计算机科学
COMPUTER SCIENCE

基于深度聚类的航空交通流识别与异常检测研究

饶丹, 时宏伟

引用本文

饶丹, 时宏伟. 基于深度聚类的航空交通流识别与异常检测研究[J]. 计算机科学, 2023, 50(3): 121-128.

RAO Dan, SHI Hongwei. Study on Air Traffic Flow Recognition and Anomaly Detection Based on Deep Clustering [J]. Computer Science, 2023, 50(3): 121-128.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220100086

基于深度聚类的航空交通流识别与异常检测研究

饶 丹 时宏伟
四川大学计算机学院 成都 610065
(3328456642@qq.com)

摘 要 针对传统的聚类算法无法捕获高维轨迹数据在低维空间中的隐含关系,且难以定义适当的相似性度量以同时考虑轨迹的局部和全局特征的问题,提出了一种基于深度神经网络的多变量轨迹深度聚类框架(MTDC)并将其用于航空交通流识别与异常检测。该框架主要包含一个非对称的自编码器和一个自定义的轨迹聚类层。自编码器由一维卷积神经网络和双向长短时记忆网络堆叠而成,用于学习原始输入在低维隐空间中的特征表示。轨迹聚类层则通过计算隐空间中样本的Q分布实现聚类。结合自编码器的重建损失和轨迹聚类Q分布定义了一个新的异常分数,用于检测异常轨迹。使用基于广播式自动相关监视(ADS-B)的真实轨迹数据进行实验,结果表明,所提框架能有效地进行航空交通流识别,并能检测出具有实际意义且可解释的异常轨迹。

关键词:轨迹聚类;异常检测;深度神经网络;自编码器;ADS-B
中图法分类号 TP391

Study on Air Traffic Flow Recognition and Anomaly Detection Based on Deep Clustering

RAO Dan and SHI Hongwei
School of Computer Science,Sichuan University,Chengdu 610065,China

Abstract Aiming at the problem that traditional clustering algorithms cannot capture the implicit relationship of high-dimensional trajectory data in low-dimensional space,and it is difficult to define appropriate similarity measures to consider both local and global features of trajectories,a multivariate trajectory deep clustering(MTDC) framework based on deep neural network (DNN) is proposed and used for air traffic flow recognition and anomaly detection. The framework mainly includes an asymmetric autoencoder and a custom trajectory clustering layer. The autoencoder is mainly composed of 1D convolutional neural network and bi-directional long short-term memory to learn the feature representation of the original input in the low-dimensional latent space. The trajectory clustering layer realizes clustering by calculating the Q distribution of samples in the hidden space. Combined with reconstruction loss of autoencoder and trajectory clustering Q distribution,a new anomaly score is defined for anomaly trajectory detection. The results of experiments using real trajectory data based on automatic dependent surveillance-broadcast(ADS-B) show that the proposed framework is effective for air traffic flow recognition and can detect anomaly trajectories that are meaningful and interpretable.

Keywords Trajectory clustering,Anomaly detection,Deep neural network,Autoencoder,ADS-B

1 引言

随着各类导航定位系统的快速发展,从不同场景中收集的海量数据为轨迹数据挖掘提供了重要支撑,进一步推动了知识发现。

轨迹聚类作为一种无监督的数据驱动方法,旨在发现具有相似轨迹的簇,自动识别轨迹中的交通流。长期以来,轨迹聚类被视为揭示运动模式、进行位置预测,以及实现其他更复杂应用的基础,对于空中交通管制具有重要意义[1]。目前已有的轨迹聚类技术通常依赖于定义合适的距离度量函数,以量化轨迹间的相似性,再应用经典聚类算法(*K*-means[2]、DBSCAN[3]、GMM[4]、谱聚类[5]等)进行聚类。基于常规的距离度量函数进行航空轨迹聚类极具挑战性,原因在于航空轨迹数据具有较高的维度,难以定义适当的相似性度量方法以同时考虑轨迹的局部和全局特征,且难以捕获低维隐空间中更丰富的依赖关系。随着深度学习在计算机视觉方面取得重大进展,DNN(Deep Neural Network)被证明能有效提取数据的重要特征[6]。本文提出了一种多变量轨迹深度聚类(Multivariable Trajectory Depth Clustering,MTDC)框架,旨在利用非对称自编码器学习原始高维轨迹数据在低维隐空间中的特征表示,并与轨迹聚类层相集成,以弥补传统聚类算法存在的缺陷。

轨迹异常检测用于发现历史轨迹数据中不符合预期的飞行模式或重大事件[7],便于安全专家进行风险评估,对于航空

到稿日期:2022-01-11 返修日期:2022-05-22
通信作者:时宏伟(shihw001@126.com)

安全监测和维护具有重要作用。本文结合 MTDC 框架中自编码器的重建误差和轨迹聚类 Q 分布定义了一个新的异常分数，以检测出更具实际意义的异常轨迹。实验中采用从 OpenSky[8] 开放数据平台获取的真实轨迹数据来验证方法的有效性。

2 相关工作

现有的轨迹聚类算法主要基于原始数据空间进行聚类，具体可分为完整轨迹聚类[9]、轨迹段聚类[10]以及轨迹点聚类[11]。目前，已有许多学者对轨迹聚类进行了大量研究。Ayhan 等[12]提出了一种基于分割、聚类、合并的飞机轨迹聚类框架，根据爬升、巡航、下降 3 个主要的飞行阶段对轨迹点进行划分和聚类，然后合并得到整个轨迹。Mahboubi 等[13]提出基于轨迹转折点聚类的交通模式识别方法，该方法在仿真数据上取得了良好的聚类效果，但难以适用于含噪音的真实轨迹数据。Li 等[14]提出了一种基于谱聚类技术的算法，对终端区域的飞行轨迹实现了有效的自动识别。

鉴于对原始数据空间进行聚类的方法往往无法捕捉轨迹的隐含特征，考虑采用深度神经网络学习原始轨迹在低维隐空间中的特征表示。Olive 等[15]结合了轨迹聚类方法来识别空域内的空中交通流量。Zeng 等[16]提出了一种基于 DAE 和 GMM 的聚类算法，通过 DAE 对原始空间中的高维轨迹数据进行特征提取，并将提取结果输入 GMM 执行聚类。相比基于原始空间聚类的方法，该方法提高了聚类效率，具有一定的优势。

异常检测是很多领域的热点研究问题[17]。在航空异常检测方面，Das 等[18]提出的 MKAD 方法可以有效检测飞行数据中的异常操作情况。Li 等[19]提出了一系列基于聚类的异常检测算法，能够在飞行过程中实时检测异常数据样本。在深度学习方向，Nielsen 等[20]提出了一种基于前馈神经网络的异常检测方法，能够从大型轨迹数据集中识别航空安全风险。

3 问题定义

3.1 轨迹聚类

轨迹本质上是一组带有空间位置信息的时间序列数据，除了基本的时空信息外，通常还携带一些特征描述信息（如速度、航向、身份标识等）。将 X_{TP} 视为 N 条轨迹的集合，则：

$$\begin{aligned}\boldsymbol{X}_{TP}&=\{\boldsymbol{x}_{TP_1},\boldsymbol{x}_{TP_2},\cdots,\boldsymbol{x}_{TP_N}\}\\ \boldsymbol{x}_{TP_i}&=\{\boldsymbol{x}_{TP_{i,1}},\boldsymbol{x}_{TP_{i,2}},\cdots,\boldsymbol{x}_{TP_{i,T}}\}\end{aligned}\tag{1}$$

其中，$\boldsymbol{x}_{TP_i}\in\mathbb{R}^{d\times T}$ 表示集合 $\boldsymbol{X}_{TP}$ 中的第 i 条轨迹，是由 T 个轨迹点组成的时间序列，d 为每个轨迹点对应的特征数。聚类即为将 $\boldsymbol{X}_{TP}$ 划分为 K 个簇 $\boldsymbol{C}=\{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_k\}$，以最大化簇内对象的相似性和簇间对象的差异性的过程。

3.2 轨迹深度聚类

深度聚类通过全连接神经网络（FCNN）、卷积神经网络（CNN）和循环神经网络（RNN）等深度神经网络（DNN）构建编码器，从而学习原始数据在低维隐空间中的特征表示。编码器是一个非线性映射函数：

$$f_\theta:\boldsymbol{X}_{TP}\rightarrow\boldsymbol{Z}_{TP}\tag{2}$$

其中，θ 为编码器的参数；$\boldsymbol{Z}_{TP}$ 为编码器从输入数据 $\boldsymbol{X}_{TP}$ 中学习到的新的表示，被称为隐空间（Latent Space），此时的聚类任务即为对当前的隐空间 $\boldsymbol{Z}_{TP}$ 进行聚类，$\boldsymbol{Z}_{TP}$ 表示为：

$$\boldsymbol{Z}_{TP}=\{\boldsymbol{z}_{TP_1},\cdots,\boldsymbol{z}_{TP_N}\}=\{f_\theta(\boldsymbol{x}_{TP_1}),\cdots,f_\theta(\boldsymbol{x}_{TP_N})\}\tag{3}$$

4 MTDC 框架

4.1 模型架构

传统自编码器（Auto-Encoder，AE）包含编码器 f 和解码器 g 两个组成部分。编码器 $f:\boldsymbol{X}_{TP}\rightarrow\boldsymbol{Z}_{TP}$ 将原始数据空间非线性映射到低维隐空间，而解码器 $g:\boldsymbol{Z}_{TP}\rightarrow\boldsymbol{X}'_{TP}$ 则将隐空间映射回原始数据空间，解码器的结构和参数与编码器相对称。图 1 给出了有 3 个隐含层的自编码器结构。通过最小化 $\boldsymbol{X}_{TP}$ 与 $\boldsymbol{X}'_{TP}$ 间的均方误差（MSE）来训练自编码器，并评估其重建能力。

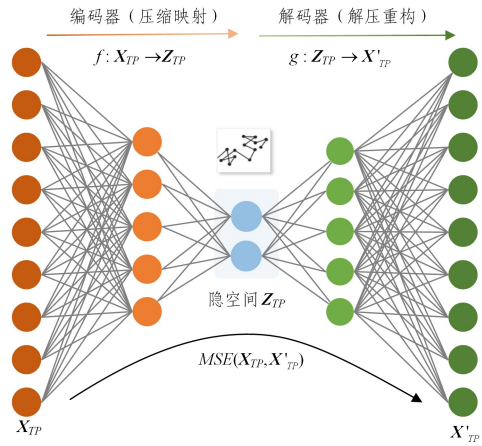


图 1 自编码器架构

Fig. 1 Autoencoder architecture

为了充分考虑轨迹的时空特性，本文将 AE 中的编码器与一维 CNN 和两个 Bi-LSTM 相结合，使编码器在降低数据维度的同时能更好地捕捉轨迹的时空特征。解码器部分首先经过 TimeDistributed 层实现二维数据到三维数据的转换，再进行上采样（UpSampling2D）和反卷积（Conv2DTranspose）操作，从而重构出原始输入。同时，本文在编码器后定义了一个轨迹聚类层（TClustering），整体构成 MTDC 框架，如图 2 所示。

MDTC 框架以单个轨迹样本为例进行描述，编码器部分包括大小为 $T\times d$ 的输入层、经过 h 个大小为 $k\times d$ 的过滤器进行一维 CNN 卷积，并采用大小为 l 的最大池化层（Max-pooling）进行下采样，从而在保留主要特征的同时对轨迹数据进行降维。使用单元数（Units）分别为 32 和 1 的两个 Bi-LSTM 网络层提取时间相关特征，以同时考虑时间序列的前向和后向数据间的时间依赖性，从而得到低维隐空间特征 $\boldsymbol{z}_{TP_i}$。$\boldsymbol{z}_{TP_i}$ 一方面作为解码器的输入，用以重建原始输入，得到输出 $\boldsymbol{x}'_{TP_i}$；另一方面作为 TClustring 层的输入，得到一个 k 维的聚类向量 $\boldsymbol{q}_i$。MTDC 架构结合非对称自编码器的重建损失与轨迹聚类层的 Q 分布和目标 P 之间的 KL 散度损失来联合训练，保证提取到的特征在保留原始输入的基本结构的同时使特征具有区分性，便于进行轨迹聚类。

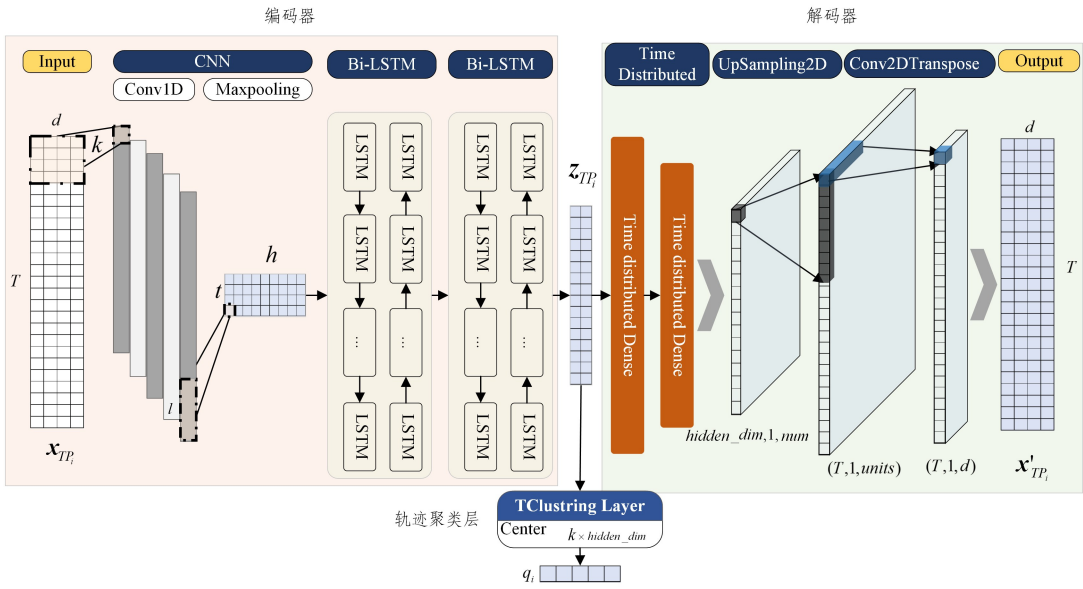


图 2　MTDC 框架结构

Fig. 2　Structure of MTDC framework

4.2　轨迹特征表示

特征学习旨在为分类、重建、可视化等寻找良好的表示。本文通过对图 2 中非对称的编码器-解码器结构进行预训练，以获取原始轨迹的初始特征表示，并采用 MSE 误差来衡量原始输入和重构结果间的差异性，称之为重建损失：

$$L_r=\frac{1}{n}\sum_{i=1}^{n}\| \boldsymbol{x}_{TP_i}-g(f(\boldsymbol{x}_{TP_i}))\|^2 \tag{4}$$

4.3　轨迹深度聚类

为了获得隐空间 $\boldsymbol{Z}_{TP}$ 的初始划分，首先对其执行聚类，得到 k 个聚类中心 $\{\boldsymbol{\mu}_j\}_{j=1}^{k}$，用于初始化轨迹聚类层权重。对于轨迹聚类层的每一个输入 $\boldsymbol{z}_{TP_i}$，使用学生 t-分布[21]作为内核度量 $\boldsymbol{z}_{TP_i}$ 与聚类中心 $\boldsymbol{\mu}_j$ 之间的相似性，得到隐空间 $\boldsymbol{Z}_{TP}$ 的 Q 分布：

$$q_{i,j}=\frac{(1+\| \boldsymbol{z}_{TP_i}-\boldsymbol{\mu}_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_{j=1}^{k}(1+\| \boldsymbol{z}_{TP_i}-\boldsymbol{\mu}_j\|^2/\alpha)^{-\frac{\alpha+1}{2}}} \tag{5}$$

其中，α 是学生 t-分布的自由度。在无监督学习中，由于无法进行交叉验证，常设 $\alpha=1$[21]。$q_{i,j}$ 可被视为轨迹 $\boldsymbol{x}_{TP_i}$ 属于集群 j 的置信度。

为了使 $q_{i,j}$ 更稳健，设定一个训练目标 $p_{i,j}$，在增强聚类效果的同时对聚类中心做标准化处理，防止大的簇干扰隐空间的特征表示。目标 P 分布定义为：

$$p_{i,j}=\frac{q_{i,j}^2/\sum_{i=1}^{N}q_{i,j}}{\sum_{j=1}^{k}(q_{i,j}^2/\sum_{i=1}^{N}q_{i,j})} \tag{6}$$

通过最小化 $p_{i,j}$ 与 $q_{i,j}$ 间的 Kullback-Leibler(KL)散度损失[22]来衡量聚类效果：

$$L_c=KL(P|Q)=\sum_{i=1}^{N}\sum_{j=1}^{k}p_{i,j}\log\frac{p_{i,j}}{q_{i,j}} \tag{7}$$

训练过程使用 SGD 优化器，通过反向传播在每一迭代时更新目标分布 P。MTDC 框架的训练目标为：

$$L_{\mathrm{MDTC}}=L_r+\gamma L_c \tag{8}$$

其中，L_r 为重建损失(见式(4))，旨在确保学习到的特征表示保持输入数据的原始结构；L_c 为聚类损失(见式(7))，旨在使隐空间中的特征表示更具区分性。$\gamma>0$ 时能防止隐空间发生扭曲。基于 MTDC 框架的轨迹深度聚类算法描述如下：

算法 1　MTDC 轨迹深度聚类算法

输入：$(\mathbf{X}_{\mathrm{TP}}),\mathbf{W},\boldsymbol{\theta},\alpha,\gamma)$

输出：$(\mathbf{X}'_{\mathrm{TP}},\mathbf{Y}_{\mathrm{pred}},\mathrm{L_r},\mathrm{L_c},\mathrm{Q})$

1. 对输入$\mathbf{X}_{\mathrm{TP}}$进行编码，提取隐空间特征表示 $\mathbf{Z}_{\mathrm{TP}}$；
2. 对$\mathbf{Z}_{\mathrm{TP}}$执行聚类，获取 k 个聚类中心$\{\boldsymbol{\mu}_j\}_{j=1}^{k}$，并初始化 TClustering 层权重；
3. 迭代训练，反向传播，微调模型；
4. for epoch in range(epochs)：
5. 　　计算$\boldsymbol{z}_{\mathrm{TP}_i}$被分配到集群 j 的概率 $q_{i,j}$
6. 　　计算目标分布 $p_{i,j}$
7. 　　if epoch % eval_epochs ==0：
8. 　　　　最小化 MTDC 损失(见式(8))
9. 　　　　更新目标分布，保存模型权重；
10. 　　end if
11. 　end for

4.4　异常检测

异常轨迹在数据集中的占比较小，因此相较于正常轨迹，其具有更大的重建误差。对于聚类任务而言，Q 分布置信度被视为每个样本属于不同集群的概率，样本的最大置信度过小时说明该轨迹不易于区分，通常被视为异常轨迹。本文同时考虑轨迹的重构效果和可区分性，联合每个轨迹样本的重构误差和 Q 分布置信度定义了一个新的异常分数：

$$\begin{aligned}&\boldsymbol{s}_i=MSE(\boldsymbol{x}_{TP_i},\boldsymbol{x}'_{TP_i})+(1-\max\{q_{i,j}(\boldsymbol{z}_{TP_i},\boldsymbol{\mu}_j)\})\\&score_i=\frac{\boldsymbol{s}_i-\boldsymbol{s}_{i_{\min}}}{\boldsymbol{s}_{i_{\max}}-\boldsymbol{s}_{i_{\min}}},score_i\in[0,1]\end{aligned} \tag{9}$$

利用 95%分位数定义阈值 δ，对于每个轨迹样本 $\boldsymbol{x}_{TP_i}$，异常判定方式为：

$$\begin{aligned}&\delta_i=percentile(score_i,0.95)\\&\begin{cases}mean(score_i(\boldsymbol{x}_{TP_i}))>\delta_i, & \boldsymbol{x}_{TP_i}\rightarrow\text{Anomal}\\ \text{otherwise}, & \boldsymbol{x}_{TP_i}\rightarrow\text{Normal}\end{cases}\end{aligned} \tag{10}$$

5 数据准备

5.1 数据集获取

旨在向研究人员提供真实的空中交通管制数据的OpenSky开放数据平台[5]迄今为止收集并存储了超过33万亿条基于ADS-B的真实航空轨迹数据，其中除经纬、高度等基本位置信息外，还包含识别号、速度、航向等共19种属性信息。

ICAO识别号为43eb7b的航空器的未处理的原始轨迹数据所含部分轨迹点的信息如表1所列。

表1 未处理的原始轨迹

Table 1 Unprocessed raw trajectory

Timestamp	icao24	Altitude	Latitude	Longitude	Track
2019-10-20 10:51:43+00:00	43eb7b	24500	46.8197	8.7825	353.7306
2019-10-20 10:51:44+00:00	43eb7b	24475	46.8218	8.7822	353.7306
2019-10-20 10:51:45+00:00	43eb7b	24475	46.8235	8.7819	353.7306
2019-10-20 10:51:46+00:00	43eb7b	24450	46.8251	8.7816	353.5532
2019-10-20 10:51:47+00:00	43eb7b	24450	46.8267	8.7814	353.5532
…	…	…	…	…	…

终端空域包含飞机起飞和降落两类场景。为验证所提方法在不同的飞行场景下的适用性。本文从OpenSky中分别获取了2019年10月至11月期间降落至苏黎世机场且距机场40海里范围内的19 480条轨迹，以及2020年9月至12月期间从阿姆斯特丹史基浦机场起飞且距机场40海里范围内的29 987条轨迹。LSZH和EHAM分别为两个机场的ICAO码，后文用其指代两个数据集的名称。LSZH机场和EHAM机场分别包含4个和12个跑道。

LSZH数据集共有5类标签，已知聚类数。此标签不参与训练，仅用于评估无监督聚类的效果。文中主要对降落轨迹量最大的14号跑道的14383条轨迹进行聚类，并检测每类轨迹中存在的异常轨迹。EHAM数据集未进行标注，聚类数未知且飞行模式相对单一，因此文中主要对起飞轨迹量最大的24号跑道的15307条轨迹进行聚类，以评估MTDC框架对未标注的数据集的有效性。数据集详细情况如表2所列。

表2 两个真实轨迹数据集：LSZH和EHAM

Table 2 Tow truthful trajectory datasets: LSZH and EHAM

数据集	LSZH	EHAM
原始轨迹量/条	19480	29987
所选跑道号	14	24
所选跑道轨迹量/条	14383	15307
轨迹量占比/%	74	51
数据集大小/(轨迹点×特征)	12853314×19	8360913×16
聚类数 k	5	未知
真实标签	含标签	未知

5.2 轨迹可视化

为了观察原始轨迹的分布，图3绘制了两个机场轨迹的投影平面图。

(a) LSZH 机场 (b) EHAM 机场

图3 LSZH机场的降落轨迹和EHAM机场的起飞轨迹

Fig. 3 Landing trajectory at LSZH airport and take-off trajectory from EHAM airport

5.3 数据预处理

由于原始轨迹的长度不一致，且存在少量空值和异常点，因此本文采用线性滤波器过滤异常点，并按每1 s取一个轨迹点进行重采样，以去除空值。

轨迹聚类时无须保留原始数据中的所有属性，因此本文选取了经度、纬度、气压高度和航向4个主要特征，将每条轨迹按128个轨迹点进行重采样，从而获得等长的轨迹数据，使其适用于神经网络模型。输入网络前将WGS84坐标系的经纬度投影到平面坐标系，并对轨迹数据进行归一化处理。

6 实验过程与结果

实验环境使用Python 3.6版本编程语言，实验平台包括PyCharm CE和Jupyter Notebook。笔记本电脑配置为MAC OS Monterey系统，4核i5 CPU，32GB内存。

6.1 评估指标

对于含真实标签的数据集，本文借助无监督聚类外部评估指标来分析聚类效果[23]。首先，将聚类数设置为真实轨迹类别数，用无监督聚类精度(Accuracy，ACC)评估MTDC聚类性能：

$$ACC=\max_{m}\frac{\sum_{i=1}^{n}\mathbf{1}y_i=m(c_i)}{n} \tag{11}$$

其中，y_i 是真实标签，c_i 是算法分配的簇，m 为聚类标签与真实标签之间的映射函数。

同时，通过标准化互信息(Normalized Mutual Information，NMI)来衡量聚类结果和真实标签间的相关性：

$$NMI(C,Y)=\frac{I(C,Y)}{\frac{1}{2}[H(C)+H(Y)]} \tag{12}$$

其中，Y 表示真实标签，C 表示集群标签，I 是互信息度量，H 为熵。

此外，对于没有真实标签的数据集，本文通过计算所有轨迹样本的平均轮廓系数(Silhouette Coefficient，SC)来评估聚类效果：

$$SC=\frac{1}{N}\sum_{i}^{N}\frac{b(\boldsymbol{x}_{TP_i})-a(\boldsymbol{x}_{TP_i})}{\max\{a(\boldsymbol{x}_{TP_i}),b(\boldsymbol{x}_{TP_i})\}},SC\in[-1,1] \tag{13}$$

其中，$a(\boldsymbol{x}_{TP_i})$ 表示样本 $\boldsymbol{x}_{TP_i}$ 与同簇内其他样本间的平均距离，$b(\boldsymbol{x}_{TP_i})$ 表示样本 $\boldsymbol{x}_{TP_i}$ 与距其最近的另一簇中的其他样本间的平均距离。

6.2　MTDC 训练

首先对非对称自编码器进行预训练，从而获取原始轨迹数据在隐空间的初始特征表示。自编码器训练过程以学习率 $\alpha=0.001$ 的 Adam 优化器进行优化，通过 MSE 损失评估自编码器的重建能力。预训练迭代次数为 1000 次、批次大小设置为 1000 时，重建误差随迭代次数的变化以及所有样本的重建误差分布如图 4 所示。

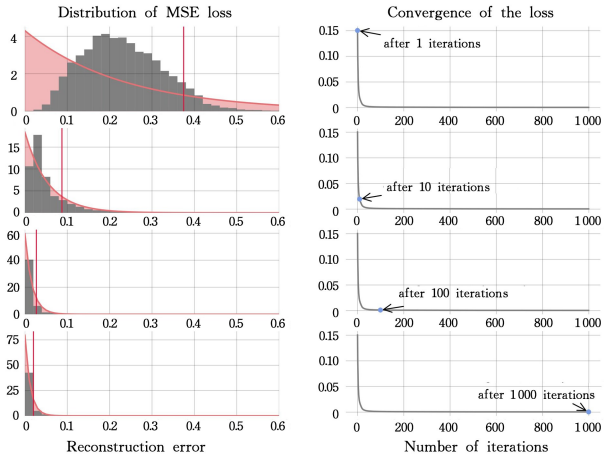


图 4　自编码器预训练中的重建误差分布

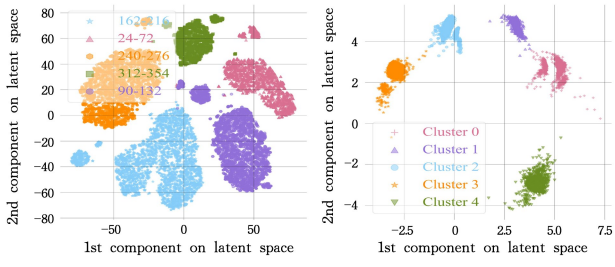
Fig. 4　Distribution of reconstruction errors during the pre-training of autoencoder

重建误差随迭代次数的增多而收敛，训练到 1000 次时，大部分样本的重建误差都接近于 0。无监督训练无法通过交叉验证评估结果的准确性，但该趋势可以表明整个训练过程的收敛性。预训练过程学习到的初始特征表示不一定适用于聚类任务。为了学习到一组更加具有判别性的特征表示，本文联合重建损失和聚类损失共同训练 MTDC 框架，同时对编码器参数进行微调。

6.3　交通流识别

6.3.1　LSZH 降落轨迹数据集

对于 LSZH 数据集，聚类数即为初始类别数。为了选择合适的 KL 损失系数，图 5 分别绘制了 $k=5$，$\gamma=0$ 和 $\gamma=0.5$ 时隐空间的二维可视化对比。可见，聚类损失的引入有利于防止隐空间发生扭曲，有益于进一步聚类。



(a)$\gamma=0$　(b)$\gamma=0.5$

图 5　$\gamma=0$ 和 $\gamma=0.5$ 时隐空间的 UMAP 可视化对比

Fig. 5　UMAP visualization comparison with $\gamma=0$ and $\gamma=0.5$

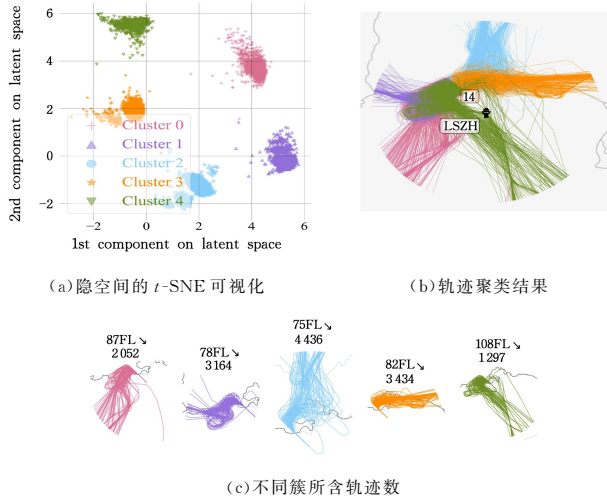
聚类损失的增加会干扰隐空间的特征表示，导致模型整体损失增大[24]。实验中测试了 γ 分别为 0.1，0.5，0.7 和 1.0 时，经过 1000 次训练后模型损失和聚类 ACC 的对比，结果如表 3 所列，易知 $\gamma=0.1$ 时效果最好。由于实验未设置验证集，无法采取网格搜索探索最优参数，因此实验中参照图像深度聚类中的常用值 $\gamma=0.1$，并将其作为 KL 散度损失系数[25-26]。

表 3　γ 对聚类效果的影响

Table 3　Effect of γ on clustering effect

γ	0.1	0.5	0.7	1.0
ACC/%	96.9	96.3	96.0	95.5
MTDC 损失均值	0.021	0.075	0.954	0.172
是否稳定收敛	是	否	否	否

图 6 绘制了 $\gamma=0.1$ 时的聚类结果，为了便于观察，图中仅绘制每类中的前 500 条轨迹。



(a)隐空间的 t-SNE 可视化　(b)轨迹聚类结果

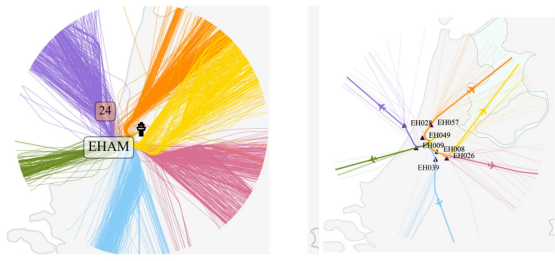
(c)不同簇所含轨迹数

图 6　LSZH 机场原始轨迹及其聚类结果

Fig. 6　Original trajectory and its clustering results at LSZH airport

6.3.2　EHAM 起飞轨迹数据集

对于无标签的 EHAM 数据集，$k=6$ 时的聚类结果如图 7(a)所示。图 7(b)绘制了轨迹聚类中心，并与 EHAM 机场进出点 ATS 航路信号标进行匹配，结果显示两者基本一致，表明 MTDC 框架对无标签轨迹聚类的有效性。



(a)聚类结果　(b)聚类中心

图 7　EHAM 机场轨迹聚类结果

Fig. 7　Trajectory clustering results at EHAM airport

6.3.3　聚类结果评估

实验中对比分析了核主成分分析(Kernel Principal Component Analysis，KPCA)降维技术以及统一流形逼近与投影(Uniform Manifold Approximation and Projection，UMAP)

降维技术对原始轨迹进行先降维再聚类的结果，并对比了基于深度学习特征提取(AE、本文 MTDC)进行聚类的结果。采用不同方法对 LSZH 和 EHAM 执行 20 次重复实验的平均结果如表 4 所列。

表 4 不同方法的聚类性能对比

Table 4 Comparison of clustering performance of different methods

方法	LSZH (k=5)		EHAM (k=6)
	ACC/%	*NMI*/%	*SC*
UMAP+k-means	87	85	0.47
KPCA+k-means	89	88	0.49
AE+k-means[15]	92	90	0.51
MTDC	**96**	**96**	**0.59**

表 4 表明，与基于降维技术的方法相比，基于深度学习特征提取的方法(AE、本文 MTDC)取得了更好的检测效果。而 MTDC 框架更能捕获轨迹时空特性，对两种场景(降落和起飞)下的轨迹均有较好的聚类准确度，多次重复实验的结果也更稳定。

6.3.4 k 值的选择

含标签的 LSZH 数据的 k 值为标签类别数，而对于不含标签的 EHAM 数据集，文中通过计算平均轮廓系数(见式(13))来选择最适合的 k 值。图 8 给出了使用不同方法对 EHAM 数据集执行聚类时，k 值的变化对聚类结果的影响。

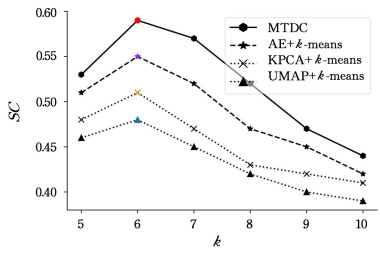


图 8 聚类结果的轮廓系数随 k 值的变化

Fig. 8 Silhouette coefficient of clustering result varies with k

6.4 异常检测

飞机着陆过程相对于起飞过程有更多的不可控因素，常出现盘旋等待、碰撞规避等行为，因此本文主要对 LSZH 机场终端的飞机降落轨迹进行异常检测。

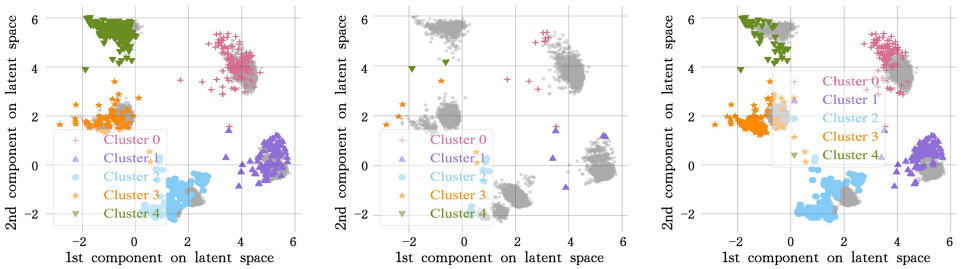
6.4.1 异常检测结果

将重建误差作为异常分数时，核心思想是异常轨迹拥有较大的重建误差。而将聚类 Q 分布置信度作为异常分数，考虑的是异常轨迹不易于区分类别。单独通过这两种方式得到的异常检测结果会有所偏颇，表 5 列出了分别使用 MES 误差、Q 分布置信度和本文提出的式(9)作为异常分数时的轨迹聚类结果准确度，以及检测出的异常轨迹的均值、最大值、最小值和异常轨迹量的详细情况。图 9 和图 10 则分别绘制了隐空间二维可视化后异常检测结果的分布以及检测结果中异常值排名前 10 的轨迹的投影平面图，从而直观地分析检测效果。

表 5 LSZH 中不同类别的异常轨迹占比情况

Table 5 Proportion of anomaly trajectories in different categories in LSZH

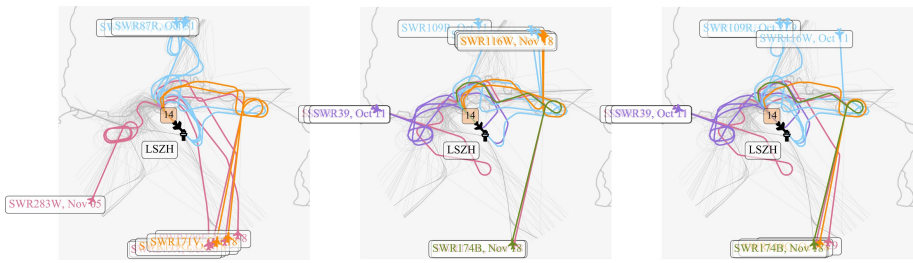
簇	*MSE* 损失			总量	Q 分布			总量	本文的异常分数			总量
	Mean	*Min*	*Max*		*Mean*	*Min*	*Max*		*Mean*	*Min*	*Max*	
0	0.058	0.011	0.980	135	0.921	0.461	0.937	16	0.066	0.035	0.560	103
1	0.024	0.002	0.571	133	0.932	0.533	0.947	5	0.044	0.026	0.518	121
2	0.024	0.000	0.982	97	0.938	0.064	0.958	10	0.040	0.019	0.854	93
3	0.027	0.002	1.000	88	0.988	0.000	1.000	6	0.017	0.000	1.000	84
4	0.068	0.015	0.732	150	0.912	0.317	0.933	2	0.076	0.042	0.512	98



(a)MSE 损失 (b)Q 分布 (c)本文提出的异常分数

图 9 不同异常分数检测出的异常轨迹分布

Fig. 9 Rajectory distribution in latent space with different outlier scores



(a)MSE 损失 (b)Q 分布 (c)本文提出的异常分数

图 10 不同异常分数检测出的前 10 条轨迹

Fig. 10 Top 10 trajectories detected with different anomaly scores

本文提出的异常分数能够避免异常检测时遗漏可区分度高而重建效果差的轨迹，以及重建效果好而可区分度差的轨迹。

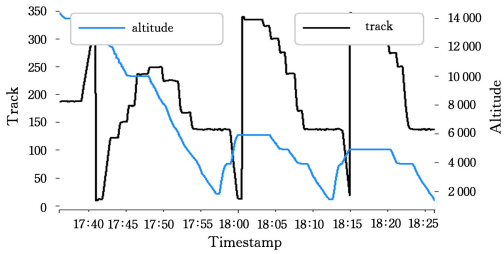
6.4.2　异常轨迹解释

为了证明异常检测结果的真实性，本文对所提出的异常分数检测出的异常值最高的前 3 条轨迹作解释分析。

(1)航班 EWG7ME

航班 EWG7ME 的轨迹主要在两个位置呈现异常。1)图 11(a)所示黑色位置处，为了规避与航班 SWR117P 发生冲突而改变航向；2)该航班在此次降落过程中尝试 3 次着陆。其航向及海拔高度随时间的变化情况如图 11(b)所示。

(a)轨迹



(b)高度和航向角随时间的变化情况(电子版为彩图)

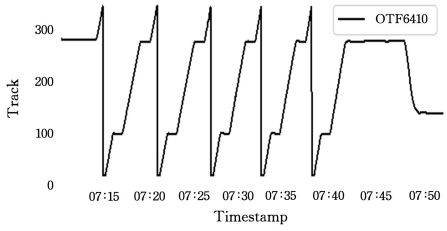
图 11　航班 EWG7ME 的轨迹随时间的变化情况

Fig. 11　Trajectory of flight EWG7ME changes over time

(2)航班 OTF6410

如图 12 所示，呼号为 OTF6410 的航班保持了 6 个等待模式后才被允许继续前往进行着陆。

(a)轨迹



(b)航向角随时间的变化情况

图 12　航班 OTF6410 的保持等待模式

Fig. 12　Hold-and-Wait mode of flight OTF6410

(3)航班 SWR1271

航班 SWR12721 检测为异常的原因在于其改变了预期的航路。迫使其改变航路的主要原因是受当天气候影响，机场区域出现积云和雷雨天气。本文从 NOAA 气象网站获取了 SWR119G 飞行当天 LSZH 机场周围 2019 年 11 月 5 日 15:00－16:00 之间的气候 METAR 信息记录，如图 13 所示。

记录 1：METAR LSZH 051550Z 01008KT 9999-**SHRA** FEW015 FEW028TCU SCT030 **BKN060** 08/06 Q0999 RETSRA NOSIG=

记录 2：METAR LSZH 051520Z 12005KT 080V180 9999 **TS** FEW022 FEW028CB **BKN038** 11/05 Q0999 TEMPO **TSRA**=

图 13　LSZH 机场周围 2019 年 11 月 5 日 15:00－16:00 之间的气候 METAR 信息记录

Fig. 13　Climate METAR information records around LSZH airport between 15:00－16:00 on November 5, 2019

图 13 中记录 1 为 2019 年 11 月 5 日下午 15:20 的气候记录，其中 SHRA 表示暴雨天气，BKN 表示云覆盖量达到 5/8 至 7/8，空中能见度低。记录 2 为 2019 年 11 月 5 日下午 14:50 分的气候记录，其中 TSRA 表示恶劣雷雨天气，BKN 表示云覆盖量达到 5/8 至 7/8，空中能见度低。此时，SWR12721 的轨迹绕行情况如图 14 所示，可以看出该时间段内有许多轨迹由于气候因素而产生变道绕行的行为。

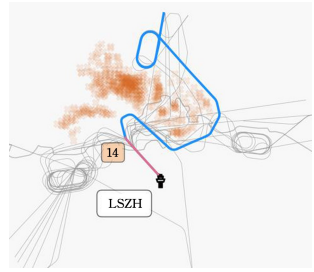


图 14　航班 SWR12721 的变道绕行

Fig. 14　Route change of flight SWR12721

结束语　针对传统聚类方法在处理大量高维数据时性能较差，且不能很好地捕获低维空间的依赖关系的问题，文本提出了一个基于 DNN 的 MTDC 深度聚类框架，并将其用于航空交通流识别与异常检测。对两个不同场景的终端数据集进行聚类的结果表明，传统轨迹聚类方法由于聚类中心的随机初始化，因此结果不稳定，平均聚类效率不佳，而本文提出的框架在相同参数下的多次实验中结果相差均小于 0.01，具有更高的准确性和稳健性。此外，本文基于 MDTC 框架，结合自编码器重建误差和轨迹聚类层的 Q 分布定义了一个新的异常分数，以检测出更加符合实际情况的异常轨迹，对检测结果的解释分析证明了异常检测结果的真实可靠性。然而，本文的研究还存在一些不足之处，在后续的研究中，将考虑不同相似性度量方法以及气候因素对聚类框架的影响。在异常检测方面，考虑使用更具鲁棒性的生成模型，从而进一步对本文提出的框架进行改进，以适应更加复杂的空域场景。

参 考 文 献

[1] WANG D, MIWA T, MORIKAWA T. Big trajectory data mining: A survey of methods, applications, and services[J]. Sensors, 2020, 20(16): 4571.

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

[3] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with

noise[C]// KDD. 1996:226-231.
[4] BARRATT S T, KOCHENDERFER M J, BOYD S P. Learning probabilistic trajectory models of aircraft in terminal airspace from position data[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 20(9):3536-3545.
[5] ENRIQUEZ M. Identifying temporally persistent flows in the terminal airspace via spectral clustering[C]// Tenth USA/Europe Air Traffic Management Research and Development Seminar(ATM2013)/Federal Aviation Administration(FAA) and EUROCONTROL. Chicago, IL, USA, 2013:10-13.
[6] LI S, ZHAO H. A Survey on Representation Learning for User Modeling[C]// IJCAI. 2020:4997-5003.
[7] FANG Z, DU Y, CHEN L, et al. E 2 DTC: An End to End Deep Trajectory Clustering Framework via Self-Training[C]// 2021 IEEE 37th International Conference on Data Engineering(ICDE). IEEE, 2021:696-707.
[8] SCHÄFER M, STROHMEIER M, LENDERS V, et al. Bringing up OpenSky: A large-scale ADS-B sensor network for research [C]// IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks. IEEE, 2014: 83-94.
[9] YANAGISAWA Y, SATOH T. Clustering multidimensional trajectories based on shape and velocity[C]// 22nd International Conference on Data Engineering Workshops(ICDEW'06). IEEE, 2006.
[10] CHEN J Y, SONG J T, LIU L X, et al. Trajectory clustering algorithm based on improved hausdorff distance[J]. Computer Engineering, 2012, 38(17):157-161.
[11] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence databases[C]// International Conference on Foundations of Data Organization and Algorithms. Springer, 1993:69-84.
[12] AYHAN S, SAMET H. Diclerge: Divide-cluster-merge framework for clustering aircraft trajectories[C]// Proceedings of the 8th ACM SIGSPATIAL International Workshop on Computational Transportation Science. 2015:7-14.
[13] MAHBOUBI Z, KOCHENDERFER M J. Learning traffic patterns at small airports from flight tracks[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 18(4):917-926.
[14] LI C Q, LI S M, MA W Y, et al. Research on automatic identification method of air traffic flow based on trajectory clustering [J]. Computer Simulation, 2021, 38(10):73-77.
[15] OLIVE X, BASORA L. Identifying anomalies in past en-route trajectories with clustering and anomaly detection methods [C]// ATM Seminar. 2019.
[16] ZENG W, XU Z, CAI Z, et al. Aircraft Trajectory Clustering in Terminal Airspace Based on Deep Autoencoder and Gaussian Mixture Model[J]. Aerospace, 2021, 8(9):266.
[17] BASORA L, OLIVE X, DUBOT T. Recent advances in anomaly detection methods applied to aviation[J]. Aerospace, 2019, 6(11):117.
[18] DAS S, MATTHEWS B L, SRIVASTAVA A N, et al. Multiple kernel learning for heterogeneous anomaly detection: algorithm and aviation safety case study[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010:47-56.
[19] LI L, HANSMAN R J, PALACIOS R, et al. Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring[J]. Transportation Research Part C: Emerging Technologies, 2016, 64:45-57.
[20] JANAKIRAMAN V M, NIELSEN D. Anomaly detection in aviation data using extreme learning machines[C]// 2016 International Joint Conference on Neural Networks(IJCNN). IEEE, 2016:1993-2000.
[21] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]// International Conference on Machine Learning. PMLR, 2016:478-487.
[22] ZENG J, KRUGER U, GELUK J, et al. Detecting abnormal situations using the Kullback-Leibler divergence[J]. Automatica, 2014, 50(11):2777-2786.
[23] RENDÓN E, ABUNDEZ I, ARIZMENDI A, et al. Internal versus external cluster validation indexes[J]. International Journal of Computers and Communications, 2011, 5(1):27-34.
[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6):84-90.
[25] GHASEDI DIZAJI K, HERANDI A, DENG C, et al. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:5736-5745.
[26] GUO X, GAO L, LIU X, et al. Improved deep embedded clustering with local structure preservation[C]// IJCAI. 2017:1753-1759.

RAO Dan, born in 1996, postgraduate. Her main research interests include big data and data mining.

SHI Hongwei, born in 1965, professor. His main research interests include intelligent decision based on big data, aviation safety big data, UAV intelligent information processing and air traffic management ATM/CNS.

(责任编辑:何杨)