基于自适应谱聚类的海量数据离群点挖掘方法

李 婷[1] 刘 凯[1]
LI Ting LIU Kai

摘 要 为降低海量数据离群点检测的误判概率，文章提出基于自适应谱聚类的海量数据离群点挖掘方法研究。先对海量数据进行降维，利用滑动窗口和最小二乘法得到数据集的映射投影，得到数据集的特征向量。然后通过谱聚类算法对序列离群点数据特征进行聚类，并通过 K 最近邻法得到数据特征矩阵。最后计算海量数据的离群因子，并与阈值比较，完成海量数据离群点挖掘。实验结果表明，所设计方法能有效检测出海量数据中的离群点，应用效果较好。

关键词 自适应谱聚类；海量数据；滑动窗口；最小二乘法；离群因子

doi：10.3969/j.issn.1672-9528.2025.08.041

0 引言

在海量数据集中存在特征和行为异常的离群点，离群点的存在使数据集信息发生变化 [1]。传统的聚类算法在处理数据时，往往难以有效识别出离群点，但数据点蕴含着重要信息。因此，对海量数据的离群点挖掘已经成为研究的热门。

近年来众多学者开展研究，如朱华等人 [2] 利用 CART 决策树和类间中心距离分裂准则，结合空间局部偏离因子，能准确地检测出数据离群点。该方法未对高维数据进行预处理，且对不平衡数据的处理能力较差，当数据分布不均匀时容易出现离群点的误判的现象，应用效果不佳。再如周燕等人 [3] 通过 Spark-MML 聚类算法和兴趣度约束与支持度自适应策略，该方法提高了异常数据挖掘的准确率，但其过于依赖数据的关联规则，且未对数据维度进行处理，且数据冗余度较高，离群点挖掘效果不佳。张忠平等人 [4] 通过快速密度峰值聚类离群因子和 KD-Tree 索引数据结构，自动选取聚类中心，提高了算法效率。该算法同样未对数据维度进行处理，且其受到数据分布和密度峰值影响，若数据密度分布不均，可能导致正常数据被误判为离群点。在该背景下，本文提出基于自适应谱聚类的海量数据离群点挖掘方法。

1 海量数据离群点挖掘设计

1.1 海量数据降维处理

海量数据的高维性会增加数据离群点挖掘的复杂度 [5-6]。因此，本文先对海量数据进行降维处理。为准确量化海量数据的不确定性，本次引入信息熵作为度量标准，信息熵表达式为：

$$S(x)=\sum_{x\in S}P(x)\log(P(x)) \tag{1}$$

式中：$S(x)$ 表示数据信息熵；$P(x)$ 表示概率函数。

信息熵能表示海量数据中各个变量具有不确定性，且这种不确定性会随着信息熵数值的变化而变化 [7]。接下来，计算海量数据的信息熵增量，表达式为：

$$\Delta(x)=S(A)-\sum_{k=1,2}\left(\frac{|C_k|}{D_0}(S(C_k))\right) \tag{2}$$

式中：$\Delta(x)$ 表示信息熵增量；A 表示数据子集；C_k 表示分割后的数据；D_0 表示海量数据集合。

信息熵增量值反映海量数据属性间的不确定性变化，利用信息熵增量能量化数据属性权重，实现数据降维。由此，得到新的数据集 $D_{\Delta(x)}$。

1.2 数据特征向量提取

完成数据降维可以降低数据的复杂度，同时保留数据中的关键信息和结构特征，使后续的数据处理和分析工作更加高效且准确。通过滑动窗口，将降维后的数据沿窗口单位滑动，获得不等长序列二者之间的子序列。计算子序列在对应窗口内的滑动相似度为：

$$\lambda=1-\frac{\left(L_{ij}\left(P_i,Z(P_i)\right)\right)}{L_{\max}} \tag{3}$$

式中：λ 表示数据序列之间的滑动相似度；Z 表示获取的子序列窗口；L 表示数据序列与滑动序列之间的距离；$L_{\max}$ 表示最大距离。基于获取的相似度，利用最小二乘法提取数据离群点的特征。

设海量数据集 $D_{\Delta(x)}$ 中存在 m 对数据样本，数据集的映射投影公式为：

$$\begin{cases}a=\lambda A\chi\\ b=\lambda B\tau\end{cases} \tag{4}$$

式中：a 与 b 分别表示数据集在低维空间中的分布。

根据式（4）计算结果，构建海量数据集的最大化函数准则，旨在通过最大化不确定数据集的某种度量，优化数据

1. 郑州工商学院信息工程学院 河南郑州 450000

的特征表示，具体用公式表示为：

$$J_{\text{pls}}(\chi,\tau)=\frac{\left(\chi^{T}\boldsymbol{F}_{a,b}\tau\right)^{2}}{\left[\chi^{T}\chi\right]\left[\tau^{T}\tau\right]} \tag{5}$$

式中：$J_{\text{pls}}(\chi,\tau)$表示海量数据集准则；$\boldsymbol{F}_{a,b}$ 表示协方差矩阵；T 表示函数系数。

根据该准则制定数据正交约束条件，同时引入拉格朗日乘子，将海量数据的特征提取问题转化为方程，具体公式为：

$$\begin{cases}\boldsymbol{F}_{a,b}\boldsymbol{F}_{a,b}\chi=\delta^{2}\chi\\ \boldsymbol{F}_{a,b}\boldsymbol{F}_{a,b}\chi=\delta^{2}\tau\end{cases} \tag{6}$$

式中：$\delta^{2}\chi$与$\delta^{2}\tau$分别表示转化后数据特征方程。

综上，通过对该方程的求解，得到数据集的特征向量。

1.3 基于自适应谱对构建数据特征矩阵

提取数据集的特征向量是数据集中每个样本或数据点的关键属性表示，能够概括和反映数据的本质特征。接下来，基于自适应谱对构建数据特征矩阵，以整合数据集的特征向量到一个矩阵中，便于后续的离群点检测 [8-9]。本次将海量数据集中的每个数据视为图像中的顶点，并根据 1.2 得到的特征向量建立海量数据之间的关系矩阵。通过自适应谱聚类对聚类序列离群点数据特征进行聚类，具体公式为：

$$\begin{cases}b=\{b_1,b_2,\cdots,b_m\}\\ |b_1|\geq|b_2|\geq\cdots|b_m|\end{cases} \tag{7}$$

式中：b_m 表示 m 个聚类离群点的聚类结果 [10]。

根据聚类结果设置聚类参数，从而得到数据的大小聚类边界，用公式表示为：

$$\begin{cases}b_{\max}=\{b_i|i\leq\beta\}\\ b_{\min}=\{b_i|i\geq\beta\}\end{cases} \tag{8}$$

式中：β 表示聚类系数。

选取一组数据组成数据集，并对剩余数据进行聚类，计算任意类到数据点的平均距离，以评估聚类效果，其表达式为：

$$D_{J_{\text{pls}}(\chi,\tau)}=\frac{1}{n}\sum_{i=1}^{m}\left|b_j-b_z\right| \tag{9}$$

式中：$D_{J_{pls}(\chi,\tau)}$表示数据点平均距离；n 表示数据总数；b_z 表示剩余数据数量。

选取欧氏距离最大的点作为新的聚类中心，并将剩余的数据点分配到这个新的聚类集合中。接着，选择当前集合中欧氏距离计算值最大的样本点作为新的中心，不断循环这一过程，直到不再产生新的聚类中心。最后，计算所有数据的相似性距离矩阵，并将其作为谱聚类的输入相似性矩阵，具体表示为：

$$\boldsymbol{R}=\begin{bmatrix}\boldsymbol{R}_{11}&\boldsymbol{R}_{12}&\cdots&\boldsymbol{R}_{1b}\\ \boldsymbol{R}_{21}&\boldsymbol{R}_{2}&\cdots&\boldsymbol{R}_{2b}\\ \vdots&\vdots&\ddots&\vdots\\ \boldsymbol{R}_{a1}&\boldsymbol{R}_{a2}&\cdots&\boldsymbol{R}_{ab}\end{bmatrix}_{D_{J_{\text{pls}}(\chi,\tau)}} \tag{10}$$

式中：$\boldsymbol{R}$ 表示相似性矩阵。

通过 K 最近邻法遍历所有的数据点，并选择每个数据最近的 k 各点作为近邻。构建邻接矩阵 $\boldsymbol{\delta}$，进而构建出度矩阵，用公式表示为：

$$\begin{cases}\boldsymbol{E}=\text{diag}\left(e_1,e_2,e_3,\cdots e_n\right)\\ e_1=\sum_{i=1}^{n}\boldsymbol{\delta}_{ab}\end{cases} \tag{11}$$

式中：$\boldsymbol{E}$ 表示度矩阵；e_1 表示邻接矩阵的每行和。

最后，计算邻接矩阵与度矩阵之间的差值得到拉普拉斯矩阵，对拉普拉斯矩阵进行标准化处理，求解拉普拉斯矩阵的前 k 个最大特征值及其对应的特征向量，并将其组合成一个新的矩阵，进行标准化处理，从而得到一个特征矩阵，其可表示为：

$$\boldsymbol{U}=\begin{bmatrix}\boldsymbol{U}_{11}&\boldsymbol{U}_{12}&\cdots&\boldsymbol{U}_{1b}\\ \boldsymbol{U}_{21}&\boldsymbol{U}_{22}&\cdots&\boldsymbol{U}_{2b}\\ \vdots&\vdots&\ddots&\vdots\\ \boldsymbol{U}_{a1}&\boldsymbol{U}_{a2}&\cdots&\boldsymbol{U}_{ab}\end{bmatrix}_{R} \tag{12}$$

1.4 海量数据离群点挖掘

数据特征矩阵作为数据的高维表示，汇总了数据集中每个样本的关键特征信息。本次以此为基础，开展海量数据离群点挖掘检测。离群点作为数据集中与整体特征显著偏离的个体，其偏离程度可以通过离群因子来量化衡量，离群因子的计算公式为：

$$\boldsymbol{U}_i=\frac{\lambda_i}{\sum_{i=1}^{M}\lambda_i}\boldsymbol{U} \tag{13}$$

式中：$\boldsymbol{U}_i$ 表示数据向量方差。

为有效识别离群点，根据数据的平均距离设定一个阈值，阈值的计算公式为：

$$\max D=\max_{i\geq1,j\geq1,i\neq j}D\left(N_i,N_j\right) \tag{14}$$

式中：N_i 与 N_j 分别表示聚类的第 i 与第 j 个类。

将离群因子与式（14）计算得到的阈值 U_i 进行比较，若离群因子超过该阈值，那么该数据点就被认定为离群点。

2 实验

2.1 实验环境

为验证所提方法性能，设置实验。由于数据规模较大，因此实验采用英特尔 ® 酷睿 ™i5-4460 处理器，并配备了 6 MB 的 L3 高速缓存，用于提高数据访问速度。此外，实验设置了一个主节点，内存容量为 8.00 GB。

本次实验选用 4 个不同的数据集，数据集的具体设置如表 1 所示。

表 1 数据集

数据集编号	数据量 /bit	维度	离群点占比 /%
数据集 1	152 487	150	19
数据集 2	239 617	230	30
数据集 3	112 500	120	20
数据集 4	205 981	200	23

4 个数据集中的数据分布均为不平衡数据。其中，数据集 1 的离群点挖掘范围为 x 轴 1~3，y 轴 3~6，数据集 2 的挖掘范围为 x 轴 4~6，y 轴 5~6，数据集 3 的挖掘范围为 x 轴 4~6，y 轴 3~4，数据集 4 的挖掘范围为 x 轴 1~6，y 轴 1~2。

在测试程序导入 4 个数据集，每个数据集的离群点挖掘周期设置为 5，挖掘时间控制在 0.1~0.2 s，信噪比的范围设置为 0~15 dB，采样频率设置为 13 kHz，数据集的带宽为 800 Hz。在此基础上结合谱聚类算法，针对当前海量数据离群点的挖掘需求设置挖掘测试指标以及相对应的初始参数和实际设置参数，具体如表 2 所示。

表 2 挖掘测试指标及参数

指标	初始参数	实际参数
异常熵	0.41	0.37
挖掘差值	2.0	1.09
挖掘特征值	15.98	16.17
基函数	+11.25	+13.25

根据表 2 设置海量数据集离群点挖掘测试指标及参数，调整数据挖掘的处理环境。为验证所设计方法的有效性，对自适应谱聚类的准确性进行测试，聚类准确性公式为：

$$A=\frac{\sum_{i=1}^{M}\delta\left(z_i,\mathrm{map}\left(b_i\right)\right)}{M} \tag{15}$$

式中：A 表示聚类准确率；δ 表示判别公式；z_i 表示数据真实标签；b_i 表示聚类算法得到的数据标签；M 表示数据点整体数量。

为验证所设计方法在海量数据聚类方面的有效性，在不同聚类数量下进行实验，计算这些不同条件下的聚类准确率，得到结果如表 3 所示。

表 3 聚类数量与准确率的关系

聚类数量	聚类准确率 /%
5 000	93.17
10 000	93.28
15 000	94.56
20 000	97.99

根据表 3 可知，本文方法对数据进行聚类，准确率均高于 92%，表明所设计方法可通过不同聚类数量对数据进行自适应调整。当聚类数量为 20 000 时，海量数据的聚类准确率最高，为 98.13%。因此，实验将聚类的参数设置为 20 000。

2.2 海量数据挖掘结果

为验证本文方法对海量数据离群点的挖掘有效性，将挖掘前后的数据分布情况进行对比分析，结果如图 1 所示。图 1 中，实心圆圈表示海量数据，空心圆圈表示数据离群点。分析图 1 可知，所提方法应用前，数据呈无规则分布，并无显著特点，而应用所提方法挖掘后的数据被分为不同的簇，且数据离群点可被直接识别。由此验证了所设计的数据挖掘方法能够将数据聚类为不同的簇，实现了数据离群点的挖掘。

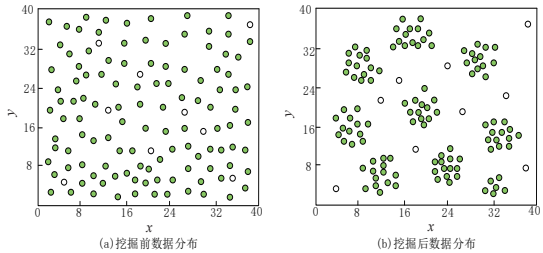

图 1 挖掘前后对比情况

2.3 海量数据挖掘性能对比

使用另外 3 种算法与所提方法一同进行离群点挖掘，得到结果如图 2 所示。

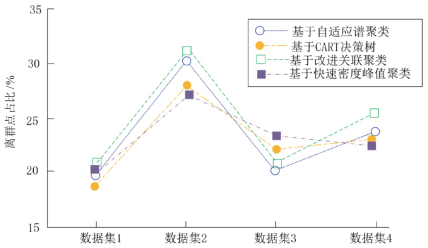

图 2 数据挖掘对比结果

根据图 2 可以看出，除了本文所设计的基于自适应谱聚类算法，另外 3 种算法的离群点占比结果与表 1 所示数据集中完全不一致，基于 CART 决策树的最大检测误差为 2%，基于改进关联聚类的最大检测误差为 2.4%，基于快速密度峰值聚类的最大检测误差为 3%。由此证明，所提方法离群点检测效果更优。为进一步验证所设计的离群点挖掘算法的性能，将所设计的方法与另外 3 种方法进行对比，得到的挖掘效果如图 3 所示。

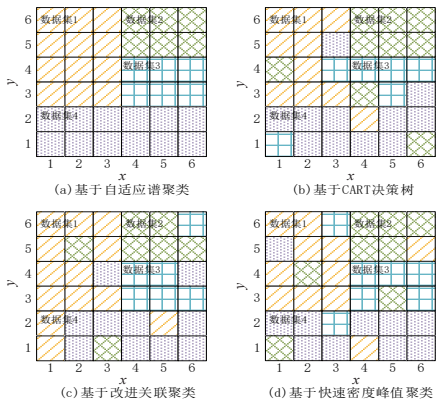

图 3 挖掘效果

根据图 3 可以看出，本文所设计的离群点挖掘方法，四个数据集在 x 轴与 y 轴上的分布与预先设定的范围一致，而对比的 3 种方法挖掘效果均与设定不一致，各数据集之间存在混淆。由此可表明，本文算法的挖掘结果能够满足预期要求，达到了理想挖掘状态，应用效果更优。

（下转第 179 页）

分析，并在此基础上成功获取了包含多种类型信息的电力多元数据集。随后，为了实现对这些数据中潜在异常值的准确检测与分析，引入了模糊 C 均值（FCM）聚类算法。在仿真实验部分，本研究将所提出的检测方法与现有的几种传统异常检测技术进行了对比分析。实验结果表明，本文所设计的方法不仅能够全面地检测出电力数据中的异常值，而且在检测的准确性方面也表现出了显著的优势。此外，该方法在处理大规模数据集时，也展现出了较高的检测效率。

展望未来，本研究计划继续深入探索电力数据的内在特性以及电网的动态行为规律。通过进一步的研究，期望能够对现有的异常检测方法进行改进和优化，从而提升其泛化能力和实用性。最终，希望本研究的成果能够为我国智能电网的建设和发展提供有力的技术支持，为智能电网的稳定运行和高效管理贡献更多的力量。

参考文献：

[1] 常荣 , 徐敏 . 基于改进 K-Means 和 DNN 算法的电力数据异常检测 [J]. 南京理工大学学报 ,2023,47(6):790-796.

[2] 林昱奂 , 胡嘉铭 , 戴伟力 , 等 . 基于改进 K- 均值联合 SVDD 的电力数据异常检测方法 [J]. 电力电容器与无功补偿 , 2023, 44(5): 99-107.

[3] 林卫伟 , 白向阳 , 孔军 . 基于伪标签的无监督电力数据异常检测框架 [J]. 计算机仿真 ,2024,41(2):131-136.

[4] 王文博 , 刘绚 , 张博 , 等 . 基于协议特征的电力工控网络流量异常行为检测方法 [J]. 电力系统自动化 ,2023,47(2):137-145.

[5] 王文博 , 刘绚 , 林海 , 等 . 基于深度学习的电力工控流量应用层报文异常检测 [J]. 电力系统自动化 ,2023,47(11):69-76.

[6] 徐子东 , 张镇勇 . 基于不平衡高维数据特征重要性的电力系统异常检测方法 [J]. 控制工程 ,2024,31(11):2029-2035.

[7] 薛晓慧 , 张文 , 张静 , 等 . 基于二次聚类的充电桩执行电价异常检测方法 [J]. 电信科学 ,2025,41(1):184-190.

[8] 李凯 , 靳书栋 , 刘宏志 , 等 . 基于 IWOA-ELM-AE 的电力资产信息管理系统异常数据检测方法 [J]. 沈阳工业大学学报 , 2024, 46(3):255-262.

[9] 王文森 , 杨晓西 , 刘阳 , 等 . 基于层次聚类分析的变压器油中溶解气体在线监测数据异常检测 [J]. 高压电器 , 2023, 59(1):142-147.

[10] 侯立 , 王健 . 改进密度峰值聚类的多源数据异常值检测算法 [J]. 计算机仿真 , 2024, 41(6): 565-569.

【作者简介】

张瑞（1978—），女，甘肃张掖人，本科，工程师，研究方向：市场营销。

王婷婷（1991—），女，甘肃张掖人，本科，工程师，研究方向：电气工程及其自动化。

辛亚峰（1992—），男，甘肃武威人，本科，工程师，研究方向：电气工程及其自动化，email: yuxq211@163.com。

（收稿日期：2025-04-15 修回日期：2025-08-12）

（上接第 175 页）

3 结语

为了提升海量数据处理与分析的效率与准确性，本文开展基于自适应谱聚类的海量数据离群点挖掘方法研究，其通过引入信息熵有效减少了海量数据冗余和相关性带来的干扰。利用谱聚类算法提高了海量数据聚类的准确性。实验选用了 4 个不同的数据集，通过对比离群点挖掘前后的数据分布情况，验证了所设计的数据挖掘方法对数据离群点的挖掘能力。实验结果表明，所设计算法应用后，其检测的离群点占比结果与真实情况完全一致，能够满足预期要求，达到理想的挖掘状态，应用效果较好。

参考文献：

[1] 庄巧蕙 . 异质信息网络快速离群点数据智能挖掘算法 [J]. 黑龙江工业学院学报 (综合版),2024,24(10):97-101.

[2] 朱华 , 乔勇进 , 董国钢 . 基于 CART 决策树的分布式数据离群点检测算法 [J]. 现代电子技术 ,2024,47(16):157-162.

[3] 周燕 , 肖莉 . 基于改进关联聚类算法的网络异常数据挖掘 [J]. 计算机工程与设计 ,2023,44(1):108-115.

[4] 张忠平 , 李森 , 刘伟雄 , 等 . 基于快速密度峰值聚类离群因子的离群点检测算法 [J]. 通信学报 ,2022,43(10):186-195.

[5] 康耀龙 , 冯丽露 , 张景安 , 等 . 基于谱聚类的不确定数据集中快速离群点挖掘算法 [J]. 吉林大学学报 (工学版), 2023, 53(4):1181-1186.

[6] 李春燕 . 基于谱聚类算法的人力资源数据集离群点快速挖掘方法 [J]. 信息与电脑 (理论版),2023,35(23):50-52.

[7] 王彩霞 , 陶健 , 舒升 . 基于机器学习的聚类序列离群点数据挖掘算法 [J]. 通化师范学院学报 ,2024,45(8):28-34.

[8] 杨志强 , 冯山 , 尹伊 , 等 . 一种多因素融合的高效离群点检测方法 [J]. 山东大学学报 (理学版), 2024,59(8): 118-126.

[9] 何旭 , 邓安生 , 葛小龙 . 基于局部相对密度的离群点检测算法 [J]. 计算机应用与软件 , 2024, 41 (12): 296-302.

[10] 周玉 , 夏浩 , 岳学震 , 等 . 基于改进 K-means 的局部离群点检测方法 [J]. 工程科学与技术 , 2024,56(4): 66-77.

【作者简介】

李婷（1996—），女，山西长治人，硕士，助教，研究方向：机器学习、数据挖掘。

（收稿日期：2025-03-26 修回日期：2025-07-31）