[文章编号] 1007－7405(2024)01－0024－08 DOI:10.19715/j.jmuzr.2024.01.04

基于参数自适应DBSCAN算法的浮标位置数据异常检测

章新亮，肖 虹，周世波

（集美大学航海学院，福建 厦门 361021）

[摘要] 针对遥测遥控系统采集浮标位置数据时易受外在因素的干扰，提出了一种K近邻优化的参数自适应DBSCAN算法，来检测浮标位置数据中的异常点。通过分析数据集的分布特性生成最优邻域距离值ε和邻域内样本点数量MinPts列表，引入卡林斯基-哈拉巴斯指数对列表中的参数进行评分，将最高评分对应的参数作为最优参数，实现DBSCAN算法的自适应聚类。实验结果表明，新算法能够自适应选择最优参数，对浮标遥测位置数据的异常点进行有效检测。

[关键词] 浮标位置；异常检测；遥测遥控系统；DBSCAN算法；K近邻算法；CH指数

[中图分类号] U 644.35

Buoy Position Data Abnormaly Detection Based on Parameter Adaptive DBSCAN Algorithm

ZHANG Xingliang, XIAO Hong, ZHOU Shibo

(Navigation College, Jimei University, Xiamen 361021, China)

Abstract: In the process of using the telemetry and remote-control system to collect data, it is easy to be disturbed by external factors and generate abnormal location data. To address this problem, a K-nearest neighbor optimized parameter adaptive DBSCAN algorithm is proposed to detect the anomalies in buoy position data. The algorithm proposed generates a list of optimal distance values ε in adjacent waters and the number of sample points MinPts through the analysis of the distribution characteristics of the dataset, and the introduction of the Calinsky-Harabas index to score the parameters in the list, and the parameter corresponding to the highest score is used as the optimal parameter to realize the adaptive clustering of DBSCAN algorithm. The experimental results show that the proposed algorithm can adaptively select the optimal parameters and realize the detection of abnormal buoy telemetry position data.

Keywords: buoy position; abnormal detection; telemetry and remote control system; DBSCAN algorithm; K-nearest neighbor algorithm; CH index

[收稿日期] 2022－06－06

[基金项目] 福建省自然科学基金项目（2020J01658）；船舶辅助导航技术国家地方联合工程研究中心开放课题（HHXY2020002）；集美大学博士启动基金项目（ZQ2019012）

[作者简介] 章新亮（1997—），男，硕士生，从事海事大数据分析研究。通信作者：肖虹（1979—），女，硕士，讲师，从事海事大数据分析及国际航运管理研究。

0　引言

浮标是引导船舶航行的人工标志，对保障船舶的航行安全有着重要意义。在浮标遥测数据采集、传输与存储过程中，由于电磁干扰、人工输入等自然因素与人为因素的影响，数据集中会存在部分异常数据[1]，这会影响到浮标数据的分析结果，因此，如何有效剔除浮标遥测数据集中的异常数据，是进行浮标漂移位置研究的关键。

针对此类剔除异常数据的问题，目前主要采用统计分布与数据挖掘这两类方法。Zhao 等[2]提出一种非线性的异常检测方法，通过核学习与稳健回归分析实现了高光谱数据的异常检测。刘首华等[3]提出一种实用的浮标数据异常值质控方法，实现了海洋浮标异常值的检测，成功率达到了 86.2%。Yu 等[4]提出一种检测多元时间序列中异常数据的方法，通过斯皮尔曼相关系数获得数据序列方向，对比分离出异常点，适用于工业生产中的时间序列数据。罗一迪等[5]提出一种基于滑动窗口和自回归移动平均模型（ARMA）的异常值检测算法，实现了浮标数据异常检测，准确率在 60% 以上。张宇等[6]提出一种基于时序相关性分析的浮标异常数据识别算法，实现了渐变异常数据的识别并处理。卢勇夺等[7]结合极值法则、莱维特法则以及局部法则识别异常值数据，实现了海洋浮标的异常数据检测。Das 等[8]运用向量自回归的方法进行预测，并通过模糊综合评价法判断传感器数据节点的异常情况。尽管采用数学统计的方法运行速度较快，准确率较高，但不适用于离散程度大且呈现分段式变化的数据集。

在数据挖掘技术日趋成熟的条件下，基于数据挖掘方法的异常数据识别逐渐成为研究的热点[9-17]。K-means 算法是最经典的一种聚类算法，实现过程简单，运行速度快，但无法识别异常数据，所以有学者在此基础上进行了改进[9-12]。Amin 等[10]提出一种 PSO-K-means 算法，有良好的鲁棒性与效率，但精确度还有待提升；蒋华等[11]提出一种结合密度函数的 K-means 算法检测海洋 Argo 浮标数据，与传统 K-means 算法相比，在迭代次数下降的基础上，聚类准确率提升了 12.8%；高书强等[12]提出一种改进的 Mini-Batch K-Means 算法，通过对数据集进行剪枝处理，再采用杰卡德相似系数值进行对比，确定异常点。

而另一种异常数据识别方法是 DBSCAN 算法[13]，通过设定的邻域距离值 ε 与邻域内最小样本点数 MinPts，能够识别出输入数据集中各种形状的簇，并分离出不属于簇的异常数据。Hossein 等[14]将经典的 DBSCAN 算法应用于网络数据异常检测，通过系数相关性的方法，实现了正常数据与异常数据分离。马良玉等[15]基于 DBSCAN 算法与降噪自编码器（SDAE）结合，提出一种识别风电机组异常数据的方法，实现了风机异常工况示警，但 DBSCAN 算法的输入参数需要人工调整。郑玉巧等[16]在此基础上进行了改进，提出一种基于四分位法确定 DBSCAN 算法参数的 QM-DBSCAN 算法，该算法能够自适应生成参数，在风机异常数据识别中效果较好，但该算法不具备普适性。Jain 等[17]提出一种改进的 DBSCAN 算法，能够检测出季节性数据中的局部异常与全局异常，为时间序列数据的异常检测提供了新的方向。基于改进 K-means 算法识别异常数据的一个不足是 K-means 算法只适用于“球状”数据集，对形状多样的数据集，效果较差，而 DBSCAN 算法需要输入邻域距离值 ε 与邻域内最小样本点数 MinPts 两个参数，如果参数设置不当，可能无法有效识别异常数据。

基于此，本文提出一种参数自适应的 DBSCAN 算法，以期实现对浮标遥测位置数据中异常数据的识别与剔除。

1　基于 K 近邻的参数自适应 DBSCAN 算法

尽管 DBSCAN 算法能够根据密度分布变化识别出异常数据，但是聚类结果受输入邻域距离值 ε 与 MinPts 影响较大，因此，为了减少人工输入参数造成的误差，本文采用 K 近邻方法生成 ε 与 MinPts 参数列表，并结合卡林斯基-哈拉巴斯（Calinski-Harabasz，CH）指数[19]进行寻优，实现 DBSCAN 算法参数的自适应。

1.1 生成 ε 与 MinPts 参数列表

参考文献［18］的 K 平均最近邻法与数学期望法，生成 ε 与 MinPts 参数候选列表的步骤如下：

步骤 1）根据初始数据集生成距离分布矩阵 $\boldsymbol{D}_{n\times n}$，矩阵中的元素表示为 x_{ij}（$i \in 1,2,\cdots,n$；$j \in 1,2,\cdots,n$），其中 i 为行数，j 为列数；

步骤 2）将距离矩阵中每行的数据按照从小到大的顺序进行排列，记为 $\boldsymbol{D}_{\mathrm{k}}$，则第一列数据全为 0；

步骤 3）取 $\boldsymbol{D}_{\mathrm{k}}$ 中每列元素的平均值生成 $\{\bar{x}_{i1},\bar{x}_{i2},\cdots,\bar{x}_{in}\}$ 集合，并将该集合作为 ε 参数集合；

步骤 4）根据 ε 参数集合，采用数学期望法生成 MinPts 参数列表，即

$$\mathrm{MinPts} = \frac{1}{n}\sum_{i=1}^{n} P_i 。 \tag{1}$$

其中：P_i 为第 i 个 ε 邻域内的样本数，n 表示样本总数。

1.2 确定 ε 与 MinPts 参数

本文采用 CH 指数对各参数下的聚类结果进行评分，最终根据评分确定最优参数进行聚类。虽然传统 CH 指数具有计算速度较快的特点，但存在聚类簇数为 1 时，评价指标无意义的情况，因此，本文对 CH 指数进行了优化，计算式为：

$$s(k) = \frac{T_{\mathrm{r}}(\boldsymbol{B}_k)}{T_{\mathrm{r}}(\boldsymbol{W}_k)} \times \frac{N+1-k}{k} 。 \tag{2}$$

其中：$T_{\mathrm{r}}(\boldsymbol{B}_k)$ 与 $T_{\mathrm{r}}(\boldsymbol{W}_k)$ 代表簇间离散矩阵 $\boldsymbol{B}_k$ 与簇内离散矩阵 $\boldsymbol{W}_k$ 的迹；N 为数据集内样本数量；k 代表簇的个数。当簇间离散度越大时，$T_{\mathrm{r}}(\boldsymbol{B}_k)$ 越大，当簇内离散度越小时，$T_{\mathrm{r}}(\boldsymbol{W}_k)$ 越小，因此，当聚类效果最优时，$s(k)$ 最大。算法运行流程如图 1 所示。

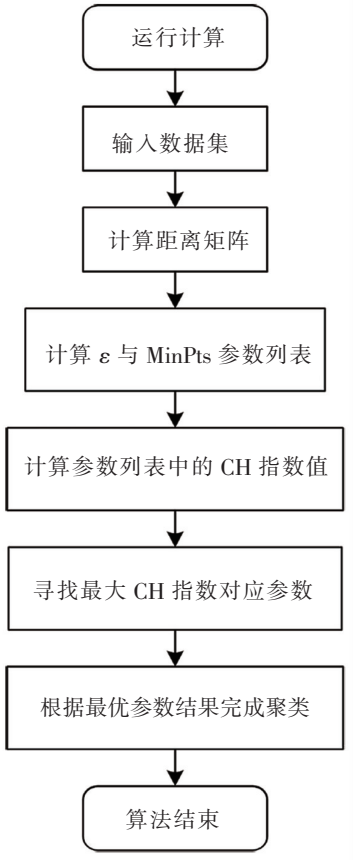


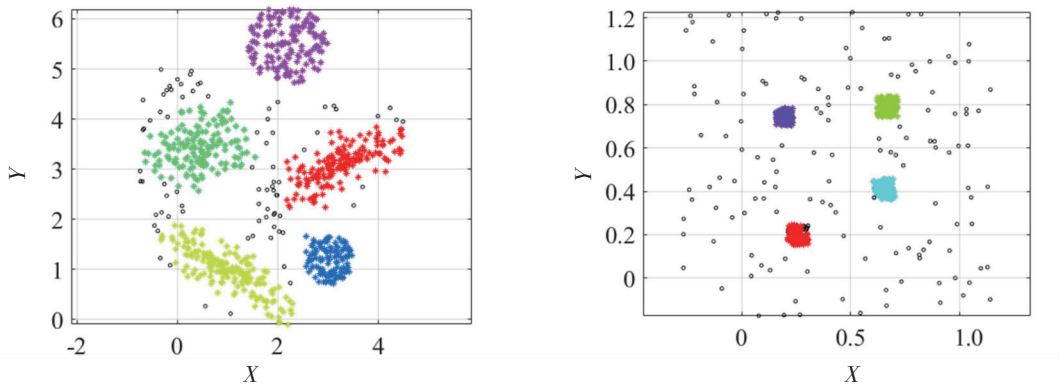
图 1 算法运行流程图

Fig.1 Algorithm flow chart

1.3 人工数据集验证

为检验参数自适应 DBSCAN 算法处理异常数据的能力，采用含噪声的 DS 和 DTN 数据集进行测试。DS 数据集是由 850 个含噪声数据样本构成的人工数据集，特点是数据样本密度分布不均匀；DTN 数据集由 623 个含噪声数据样本构成，特点是包含多个簇与噪声。

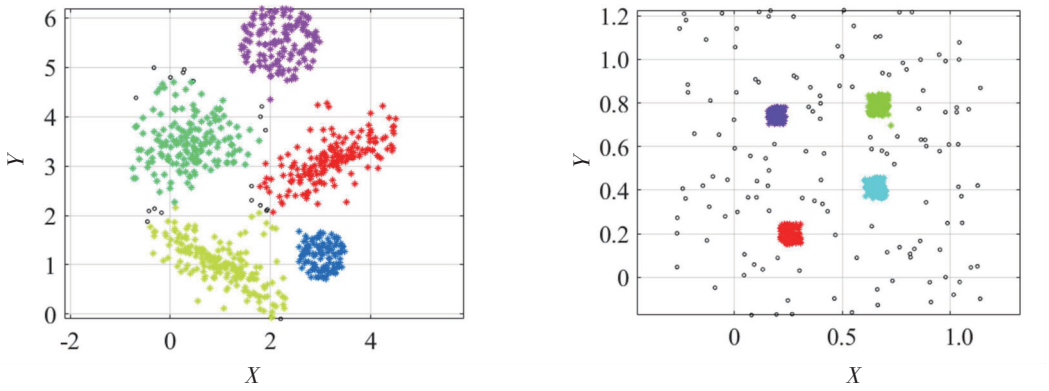
聚类分析中常见的准确率是一种评价聚类效果的指标。其取值在区间［0,1］内，越靠近 1，表明聚类效果越好。经典 DBSCAN 算法与本文所提出的参数自适应 DBSCAN 算法的聚类效果见图 2 和图 3。指标值及相关参数见表 1。



a) DS 数据集聚类效果图　　b) DTN 数据集聚类效果图

图 2 DBSCAN 算法聚类效果

Fig.2 Clustering effect of DBSCAN algorithm



a) DS 数据集聚类效果图　　b) DTN 数据集聚类效果图

图 3　参数自适应 DBSCAN 算法聚类效果

Fig.3　Parameter adaptive DBSCAN algorithm clustering effect

表 1　识别效果对比表

Tab. 1　Comparison of clustering effects

数据集	算法	ε	MinPts	准确率
DS	DBSCAN 算法	0.3000	10	0.868
	本算法	0.8300	96	0.943
DTN	DBSCAN 算法	0.0300	30	0.947
	本算法	0.0785	34	0.983

从图 2 可以看出：DBSCAN 虽然识别出了 DS 数据集的噪声点，但是在一个簇中，把非噪声点误认为噪声点；在 DTN 数据集中也存在同样的问题。而从图 3 可以看出，本文提出的参数自适应 DBSCAN 算法能较好地处理类似的情况，这一点从表 1 中的准确率也可以看出来。因此，本文提出的自适应 DBSCAN 算法在不需要输入参数的前提下，异常数据的检测效果优于普通 DBSCAN 算法。

2　实例分析

湄洲湾是我国东南沿海的一个天然良港。为了保障船舶航行安全，湄洲湾主航道配备了完善的助航浮标，并通过浮标遥测遥控系统监测浮标的状态。本文以湄洲湾港 1 号、2 号、4 号浮标在 2020 年 6 月份的 2072 个数据样本为实验对象，用本文提出的参数自适应 DBSCAN 算法进行浮标异常数据的识别。

2.1　实验环境及浮标数据分析

本文算法主体采用 Matlab 语言实现，系统采用微软 Windows 10，硬件配置 Intel Core i5-7200 处理器、内存 4 GB、硬盘 128 GB。

湄洲湾 1 号、2 号和 4 号浮标的遥测位置数据分布如图 4 所示。本文根据浮标位置数据的分布，将浮标异常数据分为两类：一类如图 4a 中零散的异常点，它与浮标数据之间有明显区别，此类异常主要是由于接受到其他信号形成的误差，故将这种误差定义为全局异常点；另一类如图 4b 中的 4 个数据异常点，其形成原因是数据传输过程中，受到数据精确值影响，部分浮标位置数据产生了误差，故将这种误差定义为局部异常点。图 4c ~ e 为单个浮标的分布数据。

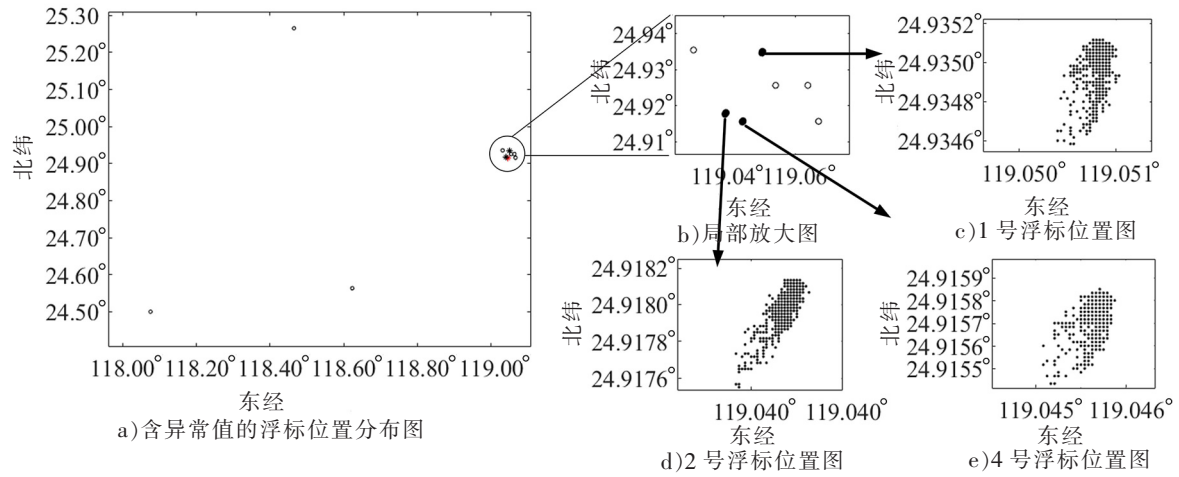


图 4　浮标数据初始点分布图

Fig.4　Buoy data initial point distribution chart

2.2　DBSCAN 算法参数生成

将浮标位置数据集输入算法程序，根据数据样本点的位置生成距离矩阵，并根据 1.1 中提出的方法生成 ε 与 MinPts 参数列表（见表 2）。

表 2　DBSCAN 算法参数与 CH 评分列表

Tab. 2　DBSCAN algorithm parameters and CH score list

ε	MinPts	ε	MinPts	ε	MinPts
⋮	⋮	⋮	⋮	⋮	⋮
0.000 020 925 2	54	0.000 234 093	598	0.000 347 466	663
0.000 021 907 1	54	0.000 235 571	600	0.000 353 435	665
0.000 022 716 4	54	0.000 236 597	601	0.000 359 243	667
0.000 023 603 9	54	0.000 238 248	603	0.000 364 983	668
0.000 024 277 1	54	0.000 240 221	603	0.000 368 683	671
0.000 024 809 1	54	0.000 242 052	604	0.000 373 735	672
0.000 025 451 6	54	0.000 243 684	606	0.000 380 008	672
0.000 025 975 2	54	0.000 245 216	606	0.000 389 678	673
0.000 026 863 4	54	0.000 246 833	608	0.000 393 429	675
0.000 027 397 5	54	0.000 247 995	609	0.000 393 782	678
⋮	⋮	⋮	⋮	⋮	⋮

由于生成参数较多，表 2 仅展示部分列表数据，具体参数变化特征如图 5 和图 6 所示。图 5 中，由于采用的 3 个浮标数据样本密度较大，且距离较远，因此，表示邻域距离阈值的 ε 参数呈现出三次分段增加的趋势；而图 6 中，由于 ε 参数的变化较小，无法引起 MinPts 参数发生明显变化，因此 MinPts 呈现出分段式上升的趋势。

用 CH 指数对参数聚类结果进行评分，CH 指数评分越高，表示聚类效果越优。根据列表参数，依次采用 CH 指数评分，结果如图 7 所示。当 ε 与 MinPts 参数只在一定范围内变化时，聚类结果基本不变，因此 CH 指数也呈现出梯度变化的结果。图 7a 为 CH 指数变化过程图，CH 指数最高值稳定在 7 128 145，此时聚类效果最优。图 7b 与图 7c 为子图 7a 中部分数据补充展示。

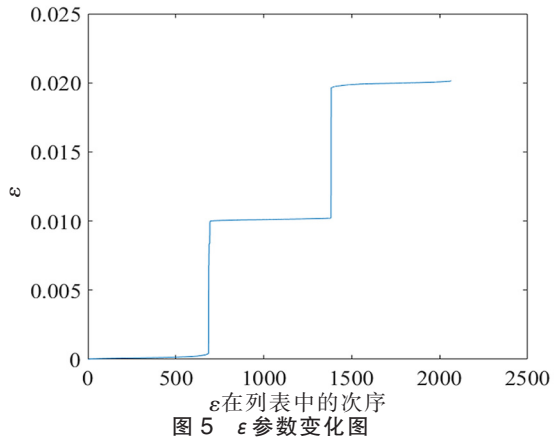


图 5　ε 参数变化图

Fig.5　Parameter ε change graph

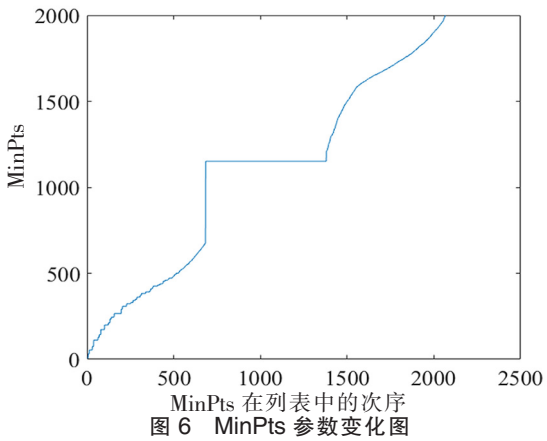


图 6　MinPts 参数变化图

Fig.6　Parameter MinPts change graph

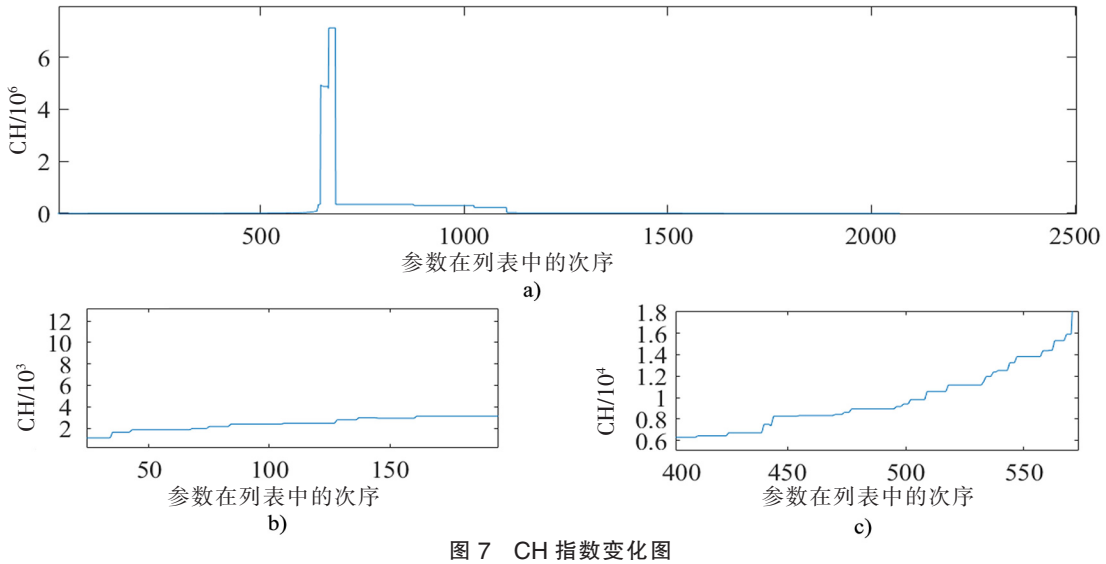


图 7　CH 指数变化图

Fig.7　CH fractional change graph

2.3　浮标位置异常识别结果分析

通过图 7 展示的可视化结果，确定最优参数处于参数列表的第 683 至第 687 位之间。为检测参数自适应 DBSCAN 算法的聚类效果，选取 CH 指数评分列表中第 17、627、683 位的参数点进行聚类结果分析对比，具体异常检测可视化结果见图 8 ~ 10。

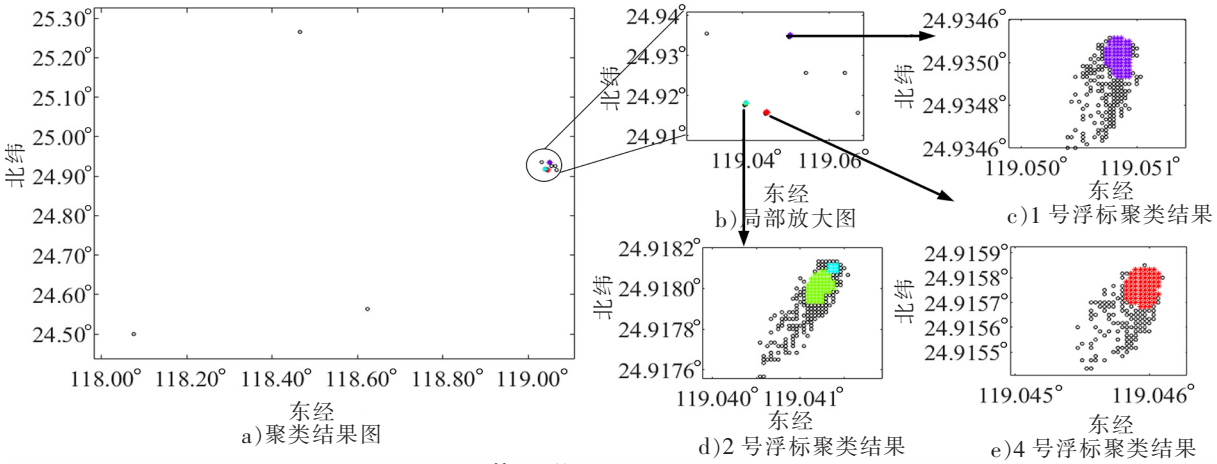


图 8　第 17 位 CH=1092.612

Fig.8　The 17th CH=1092.612

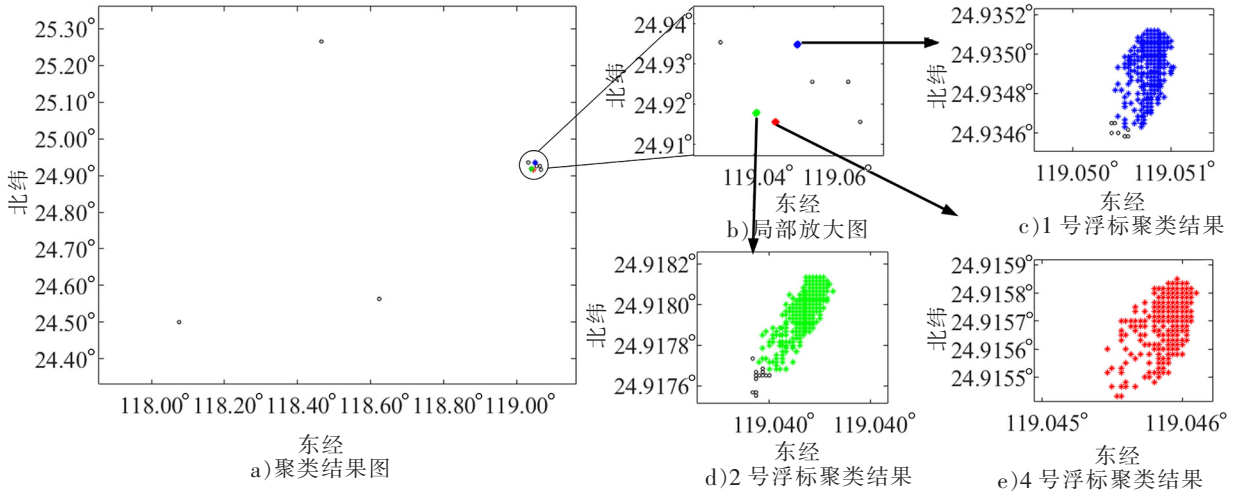


图 9　第 627 位 CH=57 343.57
Fig.9　The 627th CH=57 343.57

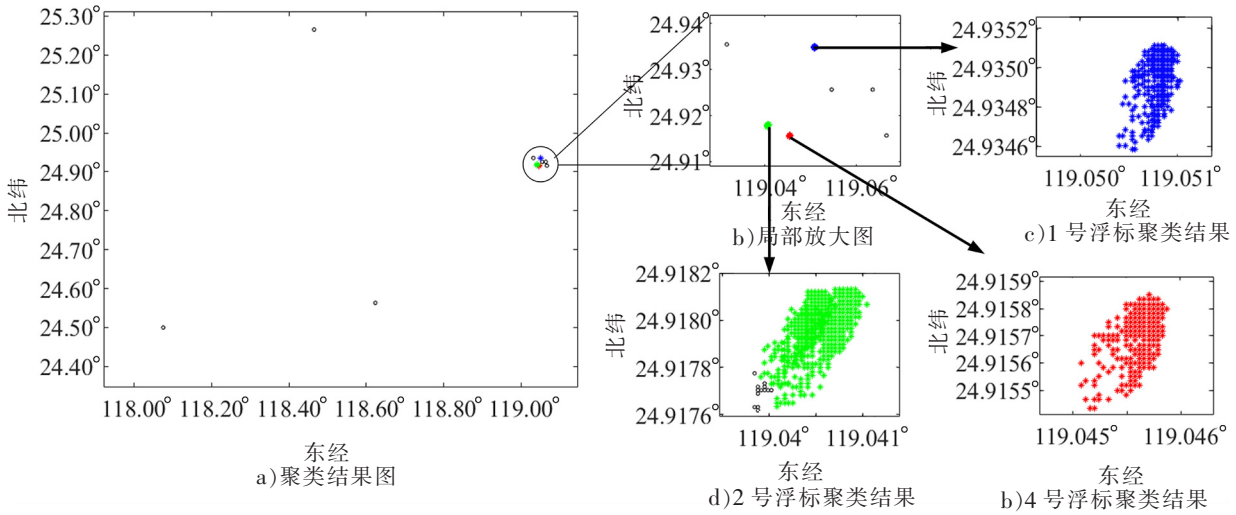


图 10　第 683 位 CH=7 128 145
Fig.10　The 683th CH=7 128 145

图 8 展示了 CH 指数评分为 1092.612 时的异常效果。从图 8 中可以看出，改进的 DBSCAN 算法基本能够识别出全局异常点，但出现部分正常的浮标数据被识别为异常的情况，且将图 8d 中单个浮标数据识别为 2 个簇，使最终聚类效果较差，异常位置点识别结果不理想。当 CH 指标评分为 57 343.57 时，能够识别出全局和局部的异常数据，但是有一些正常的浮标位置数据被划分为异常数据（见图 9）。图 10 展示了选取最优 CH 评分对应的参数聚类时的聚类结果和异常识别情况。从图 10 中可以看出，改进的 DBSCAN 算法在正确聚类的情况下，能有效识别全局和局部的异常数据，得到比较满意的效果，这说明，本文改进的 DBSCAN 算法能够有效识别浮标位置数据中的异常数据。

3　结论

本文提出一种参数自适应的 DBSCAN 聚类方法，通过 K 近邻算法与数学期望法，生成 ε 与 MinPts 参数列表，再通过 CH 指数进行评分，确定最优的 ε 与 MinPts 参数，全过程无需人工输入任何参数。将该算法应用于浮标位置数据的异常检测，实验结果表明：本文提出的自适应 DBSCAN 算法在不输入任何参数的情况下，能够根据浮标位置数据集自适应寻找参数，达到较好的聚类和异常点识别效果；但在自适应过程中，MinPts 参数取值的灵活性下降了。因此，提升 DBSCAN 算法的运行效

率，降低 ε 与 MinPts 参数之间的关联性是本文下一步的研究方向。

[参 考 文 献]

[1]XU L K,JIN Y X,XUE H,et al. Outlier detection of light buoy telemetry and telecontrol data based on improved adaptive neighborhood DBSCAN clustering[J]. Mathematical Problems in Engineering,2021,2021:5522107.

[2]ZHAO R,DU B,ZHANG L. A robust nonlinear hyperspectral anomaly detection approach[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2014,7(4):1227-1234.

[3]刘首华,陈满春,董明媚,等. 一种实用海洋浮标数据异常值质控方法[J]. 海洋通报,2016,35(3):264-270.

[4]YU Y,ZHU D,WANG J D et. al. Abnormal data detection for multivariate alarm systems based on correlation directions[J]. Journal of Loss Prevention in the Process Industries,2017,45:43-45.

[5]罗一迪,蒋华,王慧娇,等. 基于滑动窗口和 ARMA 的 Argo 剖面数据异常检测算法[J]. 计算机工程与应用,2018,54(19):254-260.

[6]张宇,周燕,陶邦一,等. 基于时序相关性分析方法的浮标异常数据识别[J]. 海洋学报,2020,42(11):131-141.

[7]卢勇夺,王朝阳,王豹,等. 我国海洋锚系浮标数据异常值检测方法研究:以 QF110 和 QF306 为例[J]. 海洋预报,2019,36(6):37-43.

[8]DAS A P,THAMPI S M,LLORET J. Anomaly detection in UASN localization based on time series analysis and fuzzy logic [J]. Mobile Netw Appl ,2020,25:55-67.

[9]林成虎,李晓东,金键,等. 基于 W-Kmeans 算法的 DNS 流量异常检测[J]. 计算机工程与设计,2013,34(6):2104-2108.

[10]AMIN K,MANEL G Z. A fuzzy anomaly detection system based on hybrid PSO-Kmeans algorithm in content-centric networks[J]. Neurocomputing,2015,149:1253-1269.

[11]蒋华,季丰,王慧娇,等. 改进 Kmeans 算法的海洋数据异常检测[J]. 计算机工程与设计,2018,39(10):3132-3136.

[12]高书强,李晨. 一种针对电力数据异常检测的改进谱聚类算法[J]. 计算机仿真,2019,36(11):239-242.

[13]ESTER M,KRIEGEL H P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the KDD. Oregon,Portland:AAAI Press,1996:226-231.

[14]HOSSEIN S E,SAYYED MM. A novel anomaly detection algorithm using DBSCAN and SVM in wireless sensor networks [J]. Wireless Personal Communications,2018,98(2):2025-2035.

[15]马良玉,孙佳明,於世磊,等. 基于 DBSCAN 和 SDAE 的风电机组异常工况预警研究[J]. 动力工程学报,2021,41(9):786-793.

[16]郑玉巧,刘玉涵,何正文,等. 基于 QM-DBSCAN 的风力机数据清洗方法[J]. 兰州理工大学学报,2021,47(6):50-55.

[17]JAIN P,BAJPAI M P,et al. A modified DBSCAN algorithm for anomaly detection in time-series data with aeasonality[J]. The International Arab Journal of Information Technology,2022,19(1):34028.

[18]李文杰,闫世强,蒋莹,等. 自适应确定 DBSCAN 算法参数的算法研究[J]. 计算机工程与应用,2019,55(5):1-7.

[19]ZHANG W B,YUE Z X,YE J M,et al. Modulation format identification using the Calinski-Harabasz index[J]. Applied optics,2022,61(3):851-857.

(责任编辑　朱雪莲　英文审校　周云龙)