一种基于降维密度聚类的船舶异常轨迹识别方法

李可欣[1]，郭健[1]，王宇君[2]，李宗明[3]，缪坤[4]，陈辉[5]
（1.信息工程大学，郑州 450001；2.32022 部队，广州 510000；3.31682 部队，兰州 730000；
4.陆军特种作战学院，广西 桂林 541000；5.31438 部队，沈阳 110031）

摘要：目的 有效分析和探索海洋船舶时空轨迹行为模式，提高船舶轨迹聚类的效率与质量，更好地检测真实船舶的异常行为。**方法** 针对当前船舶轨迹数据研究中存在的对多维特征信息利用不足、检测效率不高、检测精度较差等问题，提出一种精确度高、能自主识别分析多维特征的船舶异常轨迹识别方法。首先利用随机森林分类器评估多维特征重要性，构建轨迹特征的最优组合；然后提出一种降维密度聚类方法，将 T–分布随机邻域嵌入（T–SNE）和自适应密度聚类（DBSCAN）模型结合，通过构建特征选择层和无监督聚类层实现对数据元素非线性关系的高效提取以及对聚类参数的智能选择；最后根据聚类结果构建类簇特征向量，计算距离阈值判别轨迹相似度，实现轨迹异常检测模型的构建。**结果** 以 UCI 数据集为例，降维密度聚类方法对 4、13、30、64 维特征数据集的 F_1 分数能达到 0.9 048、0.9 534、0.8 218、0.6 627，多个聚类指标均优于 DBSCAN、K–Means 等常见聚类算法的。**结论** 研究结果表明，降维密度聚类方法能有效提取数据多维特征结构，实现聚类参数自适应，弥补密度聚类中参数难以确定的问题，有效实现对多种类型船舶轨迹异常的识别。

关键词：异常检测；时空轨迹；特征降维；密度聚类；参数自适应；T–分布随机邻域嵌入；随机森林
中图分类号：TP391.7 **文献标识码：**A **文章编号：**1001-3563(2023)11-0284-09
DOI：10.19554/j.cnki.1001-3563.2023.11.033

Trajectory Anomaly Identification Method of Vessels Based on Dimensional-density Reduction Clustering

LI Ke-xin[1], *GUO Jian*[1], *WANG Yu-jun*[2], *LI Zong-ming*[3], *MIAO Kun*[4], *CHEN Hui*[5]
(1. Information Engineering University, Zhengzhou 450001, China; 2. Unit 32022, Guangzhou 510000, China;
3. Unit 31682, Lanzhou 730000, China; 4. Army Special Operations College, Guangxi Guilin 541000, China;
5. Unit 31438, Shenyang 110031, China)

ABSTRACT: The work aims to effectively analyze and explore the space-time trajectory behavior patterns of ocean vessels, improve the efficiency and quality of vessel trajectory clustering, and better detect abnormal behaviors of real vessels. In allusion to existing problems in current vessel trajectory data research, such as insufficient utilization of multidimensional feature information, low detection efficiency, poor detection accuracy, etc., a high accuracy and multi-dimensional feature identification method for vessel abnormal trajectory was proposed. Firstly, random forest classifier was used to evaluate the importance of multidimensional features and construct the optimal combination of trajectory features. Then, a dimensional-density reduction clustering method was proposed to combine T-SNE and DBSCAN models. By constructing feature selection layer and unsupervised clustering layer, the nonlinear relation of data elements

收稿日期：2022−07−04
作者简介：李可欣（1998—），女，硕士生。
通信作者：郭健（1964—），女。

could be extracted efficiently and the clustering parameters could be selected intelligently. Finally, the cluster feature vector was constructed according to the clustering results, and the distance threshold was calculated to distinguish the trajectory similarity, and the trajectory anomaly detection model was constructed. With UCI datasets as examples, the F1 score of this method could reach 0.904 8, 0.953 4, 0.821 8 and 0.662 7 for datasets with 4, 13, 30 and 64 dimensional features, and many clustering indexes were superior to DBSCAN, K-means and other common clustering algorithms. The results show that this method can effectively extract multi-dimensional feature structure of data, realize clustering parameter self-adaptation, make up for the problem that parameters are difficult to be determined in density clustering, and effectively realize the identification of multiple types of ship trajectory anomalies.

KEY WORDS: anomaly detection; space-time trajectory; feature dimension reduction; density clustering; parameter self-adaptation; T-distributed stochastic neighbor embedding; random forests

随着经济全球化程度的不断加深，各类船舶逐渐实现高速化和大型化，持续增长的海洋运输需求与日趋饱和的航道容量之间的矛盾日益加剧，影响着海洋航运的安全与效率。为了更好地加强对海洋船舶的监控与管理，为海事监管人员提供更具针对性的解决方案，对大规模轨迹数据中的孤立、偏离、新颖数据点等进行检测。实现对海上船舶异常轨迹的识别与研究，从而实现对海域的智能高效全监控。在智慧海洋态势感知与管理方面具有重要的应用价值。

船舶自动识别系统（Automatic Identification System, AIS）包含船舶静态以及航行运动动态信息，已经成为了海上监控管理的主要数据来源。由于 AIS 信息最初是为避免碰撞而设计的，缺乏关于数据质量的元数据，如可靠性、确定性等，这使得利用 AIS 检测船舶异常成为一项非常困难的任务。AIS 数据包含地理空间特征、时序特征等一般数据所没有的特定特征，并且缺乏具有代表性的真实数据集，因此如网络流量[1]、网络安全[2]等领域的异常检测方法以及神经网络[3]、支持向量[4]等有监督模式的识别方法不适用该类数据。上述方法不仅要花费大量的时间对数据进行标记，类别不均衡也易导致检测结果的准确率降低。

针对 AIS 数据特性，近年来关于海上异常检测的研究方法可以分为基于规则的异常检测[5]以及基于学习的异常检测[6]。前者通过明确定义异常行为实现对异常的检测，具有可解释性，但需要基于大量历史数据对规则进行总结，但对一些隐式规则难以发现和描述，实际可用性较低。后者基于历史数据学习一般模式中隐藏的规则，成为海上异常检测的主导方法。基于学习的异常检测方法一般可分为 2 个阶段：学习船舶轨迹的一般模式；检测偏离模式的偏差。在第 1 阶段，以聚类分析为代表的无监督模式识别得到了广泛的应用，如 K–Means 算法[7]、DBSCAN 算法[8]、OPTICS 算法[9]、CURD 算法[10]，ST–DBSCAN 算法[11]、ST–OPTICS 算法[12]等。对于密度聚类通常只考虑空间信息这一问题，张春玮等[13]构建了船舶行为相似度模型，基于 DBSCAN 对船舶轨迹行为模式进行识别。王永明[14]综合 K–means 和 DBSCAN 算法对船舶轨迹进行聚类，以发现船舶航行轨迹异常。利用专家调查法和层次分析法对敏感水域的异常行为进行检测和排序。李楠等[15]通过聚类算法找到类簇中心点，利用轨迹信息和飞行距离构建异常因子，实现航空器异常检测。杜志强等[16]基于卡尔曼滤波，通过距离计算实现异常判别。孟祥泽等[17]采用 ST–DBSCAN 算法从老年人轨迹中提取行为模式链，结合空间环境信息构建异常分析模型。冯宏祥等[18]通过船舶轨迹更新距离的均值和标准差，实现对 AIS 误用等多种海上船舶异常的发现与数据处理。上述方法中聚类参数的选择往往基于经验，由于缺乏异常数据的标签，无法对所选参数的优劣进行评估，故难以获取最优参数。李文杰等[19]根据数据及自身分布特性生成候选集，基于参数寻优策略实现聚类过程的全自动化，但是在密度分布差异大的数据集上聚类效果差。万佳等[20]基于 KANN–DBSCAN 方法，结合去噪衰减和多密度聚类，在实现参数自适应的前提下，提升了方法在密度分布差异大数据集上的聚类效果，但是该方法仍需设置密度阈值，且计算复杂度较高。

针对上述问题，本文提出一种基于降维密度聚类的船舶异常轨迹识别方法，将 T–SNE 和自适应密度聚类结合，实现高效可靠的聚类，并根据聚类结果提取中心类簇构建类簇特征向量；最后根据不同距离阈值判别轨迹相似度，实现对异常轨迹的识别。构建海洋船舶轨迹异常模式识别模型，可以为智能海洋交通管理与优化提供科学化的数据支撑。

1　基于降维密度聚类的船舶异常轨迹识别方法

异常是指数据中不符合一般行为规范的模式。具体到海洋交通领域，异常轨迹一般包括：剧烈变速、剧烈转向、位置漂移等运动学异常以及船舶轨迹偏离一般航线、行驶在禁渔区或禁航区等规则异常。结合轨迹数据特点，设计基于自适应降维密度

聚类的船舶异常轨迹识别方法如图 1 所示。首先对 AIS 数据进行预处理，通过随机森林分类器构建最优多维特征组合；然后通过降维密度聚类生成轨迹聚类结果；根据聚类结果计算类簇特征向量，通过计算数据集中点与特征向量的位置距离和速度角度距离，生成判断相似度检测轨迹异常的距离阈值；最后结合轨迹段航行距离评估置信度，实现对轨迹异常的检测。

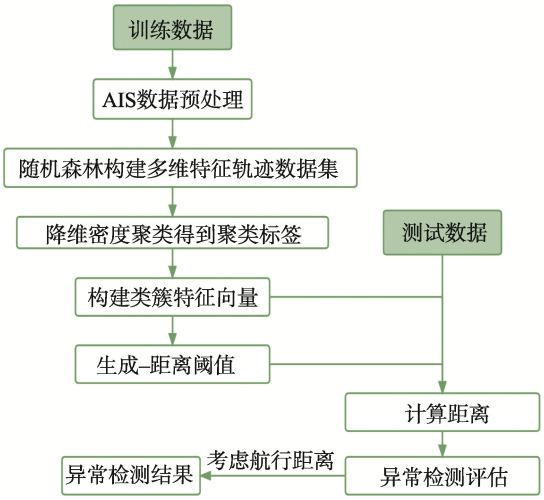


图 1 基于 DR–DBSCAN 的轨迹异常识别分析

Fig.1 Analysis of trajectory anomaly identification based on DR-DBSCAN

1.1 AIS 数据预处理

1.1.1 数据清洗

由于轨迹数据本身具有的多源异构性以及数据质量差等特点，需要对原始数据进行处理，轨迹数据处理通常需要解决以下 3 个问题：过滤清洗，去除由于采样频率、采样精度、人为失误等产生的噪声数据；降低计算量；提高轨迹数据的精度。

对轨迹数据进行缺失值删除、插值等预处理操作后，对轨迹基础信息进行分析计算构建多维特征，根据 MMSI 号将轨迹点分为完整轨迹段。船舶轨迹的集合 $M_{\_\mathrm{traj}}$、具体某一艘船舶的完整轨迹 $M_{\_\mathrm{traj}i}$ 以及船舶轨迹点信息 P 可表示为：

$$M_{\_\mathrm{traj}}=\{M_{\_\mathrm{traj_1}},M_{\_\mathrm{traj_2}},...,M_{\_\mathrm{traj_n}}\} \tag{1}$$

$$M_{\_\mathrm{traj_i}}=\{P_{i1},P_{i2},...,P_{ij}\} \tag{2}$$

$$P=\{x,y,d,t,v,C,H,A_{\mathrm{cal}},A_{\mathrm{rep}},v_{\mathrm{rep}}\} \tag{3}$$

式中：x、y 为轨迹点经纬度信息；d 为根据经纬度计算的地理空间距离；t 为此段轨迹航行的总时间；v 为 AIS 报告的船舶速度；C 为 AIS 报告的船舶航向；H 为 AIS 报告的船舶艏向；A_{cal} 为根据 H 计算的角度变化；A_{rep} 为根据 C 计算的角度变化量；v_{rep} 为根据时间距离计算的航迹平均速度。

1.1.2 多维特征构建

数据集所选取的特征属性离散性或相异性越高，数据的聚类效果则越好。原始轨迹数据包含经纬度、航行速度、航行方向等信息。为了更充分地挖掘轨迹特征，计算轨迹的航行距离、平均航行速度、加速度、转向角等特征，避免偏离数据干扰，每个特征指标分别取平均值、最大值、最小值、中值构建轨迹特征数据集。由于特征之间也存在干扰，利用随机森林分类器对轨迹数据进行分析，对多维特征轨迹进行评估，构建最佳特征组合，避免特征间的相互干扰，提高计算精度和计算效率。

1.1.3 轨迹分段和静止点提取

停止点一般是船舶的运动状态或行为模式发生变化的点，可以反映出停泊区、捕鱼区、低速作业区等停止区域，具有重要的分析意义。从清洗后的 AIS 数据中提取同时满足计算速度和报告速度均为静止状态的轨迹点，构建静止轨迹点，并依据静止点对完整轨迹段进行划分。

根据保留的特征属性信息，轨迹划分的流程分为 2 步：首先计算相邻轨迹点的距离、转向角以及速度；然后根据设定的速度阈值和最小轨迹长度，以静止点以及发生较大转向的点作为断点对轨迹段进行划分，筛除长度不符合要求的轨迹段，根据原始数据计算构造多维特征的时序子轨迹段特征。保留时序位置的子轨迹段集合 $S_{\_\mathrm{traj\_sequence}}$ 可表示为：

$$S_{\_\mathrm{traj\_sequence}}=\{S_{\_\mathrm{traj\_seq_1}},S_{\_\mathrm{traj\_seq_2}},...,S_{\_\mathrm{traj\_seq_m}}\} \tag{4}$$

$$S_{\_\mathrm{traj\_seq_i}}=list\left([x_{i1},y_{i1}],[x_{i2},y_{i2}],...,[x_{iq},y_{iq}]\right) \tag{5}$$

对子轨迹段的多维特征进行处理，将多点二维信息转化为单点二维信息，子轨迹段特征集合 $S_{\_\mathrm{traj\_features}}$ 可表示为：

$$S_{\_\mathrm{traj\_features}}=\{S_{\_\mathrm{traj\_f_1}},S_{\_\mathrm{traj\_f_2}},...,S_{\_\mathrm{traj\_f_m}}\} \tag{6}$$

1.2 降维密度聚类

1.2.1 算法原理

由于海上航行相较于陆上交通具有更高的自由度，不同海洋区域船舶航行规律具有较大差别，因此很难提前确定聚类数目。由于 AIS 数据本身具有不确定性，报告数据中包含许多错误轨迹构成的噪声点，因此本文基于 DBSCAN 算法，同时针对 DBSCAN 算法超参数难以确定的问题，提出一种充分利用数据分布特性的基于多维特征降维的聚类方法（Dimensionality Reduction-Density-Based Spatial Clustering of Applications with Noise, DR-DBSCAN）。引入 T–SNE 作为数据特征提取模块，借助多流形聚类[21]的思想，从高维数据中提取和构建更高质量和更具鲁棒性的数据特征低维有效表示。该方法的处理流程如图 2 所示。

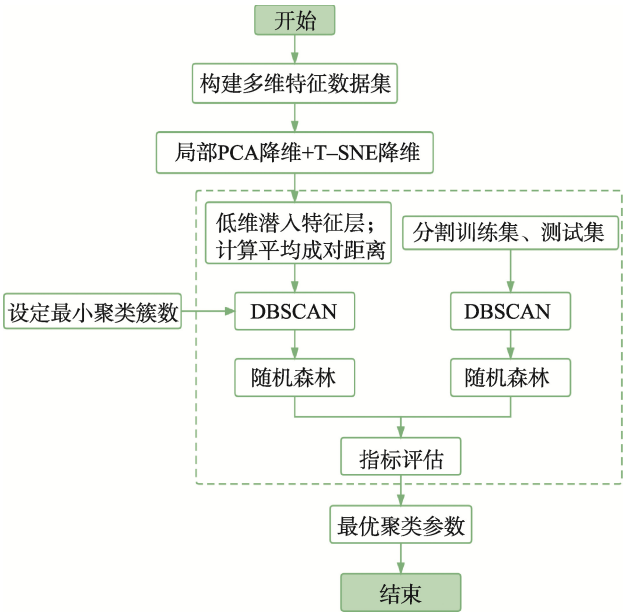


图 2　DR–DBSCAN 算法流程

Fig.2 DR-DBSCAN algorithm flow chart

对于多维特征数据集，常采用维数约减的方法降低特征间的复杂关系，减少噪声。常用的手段有特征删除、特征选择以及特征抽取。前 2 种手段往往容易导致信息丢失，PCA 和 T–SNE 都属于特征抽取的方法，在原始特征的基础上通过空间映射创建新的特征，能更好地挖掘特征间的深层联系。PCA 是一种线性降维方法，计算复杂度低但是特征表征效果较差；T–SNE 属于非线性方法，计算复杂度高但对特征映射效果较好。随机森林是一种由多个决策树组成的机器学习模型，具有很好的数据集适应能力，对高维数据、离散或连续型数据都能很好的处理，鲁棒性强。因此，本文将 2 种方法结合，在提高计算效率的同时充分挖掘特征间的相关关系，使得在聚类时能充分利用数据特征间的关系；然后利用随机森林模型学习聚类标签，并判断样本类型。

在 DR–DBSCAN 算法中，具体步骤如下：

1）将输入的多维特征数据通过局部 PCA 方法进行投影，再利用快速 T–SNE 模型将 PCA 处理后的数据转化为低维嵌入。

2）计算低维嵌入层数据的平均成对距离作为 eps 候选集设置的基数构建候选集，并将数据划分为训练集和测试集。

3）分别将低维嵌入层数据和训练集数据代入 DBSCAN 模型中进行聚类，提取集群的聚类标签，去除聚类簇数不符合设置最小聚类簇数的数据。

4）分别用嵌入层及其训练集的聚类标签训练随机森林分类器。

5）将测试集代入步骤 4 中训练的 2 个分类器，经过 K 折交叉验证得到聚类参数最优值，输出聚类结果。

1.2.2　算法分析及评价

为了更好地验证所提出算法的性能，综合考虑内部和外部聚类评估标准构建算法评价体系。外部评价指标是指基于已知标签或模型，将聚类结果与其进行比较。选取的数据集均为有标签数据，为了对聚类结果进行准确评价，引入外部聚类指标 F_1 分数、调整兰德系数（Adjusted Rand index, A_{RI}）、归一化互信息（Normalized Mutual Information, N_{MI}）作为评价指标，计算公式如下。

F_1 分数是精确率和召回率的调和平均数，F_1 越高则模型越稳健，公式见式（7）。

$$F_1 = 2 \cdot \frac{P \times R}{P + R} \tag{7}$$

式中：P 为精确率；R 为召回率。

A_{RI} 的取值范围为[−1,1]，相比兰德系数具有更高的区分度，值越大则表示聚类结果越吻合，计算式见式（8）。

$$A_{\mathrm{RI}} = \frac{R_{\mathrm{I}} - E[R_{\mathrm{I}}]}{\max(R_{\mathrm{I}}) - E[R_{\mathrm{I}}]} \tag{8}$$

式中：R_{I} 为兰德系数，取值范围为[0,1]，表示聚类标签和真实标签的比值情况。

N_{MI} 值用来衡量 2 个数据间的相关性，在聚类中用于度量 2 个聚类结果的相近程度，N_{MI} 值越大则表示划分越准确，公式见式（9）。

$$N_{\mathrm{MI}}(X,Y) = \frac{2MI(X,Y)}{H(X) + H(Y)} \tag{9}$$

式中：$H(X)$、$H(Y)$ 分别为聚类标签和真实标签的信息熵，即出现的概率；$MI(X,Y)$ 为互信息，是联合分布与乘积分布的相对熵。

内部评价指标是根据数据集的固有特征来对算法结果进行评估。引入聚类性能内部评价指标包含轮廓系数（Silhouette Coefficient，S_{C}）和 Davies–Boulding 指数（D_{BI}）。轮廓系数结合了凝聚度和分离度，取值为[−1, 1]，其值越大越好，轮廓系数的计算式见式（10）。

$$S_{C(i)} = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{10}$$

式中：为 $a(i)$簇内不相似度；$b(i)$为簇间不相似度。

D_{BI} 指数又称分类适确性指标，D_{BI} 越小说明聚类效果越好，计算式见式（11）。

$$D_{\mathrm{BI}} = \frac{1}{K}\left[\sum_{i=1}^{K} \frac{m(C_i) + m(C_j)}{d(\mu_i, \mu_j)}\right] \tag{11}$$

式中：$m(C_{\mathrm{I}})$和 $m(C_j)$为样本间平均距离；$d\,(\mu_i, \mu_j)$为簇中心点距离。

1.3　船舶轨迹异常识别

1.3.1　类簇特征向量提取

在利用 DR–DBSCAN 算法对轨迹进行聚类后，

类簇可以代表船舶的一般运动模式。通过构建类簇特征向量来提取类簇特征，避免使用每个类簇的所有轨迹点进行计算所产生的巨大运算量，导致轨迹数据异常检测的效率降低。类簇特征向量表达式可表示为式（12）。

$$\boldsymbol{l}_{\mathrm{GV}}=(m_C,m_S,m_x,m_y,m_d,m_A) \tag{12}$$

提取类簇特征向量表示船舶行为的一般模式，通过计算训练数据集中轨迹点与类簇特征向量的聚类距离，生成距离阈值，根据特征向量和距离阈值对测试集轨迹点进行异常检测。最后根据船舶轨迹中异常点的占比来判断轨迹段是否异常。类簇特征向量提取示意图如图 3 所示。首先计算类簇平均航向角；然后根据平均航向角以及类簇点的经纬度范围构建基础网格；根据不同基础网格的经纬度跨度，将基础网格划分为小网格；计算每个网格中的类簇点的平均速度、平均经纬度、平均距离以及最大转向角；保存各个网格的特征向量，构建类簇特征向量集合。

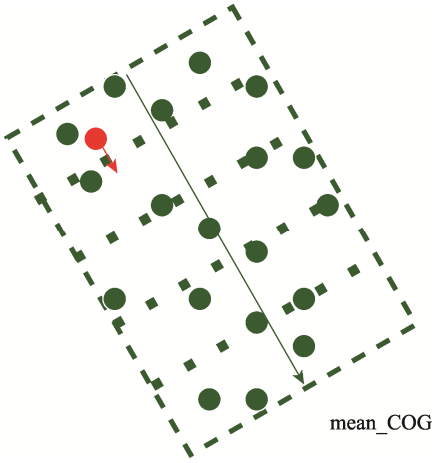


图 3 类簇特征向量提取示意图

Fig.3 Feature vector extraction of class cluster

1.3.2 距离判定阈值计算

对于一个待检测的轨迹数据 P，首先根据 P 点的经纬度坐标，利用半正矢公式计算 P 与类簇特征向量的地理距离 D_{p}。

$$D_{\mathrm{p}}=2R\times\arcsin\left[\sin^2\left(\frac{x_{\mathrm{p}}-x_{l_{\mathrm{GV}}}}{2}\right)\right]+\cos(x_{\mathrm{p}})\sin^2\left(\frac{y_{\mathrm{p}}-y_{l_{\mathrm{GV}}}}{2}\right) \tag{13}$$

式中：R 为地球半径，此处取地球平均半径 R=6 371.393 km。

保留使得所求地理距离最小的类簇特征向量 $\boldsymbol{l}_{\mathrm{GV}i}$，根据该特征向量对应的其他特征分量计算该轨迹的相对距离 $d_{\_1}$、速度转角距离 $d_{\_\mathrm{sa}}$，见式（14）—（15）。

$$d_{\_1}=\frac{D_{\mathrm{p}}}{D_{l_{\mathrm{GV}i}}} \tag{14}$$

$$d_{\_\mathrm{sa}}=\cos(C_{\mathrm{p}}-C_{l_{\mathrm{GV}i}})\frac{A_{\mathrm{p}}}{A_{l_{\mathrm{GV}i}}} \tag{15}$$

去除噪声数据和聚类异常数据后，通过计算训练数据集中轨迹数据与类簇特征向量的距离，生成各个距离阈值，实现对异常轨迹的识别与评估。

2 船舶异常检测实验分析

2.1 实验数据

本实验选取 2019 年 1 月 1 日的 AIS 数据作为训练集，设置美国西海岸、美国东海岸和墨西哥湾 3 个实验区域进行分析。

由图 4 可以看出，美西和美东均分布有较多较为重要的港口，这 2 个区域的客船和货船占比相对较多，分别为 35.45%和 28.24%。墨西哥湾北部为佛罗里达半岛，人口密度较大，该区域的船舶分布较为密集，且游艇占比较大。特殊船舶包含各种水上或水下作业船舶，如引航、搜救、挖掘、潜水等，墨西哥湾的浅大陆棚区蕴藏大量的石油和天然气，该区域特殊船舶占比较高。船舶分布与地区地理环境具有很高的相关性，根据某地区的船舶类型分布可以推论该地区的地理环境特征。

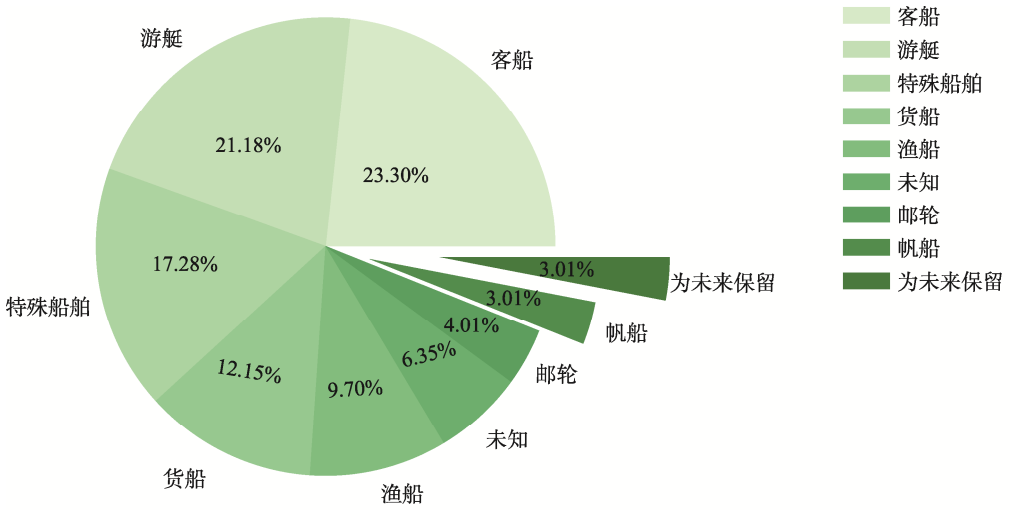
2.2 数据预处理

原始 AIS 轨迹数据共 7 516 408 条，包含船舶 13 115 艘。经过数据清洗和预处理后的 AIS 轨迹数据共 7 515 892 条，提取静止点 615 977 个。根据设定的速度阈值筛选静止点以及航向发生重大变化的点作为断点对轨迹段进行划分，保留所有轨迹长度在 10 以上的轨迹段，得到轨迹段为 5 812 条，包含船舶 4 740 艘。

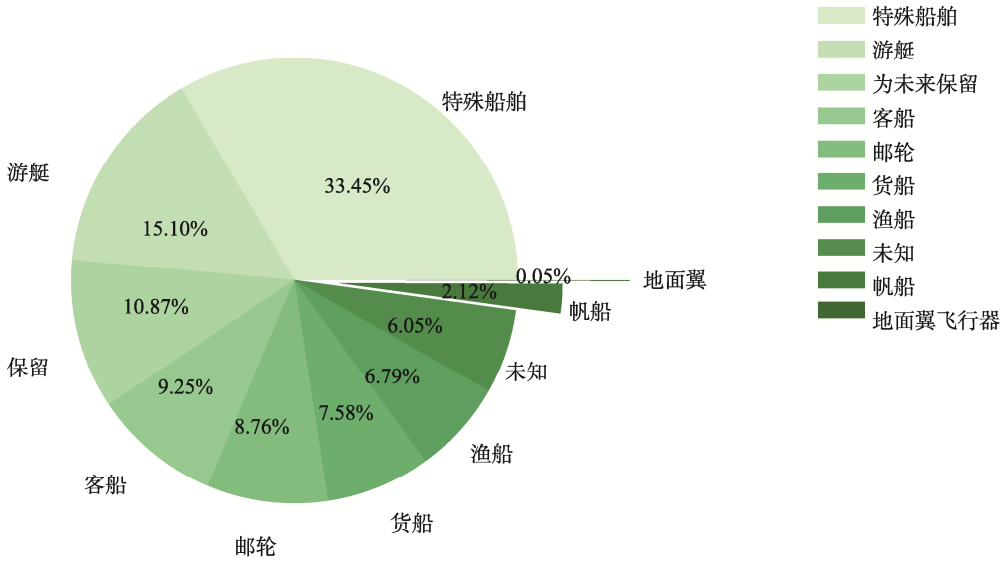
为了确保结果的准确性与有效性，使用过滤法结合随机森林模型对特征进行组合选择，以得到最佳特征组合。进行了多组对比实验，每组实验迭代运行 5 次以消除随机性，实验结果如表 1 所示。根据实验结果最终保留经纬度、报告转角以及报告速度的最大值、最小值、中位数和均值信息作为最终特征组合。

2.3 聚类分析

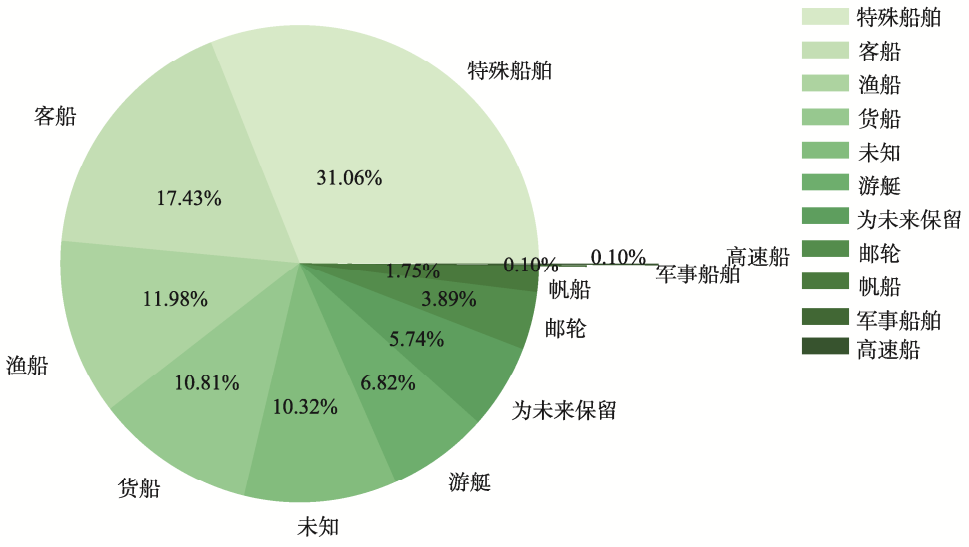
由于轨迹数据无标签，为了验证聚类方法的精确性和普适性，选取 4 个经典的具有不同维度特征的 UCI 数据集进行聚类分析，评估 DR–DBSCAN 算法解决实际问题的能力。综合考虑内部和外部聚类评估标准构建算法评价体系，通过属性数以及类别数的变化，观察相对变化下算法的聚类性能。数据集在不同算法下的聚类指标对比信息见表 2。



a 美国西海岸



b 墨西哥湾



c 美国东海岸

图 4　实验区域船舶类型分布

Fig.4 Vessel type distribution in the experimental area

表 1 轨迹特征组合评估

Tab.1 Trajectory feature combination evaluation

轨迹属性	渔船	帆船	游艇	特殊船舶	客船	货船	邮轮
all	0.64	0.50	0.74	0.94	0.64	0.70	0.38
mean_all	0.53	0.31	0.76	0.93	0.46	0.56	0.43
no_distance	0.66	0.50	0.73	0.93	0.69	0.67	0.40
no_heading	0.71	0.42	0.75	0.95	0.66	0.63	0.38
no_cog	0.67	0.38	0.71	0.92	0.64	0.66	0.50
no_angle_rep	0.64	0.54	0.74	0.93	0.63	0.69	0.40
no_angle_calc	0.69	0.42	0.74	0.93	0.61	0.59	0.45
no_speed_rep	0.59	0.31	0.76	0.93	0.54	0.68	0.43
no_speed_calc	0.71	0.50	0.75	0.93	0.66	0.66	0.45
no_lat	0.57	0.46	0.74	0.93	0.62	0.71	0.33
no_lon	0.47	0.42	0.65	0.95	0.51	0.62	0.33
最优组合	0.72	0.50	0.73	0.94	0.63	0.72	0.38

表 2 UCI 数据集聚类指标对比

Tab.2 Comparison of UCI datasets agglomeration indicators

数据集	属性数	类别数	数据量	方法	评价指标				
					F_1	A_{RI}	N_{MI}	S_C	D_{BI}
Iris	4	3	150	KMeans	0.536 8	0.539 9	0.6565	0.681 0	0.404 3
				DBSCAN	0.527 6	0.414 3	0.3442	0.325 3	2.817 2
				KANN–DBSCAN	0.532 0	0.471 9	0.5727	0.199 2	2.813 5
				MDA–DBSCAN	0.545 9	0.539 6	0.6674	0.132 8	0.498 4
				DR–DBSCAN	**0.904 8**	**0.759 2**	**0.8057**	**0.554 2**	**0.658 4**
Digits	64	10	1797	KMeans	0.320 6	0.473 2	0.6318	0.162 1	1.991 5
				DBSCAN	0.414 5	0.550 7	0.7307	0.093 5	2.157 9
				KANN–DBSCAN	0.098 5	0.109 7	0.3244	0.0398 8	4.017 5
				MDA–DBSCAN	0.120 5	0.178 6	0.2840	0.007 6	4.706 9
				DR–DBSCAN	**0.662 7**	**0.892 5**	**0.9211**	**0.073 4**	**1.822 4**
Cancer	30	2	569	KMeans	0.549 8	0.255 0	0.3543	0.467 4	0.624 7
				DBSCAN	0.670 0	0.287 5	0.3086	0.553 9	1.180 2
				KANN–DBSCAN	0.501 8	0.068 4	0.0979	0.359 6	0.245 3
				MDA–DBSCAN	0.567 3	0.065 9	0.1049	0.585 5	0.357 0
				DR–DBSCAN	**0.821 8**	**0.405 9**	**0.4085**	**0.400 7**	**0.744 9**
wine	13	2	178	KMeans	0.165 6	0.371 1	0.4288	0.571 1	0.534 2
				DBSCAN	0.535 0	0.382 1	0.3852	0.402 6	0.597 9
				KANN–DBSCAN	0.279 2	0.297 1	0.3820	0.323 5	2.436 5
				MDA–DBSCAN	0.422 4	0.305 1	0.4078	0.134 7	1.040 0
				DR–DBSCAN	**0.953 4**	**0.861 2**	**0.8432**	**0.280 6**	**1.315 7**

从 3 个外部聚类指标 F_1、A_{RI} 和 N_{MI} 来看，DR–DBSCAN 在 4 个数据集上均有较好得分，明显优于其他几种算法，但内部聚类指标 D_{BI} 评估结果相对较差。说明本文算法 DR–DBSCAN 能深入挖掘数据内部特征，而不是单纯从点迹的空间分布上挖掘信息，因此能在数据分布较为离散的情况下，实现较高的分类准确度。综合实验结果分析，本文算法 DR–DBSCAN 通过数据集低维嵌入特征层的构建，深入挖掘数据集特征分布特性，能够得到更符合数据特性的密度阈值。本文算法相较于一般的密度聚类方法，在实现参数自适应的同时能较好地处理多维数据集，在几个密度分布不均匀的多维数据集上均有较好的表现。

异常检测的实质就是学习一般行为模式，发现与一般模式相异的数据。DR–DBSCAN 算法能根据数据特征，拟合数据分布特性，构建数据分布一般模式的类簇，从而可以实现异常数据的识别。

在 3 个试验区中，美国西海岸区域包含轨迹数据 897 条；墨西哥湾区域包含轨迹 2 033 条；美国东海岸区域包含轨迹数据 1 027 条。根据随机森林分类器所构建的轨迹数据特征组合，对 3 个实验区域的轨迹段进行聚类，去除掉无法聚类的噪声点或异常轨迹，聚类结果与船舶类型分布较为类似。根据每个区域的聚类结果，划分网格并提取类簇特征向量，计算距离阈值。3 个区域的位置距离阈值分别为美国西海岸 2.249 27、墨西哥湾 1.805 97、东海岸 1.740 78；速度方向距离阈值分别为美国西海岸 1.777 7、墨西哥湾 1.952 8、东海岸 1.705 02。

2.4　异常检测

根据聚类结果获取相应区域的距离阈值，将轨迹点超出阈值范围的视为异常点。对于一条轨迹，当异常点占比超过 70%，则该轨迹视为异常轨迹。选取 2019 年 1 月 1 日的 AIS 数据进行异常检测，经过数据划分后，根据距离阈值判断异常轨迹点。经过距离计算和异常判定，美国西海岸轨迹数据 898 条，检测出异常轨迹 57 条，其中速度或方向异常速度的轨迹 19 条，位置异常轨迹 38 条；墨西哥湾轨迹数据 2 160 条，检测出异常轨迹 60 条，其中速度或方向异常速度的轨迹 17 条，位置异常轨迹 43 条；美国东海岸轨迹数据 1 054 条，检测出异常轨迹 45 条，其中速度或方向异常速度的轨迹 23 条，位置异常轨迹 312 条。

美国西海岸和美国东海岸区域沿岸为主要航道，美西向东为内河流域，向西为大西洋；美东向西为内河流域，向东为太平洋，二者内河与沿岸区域航道明显，大洋区域轨迹较为离散。墨西哥湾区域内河流域和离港航线分布较为清晰，但在中部区域轨迹分布较为杂乱。异常检测通过计算距离阈值，将超出阈值范围的判定为异常。位置异常可以解释为分布明显偏离航道的轨迹，或同一艘船舶短时间内位置出现显著飘移等情况。将行驶方向与一般轨迹不一致或突然发生较大转向或速度变动的轨迹标记为速度方向异常；将轨迹方向与周围轨迹相异的视为方向异常；将轨迹发生较大转向的可视为转向异常。2 种方向异常都可以由轨迹分布和轨迹形状进行判断。

3　结语

本文提出了一种基于降维密度聚类的船舶异常轨迹识别方法。利用随机森林分类器对轨迹多维特征的重要性进行评估，构建轨迹特征的最优特征组合。基于 DR–DBSCAN 聚类算法对历史 AIS 数据进行聚类分析，学习船舶的一般行为模式构建船舶类簇特征向量并计算距离阈值。在保证聚类精度的前提下，有效提高了聚类效率，减少了调参过程中对人工的依赖。采用 4 种经典 UCI 数据集验证 DR–DBSCAN 算法的精确度和有效性，并使用 2019 年 1 月 1 日的真实航迹数据在 3 个不同的实验区域进行分析，以减少水文地理环境对方法有效性与准确性的干扰。结果表明，该方法能够有效检测出船舶的位置异常、速度方向异常，对加强船舶交通行为分析和船舶交通监管具有重要意义。由于硬件设施限制，本文选取数据集时空范围较小，对多种类型的异常划分和定义不够详细。未来可以进一步修改模型架构在更大区域进行验证，更为明确地划分各种类型的异常，并将模型拓展至陆空交通运输领域，更好地分析判断不同的轨迹异常行为，为海陆空运输及交通管理提供数据支撑。

参考文献：

[1] SHERRY N L. Anomaly detection in aircraft data using Recurrent Neural Networks (RNN)[J]. Integrated Communications Navigation and Surveillance (ICNS), 2016(4): 19-21.

[2] RADFORD B J, APOLONIO L M, TRIAS A J, et al. Network Traffic Anomaly Detection Using Recurrent Neural Networks[EB/OL]. 2018: arXiv: 1803.10769. https://arxiv.org/abs/1803.10769

[3] 李恒，张氢，秦仙蓉，等. 基于短时傅里叶变换和卷积神经网络的轴承故障诊断方法[J]. 振动与冲击，2018, 37(19): 124-131.
LI Heng, ZHANG Qing, QIN Xian-rong, et al. Fault Diagnosis Method for Rolling Bearings Based on Short-Time Fourier Transform and Convolution Neural Network[J]. Journal of Vibration and Shock, 2018, 37(19): 124-131.

[4] 朱建军，安攀峰，万明. 工控网络异常行为的 RST-SVM 入侵检测方法[J]. 电子测量与仪器学报，2018, 32(7): 8-14.
ZHU Jian-jun, AN Pan-feng, WAN Ming. Intrusion De-

tection Method of RST-SVM for Abnormal Behavior in Industrial Control Network[J]. Journal of Electronic Measurement and Instrumentation, 2018, 32(7): 8-14.

[5] KAZEMI S, ABGHARI S, LAVESSON N, et al. Open Data For Anomaly Detection in Maritime Surveillance[J]. Expert Systems with Applications, 2013, 40, (14): 5719-5729.

[6] ZHAO Liang-bin, SHI Guo-you. Maritime Anomaly Detection Using Density-based Clustering and Recurrent Neural Network[J]. The Journal of Navigation, 2019, 72(4): 894-916.

[7] MACQUEEN J. Some Methods for Classification and Analysis of Multivariate Observations[C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1(233): 281-297.

[8] ESTER M, KRIEGEL H P, SANDER J, et al. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[R]. Proceedings of Second International Conference on Knowledge Discovery and Data Mining, 1996: 226-231.

[9] Ankerst M., Breunig M. M., Kriegel H, et al. OPTICS: Ordering Points to Identify the Clustering Structure[R]. SIGMOD ’99 Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data, 28(2), 49-60.

[10] MA S, WANG T, TANG S, et al. A New Fast Clustering Algorithm Based on Reference and Density[C]// Web-Age Information Management. Springer, Berlin, Heidelberg, 2003.

[11] BIRANT D, KUT A. ST-DBSCAN: An Algorithm for Clustering Spatial-temporal Data[J]. Data and Knowledge Engineering, 2007, 60(1): 208-221.

[12] AGRAWAL K P, GARG S, SHARMA S, et al. Development and Validation of OPTICS based Spatio-temporal Clustering Technique[J]. Information Sciences, 2016, 369: 388-401.

[13] 张春玮, 马杰, 牛元淼, 等. 基于行为特征相似度的船舶轨迹聚类方法[J]. 武汉理工大学学报(交通科学与工程版), 2019, 43(3): 517-521.
ZHANG Chun-wei, MA Jie, NIU Yuan-miao, et al. Clustering Method of Ship Trajectory Based on Similarity of Behavior Pattern[J]. Journal of Wuhan University of Technology (Transportation Science & Engineering), 2019, 43(3): 517-521.

[14] 王永明. 基于大规模 AIS 数据的船舶异常行为检测与预警[D]. 大连: 大连海事大学, 2020.
WANG Yong-ming. Vessel Abnormal Behavior Detection and warning based on large-scale AIS Data[D]. Dalian: Dalian Maritime University, 2020.

[15] 李楠, 强懿耕, 樊瑞, 等. 基于异常因子的航空器飞行轨迹异常检测研究[J]. 安全与环境学报, 2021, 21(2): 643-648.
LI Nan, QIANG Yi-geng, FAN Rui, et al. On the Abnormal Detection of the Aircraft Flight Trajectory Based on the Abnormal Factor Statistics[J]. Journal of Safety and Environment, 2021, 21(2): 643-648.

[16] 杜志强, 谭玉琪, 仇林遥. 基于卡尔曼滤波的船舶轨迹异常行为快速检测方法[J]. 地理信息世界, 2021, 28(4): 112-118.
DU Zhi-qiang, TAN Yu-qi, QIU Lin-yao. On the Rapid Detection of Abnormal Ship Trajectories by Kalman Filter[J]. Geomatics World, 2021, 28(4): 112-118.

[17] 孟祥泽, 胡啸峰, 沈兵. 社区老年人空间行为轨迹异常分析方法[J]. 科学技术与工程, 2021, 21(9): 3676-3681.
MENG Xiang-ze, HU Xiao-feng, SHEN Bing. Abnormal Analysis Method of Old People' s Spatial Behavior Trajectory in Community[J]. Science Technology and Engineering, 2021, 21(9): 3676-3681.

[18] 冯宏祥, ANNA MujalColilles, 杨忠振. 基于距离分布的 AIS 异常数据处理方法[J]. 中国航海, 2021, 44(4): 26-31.
FENG Hong-xiang, ANNA M, YANG Zhong-zhen. Outlier Processing of AIS Data According to Distance Distribution[J]. Navigation of China, 2021, 44(4): 26-31.

[19] 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.
LI Wen-jie, YAN Shi-qiang, JIANG Ying, et al. Research on Method of Self-Adaptive Determination of DBSCAN Algorithm Parameters[J]. Computer Engineering and Applications, 2019, 55(5): 1-7.

[20] 万佳, 胡大裟, 蒋玉明. 多密度自适应确定 DBSCAN 算法参数的算法研究[J]. 计算机工程与应用, 2022, 58(2): 78-85.
WAN Jia, HU Da-sha, JIANG Yu-ming. Research on Method of Multi-Density Self-Adaptive Determination of DBSCAN Algorithm Parameters[J]. Computer Engineering and Applications, 2022, 58(2): 78-85.

[21] 边荣正, 张鉴, 周亮, 等. 面向复杂多流形高维数据的t-SNE降维方法[J]. 计算机辅助设计与图形学学报, 2021, 33(11): 1746-1754.
BIAN Rong-zheng, ZHANG Jian, ZHOU Liang, et al. T-SNE for Complex Multi-Manifold High-Dimensional Data[J]. Journal of Computer-Aided Design & Computer Graphics, 2021, 33(11): 1746-1754.

责任编辑：曾钰婵