线 (RC) 定义为：

$$A_{x1}B_{x2}-B_{x1}A_{x2}=0 \tag{4}$$

式中：

$$A_{xi}=\frac{\partial}{\partial x_i}(-\frac{1}{2}(X-\mu_1)\sum_1^{-1}(X-\mu_1)^T) \tag{5}$$

$$B_{xi}=\frac{\partial}{\partial x_i}(-\frac{1}{2}(X-\mu_2)\sum_2^{-1}(X-\mu_2)^T) \tag{6}$$

因此，两个高斯分布之间的基于重叠度的聚类数目判断指标（GO）计算公式为：

$$\mathrm{GO}(G_1,G_2)=\begin{cases}1, & P_{\mathrm{peaks}}(x)=1\\ \dfrac{p(X_{\mathrm{sadelle}})}{p(X_{\mathrm{submax}})}, & P_{\mathrm{peaks}}(x)=2\end{cases} \tag{7}$$

式中：X_{sadelle} 为 PDF 中的鞍点；X_{Submax} 是较低的峰值；GO 计算为鞍点的 PDF 与较低峰值的 PDF 之比。这样做是因为鞍点的 PDF 与混合因子 λ_i 有关。根据上面的等式，整个数据集的重叠度计算为：

$$\mathrm{GO}=\frac{1}{k^2}\sum_{j=1}^{k}\sum_{i=1}^{k}\mathrm{GO}(G_i,G_j) \tag{8}$$

式中：G_i、G_j 代表遍历所有子簇的符号代表，k 为所有子簇个数。

应该注意的是，从第一个平均值开始，沿着曲线一直到另一个平均值，过程中最小值点是鞍点。如此定义的转基因程度描述了两个集群之间的重叠程度，从而确定两个子集群

是否应分为同一集群或两个不同的集群。GO ⊂ (0,1)，越接近 1，两个聚类之间的重叠程度越高，两个聚类合二为一的可能性越大，证明聚类无效。相反，GO 值越小，聚类效果越好。因此，当簇数更接近真实值时，GO 值的变化会比之前更大，凭此可以进一步通过 GO 来确定最优簇数。

2 仿真验证

2.1 数据集与基准模型

本文选择了 4 种实验结果相对较好的方法作为基线方法。它们是 BIC 值法 [8]、轮廓系数法、SSE 和 AP 算法 [9]。其中 SSE 法、BIC 法和轮廓系数法是通过计算模型类内信息来判断聚类数量的方法。SSE 的原理是计算拟合数据与原始数据对应点误差的平方和，平方和越小越好。BIC 法和轮廓系数法分别通过观察最小值和拐点值来判断最优簇数 k。而 AP 算法是一种完整的聚类算法，它不需要事先定义类的个数，而是在迭代过程中不断寻找合适的聚类中心，并在数据点之间自动识别类中心的位置和个数，使所有数据点与最近的类代表点的相似度之和最大化。

本文选择的数据集为多点聚类数据集，旨在研究聚类的基本基准。数据集由未标记的数据组成，每个数据集由 5000 个二维坐标点组成，便于重点研究重叠度。本文选择了 4 个未标记的数据集，编号为 s1~s4。编号为 s1-s4，每个数据集包括 15 个类，重叠程度依次变高。

2.2 实验结果

数据集 s1 的 GO 值变化曲线图如图 2 所示。可以清晰地看出在达到正确聚类个数时，GO 值下降明显。通过对 GO 值变化趋势的判断，可以清晰得出聚类数目值和聚类结果。

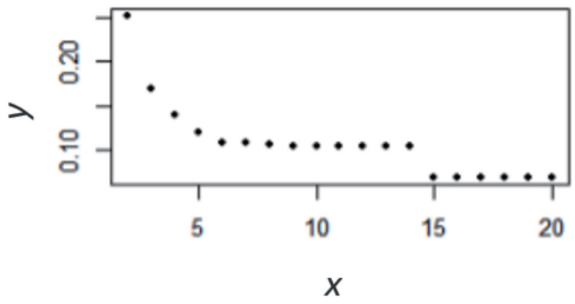

图 2 s1 数据集 GO 值变化展示

对比试验结果如表 1 所示。从表 1 中可以看出，每种方法在 10 次实验下预测聚类 k 个数的准确率。对于高重叠的未标记数据集，实验重叠模型在所有实验中都优于基线方法。当数据重叠非常低时，GO 产生的结果与某些算法一样好。数据重叠越高，GO 的效果越好。通常，GO 具有更高的预测精度，预测结果的值范围更接近真实值。

表 1 实验结果

数据集	BIC 值		轮廓系数		误差平方和 SSE		AP 算法		GO 值	
	预测均值	准确率	预测均值	准确率	预测均值	准确率	预测均值	准确率	预测均值	准确率
s1	16.0	9/10	15.0	10/10	18.5	3/10	15.9	2/10	15.0	**10/10**
s2	17.5	7/10	18.1	5/10	13.9	2/10	40.1	0/10	14.8	**9/10**
s3	17.0	3/10	16.4	9/10	14.1	4/10	53.5	0/10	15.4	**7/10**
s4	20.2	0/10	13.8	5/10	17.9	3/10	50.9	0/10	16.1	**7/10**

2.3 消融实验

首先将算法分为预聚类部分和计算参数部分。在预聚类部分使用的算法是一种改进的基于 GMM 的聚类算法，计算参数部分是计算重叠。在消融实验中，分别使用基于 K-means 的聚类算法和计算 AIC 值代替 GMM 和重叠度计算这两部分。实验结果如表 2 所示，将 GMM 部分的模型替换标记为 GO-G，将重叠部分的模型替换标记为 GO-O。可以看出，没有重叠部分的模型效果很差，这充分证明了重叠部分模型的重要性。

表 2 消融实验结果

缺失值 / 数据集		s1		s2		s3		s4	
GO-G	GO-O	预测均值	准确率	预测均值	准确率	预测均值	准确率	预测均值	准确率
×	√	15.0	10/10	14.5	9/10	16.2	6/10	15.6	7/10
√	×	15.0	10/10	15.5	9/10	16.9	5/10	20.1	1/10

3 结语

聚类算法以其实用性、鲁棒性和较好的性能成为数据挖掘领域的研究热点。本文从重叠的角度构建了具有强聚类能力和高执行效率的聚类判断指标。并从数学角度验证了重叠的意义，得出重叠在描述重叠现象中具有不可或缺的地位的结论。在此基础上，提出了一种基于 GMM 聚类改进的聚类算法和一种基于重叠度的指标 GO。该指标可以有效地帮助研究者确定数据集的聚类中心个数。与现有的用于确定聚类中心数量的算法相比，该算法具有更少的超参数，更加实用。它的结果更准确，误差范围更小。

参考文献：

[1] DING J, TAROKH V, YANG J Y. Bridging AIC and BIC: a new criterion for autoregression[J]. IEEE transactions on information theory, 2018, 64(6):4024-4043.

[2] SANTOS A ,PAULA H. Microservice decomposition and evaluation using dependency graph and silhouette coeffic ient[C]// SBCARS'21: Proceedings of the 15th Brazilian Symposium on Software Components, Architectures, and Reuse. New York: ACM, 2021: 51-60.

基于 ICOA 优化 XGBoost 的光伏阵列故障诊断方法

董建业[1] 李红月[1]
DONG Jianye LI Hongyue

摘 要 基于 XGBoost 模型是识别光伏阵列故障类型的一项重要技术，由于 XGBoost 参数初始化设定的主观性和随机性，导致模型在训练和学习时准确度低。针对传统 XGBoost 算法的不足，文章提出了一种优化长鼻浣熊算法 (ICOA) 优化 XGBoost 初始参数的故障诊断方法。采用 Logistic-Tent 混沌映射、自适应权重因子、Levy 飞行和透镜成像学习策略来优化长鼻浣熊算法 (COA)，降低了算法易陷入局部极值点的可能性。利用 ICOA 算法对 XGBoost 分类算法进行优化，构建 ICOA-XGBoost 光伏阵列故障诊断模型，并与其他优化算法模型进行实例对比分析，验证了改进后的算法在识别光伏阵列故障类别上的有效性和实用性。

关键词 光伏阵列；故障诊断；改进长鼻浣熊算法；XGBoost 算法

doi：10.3969/j.issn.1672-9528.2024.11.012

0 引言

随着能源多元化发展的持续推进，分布式光伏发电得到广泛发展。但光伏阵列易发生老化、局部遮荫、开路和短路等故障，因此提高故障诊断准确率，不仅能提高发电系统的可靠性，而且能有效降低运维成本 [1]。

目前，很多机器学习分类算法应用于光伏阵列故障诊断中。吴亚钧等人通过加入 Levy 飞行策略和黄金分割系数改进蜣螂算法，该方法寻优速度快且稳定 [2]。王一鸣等人提出麻雀算法优化支持向量机算法，有效解决支持向量机算法收敛速度慢的问题 [3]。罗凯元等人引入融合佳点集、分段自适应逃逸能量机制和黄金正弦策略进一步优化哈里斯鹰算法，显著提高故障诊断准确率 [4]。

综合上述分析，本文提出一种改进长鼻浣熊算法（coati optimization algorithem，ICOA）优化 XGBoost 算法的初始参数，有效提高模型诊断准确率。首先，提出优化策略并进行改进；然后，将 ICOA 算法与 COA、WOA 和 DBO 算法进

1. 安徽理工大学电气与信息工程学院 安徽淮南 232001

[3] DUAN J Y, GUO L L. Variable-length subsequence clustering in time series[J].IEEE transactions on knowledge and data engineering. 2022, 34(2): 983-995.

[4] CARON M, BOJANOWSKI P, JOULIN A, et al. Deep clustering for unsupervised learning of visual features[C]// Computer Vision-ECCV 2018. Berlin: Springer, 2018: 139-156.

[5] CHEN F W, PAN S R, JIANG J , et al. DAGCN: dual attention graph convolutional networks[C]//2019 International Joint Conference on Neural Networks. Piscataway, NJ : IEEE, 2019: 1-8.

[6] BONCHI F, GIONIS A, UKKONEN A. Overlapping correlation clustering[J]. Knowledge and information systems, 2013, 35: 1-32.

[7] WANG X S, LI L J, CHENG Y H. An overlapping module identification method in protein-protein interaction networks[J]. BMC bioinformatics, 2012, 13(S-7):S4.

[8] 陈国艳 , 张颖 , 梁德群 . 基于 BIC 准则的图像分割算法 [J]. 辽宁工程技术大学学报 (自然科学版), 2016,35(11):1359-1362.

[9] 赖健琼 . 自适应 AP 聚类算法研究 [J]. 计算机时代 , 2022 (4): 38-42.

【作者简介】

张娟（1981—），通讯作者（zhangjuan@mail.taiji.com.cn），女，天津人，本科，工程师，研究方向：数据挖掘、数据中台。

李娜（1981—），女，江西宜丰人，本科，助理研究员，研究方向：数字经济。

（收稿日期：2024-08-09）