

9-17-2021

Hurricane Trajectory Outlier Detection Method Based on Variational Auto-encode

Wanting Qin

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China;

Songyang Lao

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China;

Tang Jun

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China;

Lu Cong

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China;

Follow this and additional works at: <https://dc-china-simulation.researchcommons.org/journal>



Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, Numerical Analysis and Scientific Computing Commons, Operations Research, Systems Engineering and Industrial Engineering Commons, and the Systems Science Commons

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

Hurricane Trajectory Outlier Detection Method Based on Variational Auto-encode

Abstract

Abstract: Hurricanes often cause incalculable human and economic losses, and the trajectory outlier detection can provide the auxiliary information or abnormal warning of the disaster. *On deep learning, a method of hurricane trajectory outlier detection based on variable auto encoder (VAEOD) is proposed in this paper. The trajectory is divided into equal sequence sub trajectories based on the sliding window as the input of VAE. The trajectory reconstruction model is trained by the VAE. The parallel, vertical and angle distance of reconstructed trajectory and the input trajectory are compared to find out the outlier trajectory segments.* The simulation experiment on real hurricane data shows that the VAEOD method is more rational and practical than the classical TRAOD method.

Keywords

variational auto-encoder, space distance, sliding window, outlier detection

Recommended Citation

Qin Wanting, Lao Songyang, Tang Jun, Lu Cong. Hurricane Trajectory Outlier Detection Method Based on Variational Auto-encode[J]. Journal of System Simulation, 2021, 33(9): 2191-2201.

基于变分自编码器的飓风轨迹异常检测方法

秦婉亭, 老松杨, 汤俊, 卢聪

(国防科技大学 系统工程学院, 湖南 长沙 410073)

摘要: 飓风常会引起无法估计的人身和经济损失, 飓风轨迹异常检测可以为灾情提供辅助信息或异常告警。从深度学习的角度出发, 提出了基于变分自编码器的飓风轨迹异常检测方法(Variational Auto-Encoder Outlier Detection, VAEOD)。利用滑动窗口将长度不等的轨迹序列变成等长的子轨迹序列作为变分自编码器的输入, 通过变分自编码器训练轨迹重构模型, 将重构的轨迹与输入的轨迹通过平行、垂直和角度距离进行对比找出异常的轨迹段。通过真实的飓风数据进行仿真实验发现, VAEOD 方法比经典的轨迹异常值检测算法(Trajectory Outlier Detection Algorithm, TRAOD)检测的结果更合理, 实用性更强。

关键词: 变分自编码器; 空间距离; 滑动窗口; 异常检测

中图分类号: TP391.9

文献标志码: A

文章编号: 1004-731X (2021) 09-2191-11

DOI: 10.16182/j.issn1004731x.joss.20-0369

Hurricane Trajectory Outlier Detection Method Based on Variational Auto-encode

Qin Wanting, Lao Songyang, Tang Jun, Lu Cong

(College of Systems Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: Hurricanes often cause incalculable human and economic losses, and the trajectory outlier detection can provide the auxiliary information or abnormal warning of the disaster. *On deep learning, a method of hurricane trajectory outlier detection based on variable auto encoder (VAEOD) is proposed in this paper. The trajectory is divided into equal sequence sub trajectories based on the sliding window as the input of VAE. The trajectory reconstruction model is trained by the VAE. The parallel, vertical and angle distance of reconstructed trajectory and the input trajectory are compared to find out the outlier trajectory segments. The simulation experiment on real hurricane data shows that the VAEOD method is more rational and practical than the classical TRAOD method.*

Keywords: variational auto-encoder; space distance; sliding window; outlier detection

引言

飓风、台风和旋风被认为是发生在热带海洋上的热带气旋。飓风一般是指发生在大西洋和北太平洋东部地区产生的强大而深厚的热带气旋^[1-2]。台风通常是指发生在西北太平洋和中国南海的热带气旋, 而发生在印度洋、阿拉伯海和孟加拉湾地区的热带气旋常被称作旋风。2019 年 10 月 12 日,

台风“海贝思(Hagibis)”在日本静冈县的伊豆半岛登陆, 此次台风造成日本 80 人死亡, 11 人失踪, 3 900 多人受到影响而避难, 造成的损失不计其数。飓风这类自然灾害发生后所带来的损失是巨大的, 如果仅仅是在灾害来临时采取紧急处置措施已经无法满足当前的需求。因此为了减少飓风所带来的危害, 急需从事后处置向事前异常预警转型。轨迹

收稿日期: 2020-06-17 修回日期: 2020-08-23

基金项目: 青年人才托举项目(17JCJQQQ048); 学校科研计划重点项目(ZK18-02-12); 湖湘青年人才(2018RS3079); 湖南省研究生创新项目(CX20190046)

第一作者: 秦婉亭(1995-), 女, 博士生, 研究方向为轨迹数据挖掘。E-mail: qinwt01@sina.com

异常检测的目的主要是从大量的轨迹数据中发掘不同于普通的, 常见的具有共性特征的异常模式^[3-5], 如形状或行进时间与其他轨迹有显著差异^[3], 异常数据表示数据内部的隐含特征, 通常伴随着有趣的事件^[4-5]。在对轨迹异常模式挖掘的过程中可以有效地发现潜在风险冲突, 提前准备措施, 避免一些意外事故的发生^[6-7]。通过对飓风轨迹进行监测和记录, 发现飓风运动过程中的异常现象可以为飓风的灾情预报提供辅助。

常见的异常检测方法有基于距离的检测方法^[8-9]、基于密度的检测方法^[10-11]、基于网格划分的检测方法^[12-13]和基于分类的检测方法^[14-15]。基于距离的检测方法是最常用的一种方法, 该方法依据某种距离度量标准, 在历史轨迹集中查找待检测轨迹的相邻轨迹数量, 当相邻轨迹数量少于某一阈值时判定为异常。基于距离检测方法的阈值选取通常是一个全局阈值, 这种方式无法考虑到局部的异常, 因此有学者提出了基于密度的检测方法。基于网格划分的检测方法则是将移动对象的运动区域划分成同等大小网格区域, 从这些网格区域中识别出异常的网格序列。基于分类的检测方法大致分成 2 个阶段: 训练阶段和测试阶段。训练阶段通过使用带标签的数据训练学习分类器; 测试阶段依据训练好的分类器将测试数据分成正常模式和异常模式, 但是这种方法通常需要数据携带异常标签。相对于无监督学习, 基于分类的检测方法更加的准确。然而, 在标注数据的时候需要专家进行人工标注, 若想获得精确的数据标签需花费大量的人力和时间。文献[15]提出了基于支持向量机(Support Vector Machine, SVM)的分类方法, 它解决了数据可以没有标签的问题, 该方法的理论前提是异常数据在全体数据集中只占极少数, 如果目标轨迹被分类器判定为异常, 则它是异常轨迹。

神经网络在轨迹数据挖掘中得到了广泛的应用, 主要集中在轨迹预测、轨迹聚类分析、轨迹相似度分析等领域^[16], 和传统方法相比, 神经网络可以从数据中自动的学习数据中的特征。本

文以飓风轨迹数据为研究对象, 通过无监督学习的方式实现飓风轨迹的异常检测。自编码器^[17]是由编码器和解码器构成的无监督学习算法, 旨在从大量无标记的数据中学习数据的有效信息, 实现对输入数据的非线性压缩和重构, 该方法可以很好地适用数据没有异常标签的轨迹异常检测问题。本文结合一种改进的自编码器即变分自编码器(Variational Auto-Encoder, VAE)的特性提出了基于变分自编码器的飓风轨迹异常检测方法。变分自编码器^[18-19]是一种优化的自编码器, 在保持自编码器基本功能的基础上, 将深度学习方法与贝叶斯结合在一起。基于变分自编码器的飓风轨迹异常检测方法通过变分自编码器训练轨迹重构模型, 将重构的轨迹和原始轨迹在空间距离上进行对比来检测异常的轨迹段。

1 轨迹距离和自编码器

已知轨迹数据集记录了移动对象在不同时刻下的空间位置信息, 这些有序的轨迹点组成的集合称为轨迹 Trj 。

定义 1 轨迹, 指在地理空间位置上移动的对象, 其位置信息随着时间的变化而变化, 由 m 个离散位置信息所构成的有序序列:

$$Trj = \{(lon_1, lat_1), (lon_2, lat_2), \dots, (lon_m, lat_m)\} \quad (1)$$

式中: lon_m, lat_m 分别为轨迹 Trj 第 m 个轨迹点的坐标, 即经度和纬度。

定义 2 子轨迹, 指在某条轨迹范围内轨迹点的有序集合:

$$seq = \{(lon_m, lat_m), (lon_{m+1}, lat_{m+1}), \dots, (lon_{m+k-1}, lat_{m+k-1})\} \quad (2)$$

式中: k 为子轨迹的长度。

定义 3 轨迹段, 指轨迹空间位置上相邻坐标点连接形成的线段:

$$L_i = \{s_i(lon_i, lat_i), e_i(lon_{i+1}, lat_{i+1})\} \quad (3)$$

式中: s_i 为轨迹段 L_i 的起始坐标; e_i 为轨迹段 L_i 的终止坐标。轨迹和子轨迹由无数个轨迹段组成。

1.1 轨迹距离定义

常见的距离度量方法如欧式距离, 常用于寻找异常的轨迹点, 而不能用于寻找异常的轨迹段。本文主要目的是寻找异常轨迹段, 因此对 2 个轨迹段之间的距离采用空间距离的度量方法。常见的空间距离度量方式有垂直距离(d_{\perp})、平行距离($d_{//}$)和角度距离(d_{θ})^[8-9]。轨迹段 $L_i=s_i e_i$ 和 $L_j=s_j e_j$, 其中 L_i 是较长的轨迹段, L_j 是较短的轨迹段, 则 L_i 和 L_j 的垂直距离、平行距离和角度距离如图 1 所示。

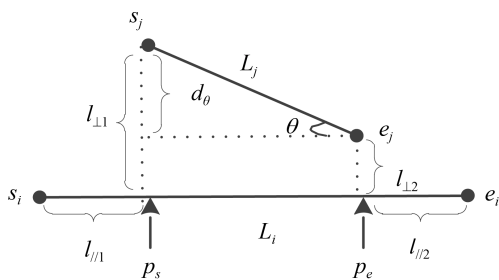


图 1 轨迹段空间距离

Fig. 1 Space distance definition of trajectory segment

定义 4 垂直距离(d_{\perp}), L_i 和 L_j 的垂直距离为

$$d_{\perp}(L_i, L_j) = \frac{l_{\perp 1}^2 + l_{\perp 2}^2}{l_{\perp 1} + l_{\perp 2}} \quad (4)$$

式中: $l_{\perp 1}$ 为点 s_j 到轨迹段 L_i 的垂直距离; $l_{\perp 2}$ 为点 e_j 到轨迹段 L_i 的垂直距离。

定义 5 平行距离($d_{//}$), L_i 和 L_j 的平行距离为

$$d_{//}(L_i, L_j) = \min(l_{//1}, l_{//2}) \quad (5)$$

式中: $l_{//1}$ 为点 s_i 到 p_s 的距离; p_s 为点 s_j 到轨迹段 L_i 的垂足; $l_{//2}$ 为点 e_j 到 p_e 的距离; p_e 为点 e_j 到线段 L_i 的垂足。

定义 6 角度距离(d_{θ}), L_i 和 L_j 的角度距离为

$$d_{\theta}(L_i, L_j) = \begin{cases} \|L_j\| \times \sin \theta, & \text{if } 0^\circ < \theta < 90^\circ \\ \|L_j\|, & \text{if } 90^\circ < \theta < 180^\circ \end{cases} \quad (6)$$

其中: $\|L_j\|$ 为轨迹段 L_j 的长度; θ 为轨迹段 L_j 和轨迹段 L_i 之间的夹角。本文把线段之间的角度距离简化为了 $\|L_j\| \times \sin \theta$ 。

1.2 自编码器

无监督学习无需数据标签, 可以自动学习数据的特征表达, 因此这种优势广泛应用在异常检测当中, 如图片异常检测、视频异常检测和时序数据异常检测^[20-21]。自编码器是一种把输入复制到输出的神经网络模型^[22], 在没有目标的情况下可以寻找数据变换关系。自编码器的输入和输出形式均是 $X=[X_1, X_2, \dots, X_n]$, 其中 X_i 表示轨迹点, 形状为 $seq_len \cdot dim$, 其中 seq_len 表示轨迹的长度, dim 表示轨迹的维度, 本文轨迹的维度是即经度和纬度 2 维。自编码器包括编码器和解码器两部分。编码器: 将轨迹输入映射到潜在空间表示, 如式(7)所示; 解码器: 通过隐藏空间的表示重构输入作为输出, 如式(8)所示。由于潜在空间的数据相对原始空间存在信息压缩, 因此自编码器可以学习数据的主要信息, 并根据这些主要信息重构数据。基于长短期记忆网络(Long Short-Term Memory, LSTM)的编码器/解码器基本结构如图 2 所示。

$$Z = \sigma_e(W^{(1)}X + b^{(1)}) \quad (7)$$

$$X' = \sigma_d(W^{(2)}z + b^{(2)}) \quad (8)$$

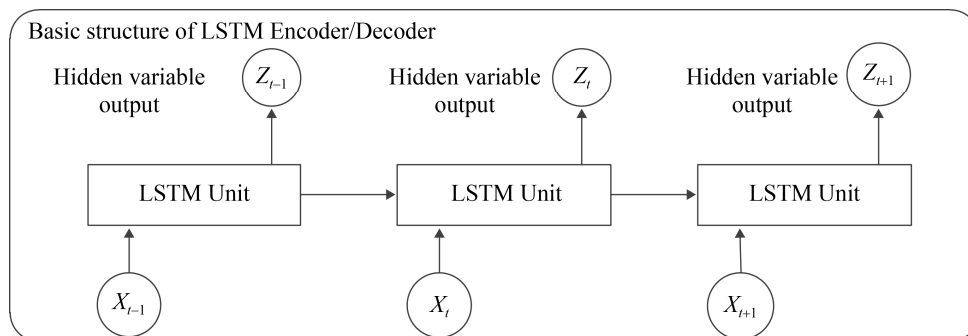


图 2 基于 LSTM 的编码器/解码器基本结构

Fig. 2 Basic structure of encoder/decoder based on LSTM

2 基于变分自编码器的轨迹异常检测

本文提出了一种基于变分自编码器的轨迹异常检测方法 (Variational Auto-Encoder Outlier Detection, VAEOD)。VAEOD 方法主要包括 3 部分: ①基于滑动窗口划分轨迹序列; ②利用变分自编码器训练轨迹重构模型; ③通过重构轨迹与原始轨迹之间的空间距离识别异常轨迹段。

2.1 基于滑动窗口的轨迹序列分段

轨迹序列是一个长度不等的时序序列, 要把轨迹序列变成变分自编码器可接受的输入格式, 则需要将轨迹划分成等长的子轨迹进行输入。基于滑动窗口的轨迹分段将原始时间序列创建为固定大小的段, 不仅能够保证将持续时间较长的数据模式完整的分割出来, 还保持了原有时间序列数据在时序上的依赖性。该分段方式具体实现为: 设包含 N 个多维的轨迹序列样本 $X = [X_1, X_2, \dots, X_n]$, 其中单个长度为 T 的样本为 $X^n = (x_0^n, x_1^n, \dots, x_i^n, \dots, x_T^n)$, x_i^n 表示第 n 个样本 i 时刻的轨迹点, 包含经度和纬度。样本 X^n 应用滑动窗口, 其中窗口为 w , 时间步为 s , 得到的分段为

$$f(X^n) = \begin{bmatrix} X_1^n \\ X_2^n \\ \vdots \\ X_i^n \\ \vdots \\ X_d^n \end{bmatrix} = \begin{bmatrix} [x_1^n, x_2^n, \dots, x_{1+w}^n] \\ [x_{1+s}^n, x_{2+s}^n, \dots, x_{1+sw}^n] \\ \vdots \\ [x_{1+(i-1)s}^n, x_{2+(i-1)s}^n, \dots, x_{1+(i-1)s+w}^n] \\ \vdots \\ [x_{1+(d-1)s}^n, x_{2+(d-1)s}^n, \dots, x_T^n] \end{bmatrix} \quad (9)$$

2.2 基于变分自编码器的轨迹重构

变分自编码器(VAE)是一类重要的生成模型, 它的功能和结构同自编码器相似, 可以实现对轨迹的重构^[18]。自编码器将数据压缩成潜在空间的固定编码, 而 VAE 是将数据转换成统计分布的参数, 即平均值和标准差(μ, σ)。VAE 的本质是假设这些轨迹数据都是由统计过程生成的, 而 VAE 的编码和解码应该将这些随机过程考虑进去, 然后 VAE

利用平均值和方差这 2 个的参数从隐变量空间中随机选取一个元素, 并将这个元素解码为原始的输入^[19]。本文通过变分自编码器建立轨迹重构模型, 变分自编码器的概率图模型如图 3 所示。

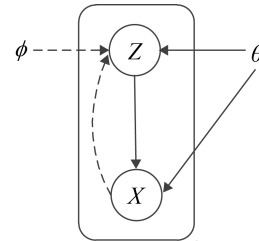


图 3 变分自编码器概率图模型
Fig. 3 VAE probability model

本文训练的轨迹数据是 X , 而最终生成的 X 是由隐变量 Z 产生, 因此由 $Z \rightarrow X$ 生成模型 $p_\theta(X|Z)$, 从自编码器的角度来看就是解码器。而 $X \rightarrow Z$ 是识别模型 $q_\phi(Z|X)$, 类似于自编码器的编码器。当 $p_\theta(X|Z)$ 训练好了以后, 将 VAE 应用于生成轨迹。

VAE 的公式原理详见文献[18]。首先假设所有的数据 $X = [X_1, X_2, \dots, X_n]$ 都是独立同分布的, 利用对数最大似然法进行参数估计:

$$\lg p_\theta(X_1, X_2, \dots, X_N) = \sum_{i=1}^N \lg p_\theta(X_i) \quad (10)$$

VAE 用识别模型 $q_\phi(Z|X_i)$ 去逼近真实的后验概率 $p_\theta(Z|X_i)$, 衡量 2 个分布的相似程度一般采用 KL 散度:

$$\begin{aligned} KL(q_\phi(Z|X_i) \| p_\theta(Z|X_i)) &= \\ E_{q_\phi(Z|X_i)} \lg \frac{q_\phi(Z|X_i)}{p_\theta(Z|X_i)} &= \\ E_{q_\phi(Z|X_i)} \lg \frac{q_\phi(Z|X_i)}{p_\theta(Z, X_i)} + \lg p_\theta(X_i) & \quad (11) \end{aligned}$$

结合式(10), (11)得到单个数据的边缘概率:
 $\lg p_\theta(X_i) =$

$$KL(q_\phi(Z|X_i) \| p_\theta(Z|X_i)) + \ell(\theta, \phi; X_i) \quad (12)$$

式中: $\ell(\theta, \phi; X_i) = -E_{q_\phi(Z|X_i)} \lg \frac{q_\phi(Z|X_i)}{p_\theta(Z, X_i)}$ 。

由于 KL 散度非负, 当 2 个分布一致时, KL 散度为 0, 于是 $\lg p_\theta(X_i) \geq \ell(\theta, \phi; X_i)$, 因此

$\ell(\theta, \phi; X_i)$ 称为对数似然函数的变分下界。现在从优化 $\lg p_\theta(X_i)$ 转化为优化它的下界 $\ell(\theta, \phi; X_i)$ 。变分自编码器的优化目标:

$$\ell(\theta, \phi; X_i) = -KL(q_\phi(Z | X_i) \| p_\theta(z)) + E_{q_\phi(Z | X_i)} \lg p_\theta(X_i | Z) \quad (13)$$

式(13)采用蒙特卡罗法进行参数估计时采用一个重参数化的技巧, 已知 $Z \sim q_\phi(Z | X)$, 那么可以引入一个新的分布 $Z = g_\phi(\xi, X)$, ξ 称为辅助变量, 这样采用蒙特卡罗法进行参数估计时保证 $q_\phi(Z | X_i)$ 的可导。则对式(13)得到的估计为

$$\tilde{\ell}(\theta, \phi; X_i) = -KL(q_\phi(Z | X_i) \| p_\theta(z)) + \frac{1}{L} \sum_{i=1}^L (\lg p(X_i | Z_{(i,l)})) \quad (14)$$

式中: $Z_{(i,n)} = g_\phi(\xi_{(i,n)}, X_i), \xi_{(i,n)} \sim p(\xi)$ 。

对数似然函数的下界可以通过 mini-batch 来估计:

$$\ell(\theta, \phi; X) \approx \tilde{\ell}^M(\theta, \phi; X^M) = \frac{N}{M} \sum_{i=1}^M \tilde{\ell}(\theta, \phi; X_i) \quad (15)$$

VAE 的本质是通过不断的训练为每一个 X_i 找到一个合适的概率分布参数 $\theta=(\mu, \sigma)$ 。VAE 采用图 2 所示的 LSTM 作为神经网络的基本单元, VAE 的神经网络结构如图 4 所示。

2.3 轨迹异常检测

正常轨迹通常具有很好的连续性、平滑性^[21], 经 VAE 重构的轨迹与原始轨迹应保持较好的一致性; 而异常的轨迹通常不是平滑的, 通常表现在速度发生明显变化或者方向偏离正常的运动模式, 经过 VAE 重构的轨迹无法较好地拟合原始轨迹。因此通过对原始轨迹和重构轨迹之间的空间距离进行对比, 如果距离大于一定的阈值 ϵ 则被认为是异常, 反之则是正常。轨迹段 $L_i=s_i e_i$ 和 $L_j=s_j e_j$ 的相似度从垂直距离、平行距离和角度距离综合度量:

$$dist(L_i, L_j) = \alpha_1 d_\perp + \alpha_2 d_{//} + \alpha_3 d_\theta \quad (16)$$

式中: $\alpha_1, \alpha_2, \alpha_3$ 分别为垂直距离、平行距离和角度距离的权重 $\alpha_1+\alpha_2+\alpha_3=1$ 。本文为了方便将权重设置为相等。

变分自编码器的输入和输出形式均是 $X^n = (x_0^n, x_1^n, \dots, x_i^n, \dots, x_T^n)$, 它由轨迹点组成轨迹序列。为了能够找到异常的轨迹段, 需要把输入和输出的轨迹点序列变成轨迹段组成的序列, 即 $L^n = (l_0^n, l_1^n, \dots, l_{i-1}^n, \dots, l_{T-1}^n)$, 其中 $l_{i-1}^n = (x_{i-1}^n, x_i^n)$ 。在轨迹异常检测阶段通过计算输入和输出的每一个轨迹段的空间距离来判定该轨迹段是否异常。基于 VAE 的轨迹异常检测方法如图 5 所示。

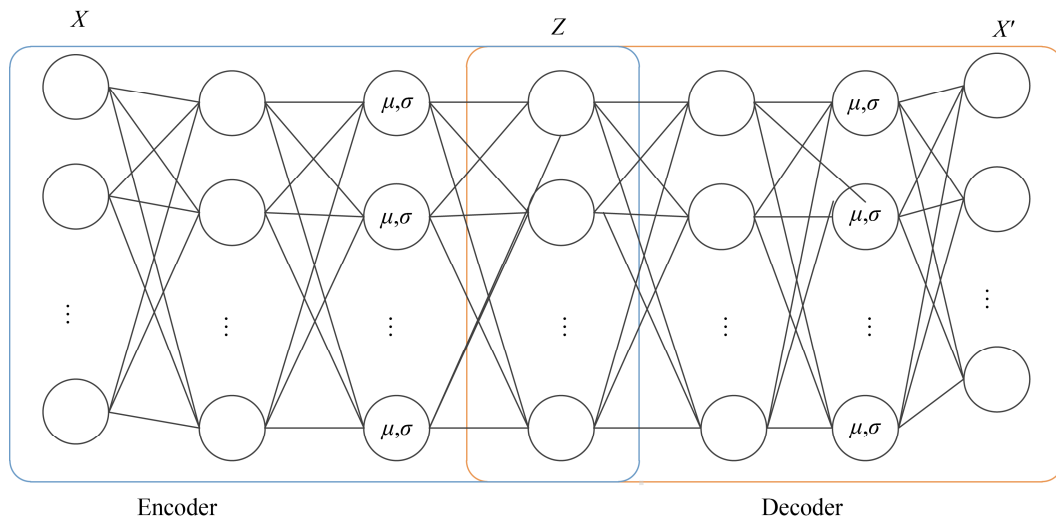


图 4 VAE 的神经网络结构
Fig. 4 Neural network architecture of VAE

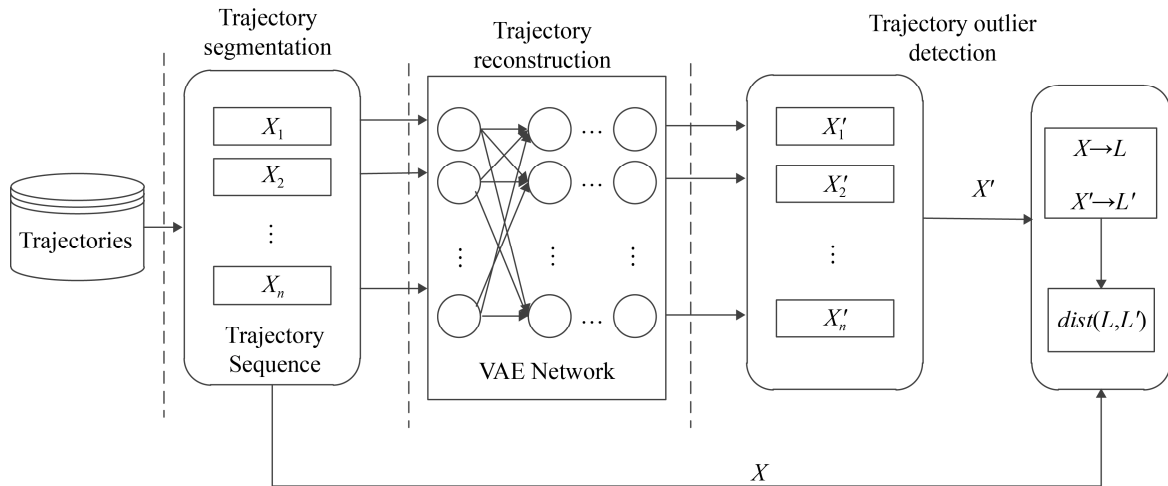


图 5 基于 VAE 的轨迹异常检测方法流程图

Fig. 5 Flow chart of trajectory outlier detection method based on VAE

3 实验结果与讨论

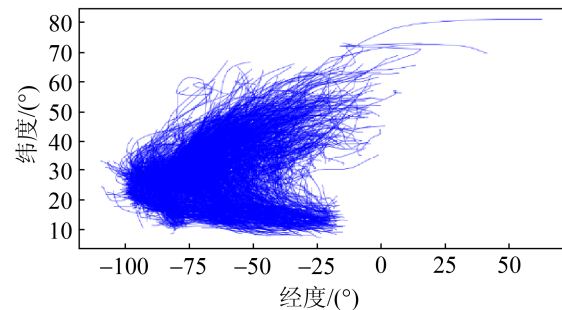
3.1 数据集和实验环境描述

本文采用美国国家飓风中心发布的 1851—2018 年的大西洋飓风数据进行仿真实验(www.nhc.noaa.gov/data), 该数据集的信息主要包括飓风经纬度、最大持续风速、中心最低气压、采样时间等信息, 主要利用飓风经纬度坐标进行实验。1851—2018 年的大西洋飓风数据总共包含 1 830 条飓风轨迹, 51 443 个采样点。实验选取 1851—2015 年这 165 年间的飓风轨迹作为训练集来训练 VAE 轨迹重构模型, 2016—2018 年的飓风轨迹作为验证集进行异常轨迹检测。1851—2015 年大西洋飓风轨迹和 2016—2018 年大西洋飓风轨迹如图 6 所示(东经用正值表示, 西经用负值表示)。实验硬件平台为 CPU Intel(R) Core(TM) i7-9850, 16G 内存, GPU QUADRO RTX3000, 操作系统为 Windows10, 实现语言为 python。

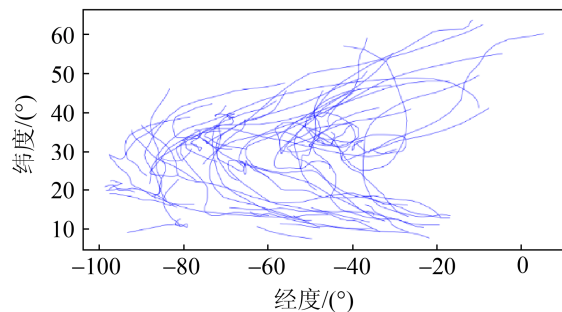
3.2 2016—2018 年飓风轨迹异常检测

本实验选取 1851—2015 年的飓风轨迹作为历史数据训练轨迹重构模型。首先采用滑动窗口将训练的飓风轨迹划分成相同长度的子轨迹序列, 窗口大小设置为 $w=5$, 时间步设置为 $s=1$ 。VAE 的网络结构为 $256 \times 64 \times 256$, 训练的轮次设置为 $epochs=$

20。将 2016—2018 年的飓风轨迹作为测试集进行异常检测, 采用滑动窗口将测试轨迹进行划分, 窗口大小设置为 $w=5$, 时间步设置为 $s=5$, 距离阈值设置为 $\epsilon=2.5$ 。基于变分自编码器的轨迹异常检测方法对 2016—2018 年的飓风轨迹异常检测结果如图 7 所示。



(a) 1851—2015 年



(b) 2016—2018 年

图 6 原始飓风轨迹图

Fig. 6 Original hurricane trajectory

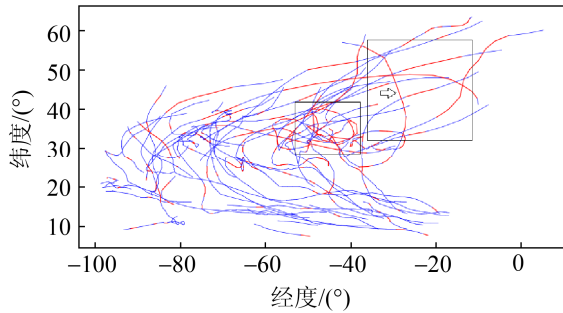


图 7 $\varepsilon=2.5$, 2016—2018 年的飓风轨迹异常检测结果图
Fig. 7 When $\varepsilon=2.5$, outlier detection results of hurricane trajectory 2016—2018

图 7 中红色线段代表检测出的异常轨迹段。此时检测出的异常轨迹段有 330 条, 占 2016—2018 年飓风总轨迹段的 19.4%。从图 7 中可以看出检测出来的异常轨迹大部分是在空间位置有些异常的飓风轨迹。从图 7 中箭头所指出的飓风轨迹可以明显看出轨迹路线和其他的轨迹有所不同, 具有明显的突变。尤其是图 7 中左边方框所圈出的区域, 这些飓风轨迹的突变情况更为明显。从图 6(a) 可以看出飓风轨迹一般出现的区域在 $25^{\circ}\sim 100^{\circ}\text{W}$, $10^{\circ}\sim 50^{\circ}\text{N}$ 左右, 图 7 中右边方框圈出的异常轨迹,

基本上处于飓风在历史上很少出现的区域。因为 VAE 是对历史知识的学习, 所以在历史上出现很少的飓风轨迹 VAE 不能完美地实现轨迹重构, 这些区域的轨迹被识别成了异常的轨迹。

本实验把距离阈值分别设置为 $\varepsilon=1.5, 1.8, 2.1, 2.4, 2.7, 3.0$, 2016—2018 年飓风轨迹的异常检测结果如图 8 所示。

在图 8 中, $\varepsilon=1.5, 1.8, 2.1, 2.4, 2.7, 3.0$ 时检测出的异常轨迹的数量分别为: 891, 687, 503, 374, 274, 189。异常轨迹的数量会随着阈值范围的扩大而减少。虽然阈值会对轨迹异常检测产生影响, 但是在合适的阈值范围内可以看出明显异常的轨迹可以被检测出来的。随着距离阈值的增加, 一些比较离散的异常轨迹段会逐渐消失, 但是明显异常的轨迹段则仍然会保留下来。

当采用不同的训练轮次时, 会对轨迹重构模型产生影响。训练轮次设置为 $epochs=20, 30, 40, 50$, 距离阈值设置为 $\varepsilon=3.0$ 时, 2016—2018 年飓风轨迹的异常检测结果如图 9 所示。

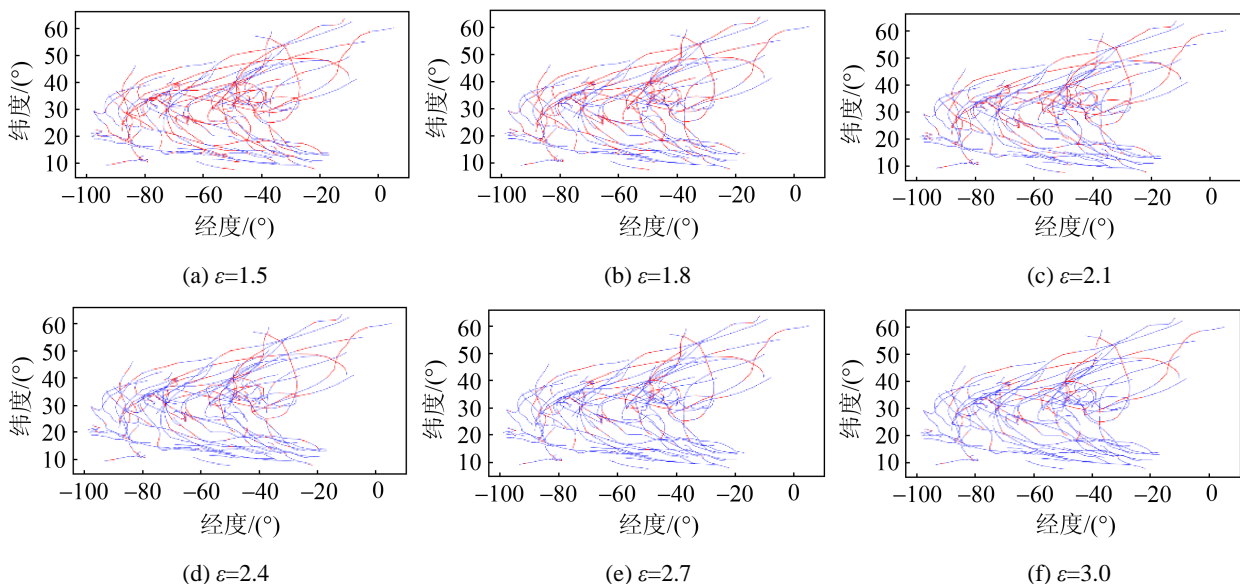


图 8 不同 ε 时, 2016—2018 年的飓风轨迹异常检测结果图
Fig. 8 Outlier detection results of hurricane trajectory from 2016—2018 with different ε

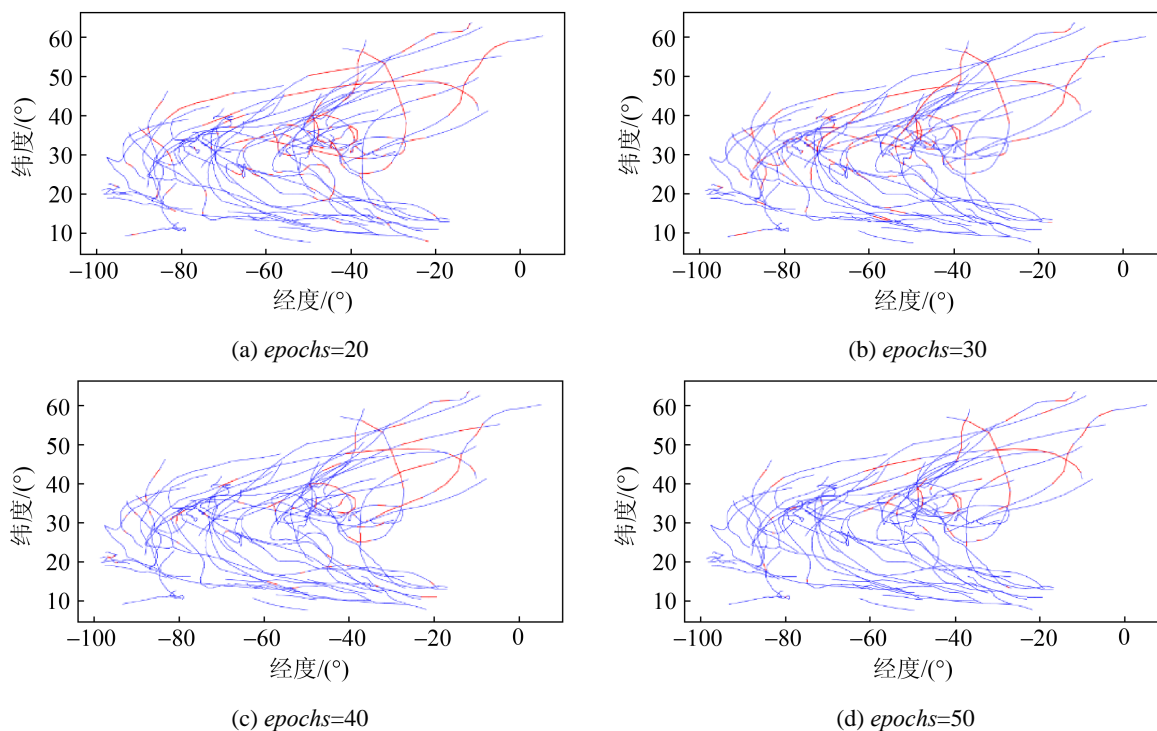


图 9 2016—2018 年飓风轨迹的异常检测结果图
Fig. 9 Outlier detection results of hurricane trajectory from 2016—2018

图 9(a)检测出的异常的轨迹段数量为 189; 图 9(b)检测出的异常轨迹段数量为 101; 图 9(c)检测出的异常轨迹段数量为 89; 图 9(d)检测出的异常轨迹段数量为 65, 异常的轨迹段会随着训练轮次的增加而减少。不同训练轮次的轨迹重构模型对轨迹重构的程度是不同的。训练的次数越多, 重构出的轨迹与原始轨迹就会越相似, 则检测出的异常轨迹段就会减少。虽然随着训练次数的增多, 一些细小的、异常程度不明显的轨迹段会逐渐消失, 但是异常程度高的轨迹仍然会被检测出来。因为这些异常程度明显的轨迹出现的次数少, 突变性大, 轨迹重构模型多轮次的学习也无法完美地拟合这些轨迹。

3.3 单个飓风轨迹异常检测

大部分的轨迹异常检测方法需要大量的数据, 通过当前轨迹与其邻域的轨迹进行对比找到异常的轨迹。这些方法使得每次异常检测都需要众多轨迹同时作为输入来进行大量的计算, 因此时间开销比较大, 无法实现单个轨迹的异常检测。本文提出

的 VAEOD 方法只需要单个的轨迹作为输入便可以完成异常检测。选取 2016 年飓风“MATTHEW”、2017 年飓风“HARVEY”和“MARIA”, 2018 年飓风“FLORENCE”这 4 条飓风轨迹进行单个轨迹异常检测实验。1851—2015 年的飓风轨迹作为训练集训练 VAE 轨迹重构模型。首先采用滑动窗口将训练的飓风轨迹划分成相同长度的序列片段, 窗口大小设置为 $w=5$, 时间步设置为 $s=1$ 。VAE 的神经网络结构为 $256 \times 64 \times 256$ 。训练的轮次设置为 $epochs=20$, 距离阈值设置为 $\epsilon=2.5$ 。单个飓风轨迹以窗口大小为 $w=5$ 逐步进行异常检测, 异常检测结果如图 10 所示。

MATTHEW 检测出 2 个轨迹变化比较大的位置; HARVEY 检测出 1 个轨迹变化比较大的位置; MARIA 检测出 1 个变化比较大的位置; FLORENCE 检测出 2 个轨迹变化比较大的位置。本实验选取的这几条飓风影响力比较大, 每一条飓风均产生了不同程度的伤亡, 因此如果能够有效地检测出其异常的部分, 则可以尽最大可能地避免人

员伤亡和经济损失。MARIA 开始是从东南向西北的方向运动, 然后变成了西北向东南的运动的趋势, 检测出这种做运动趋势的改变则可以提前对飓风的变化进行预防。

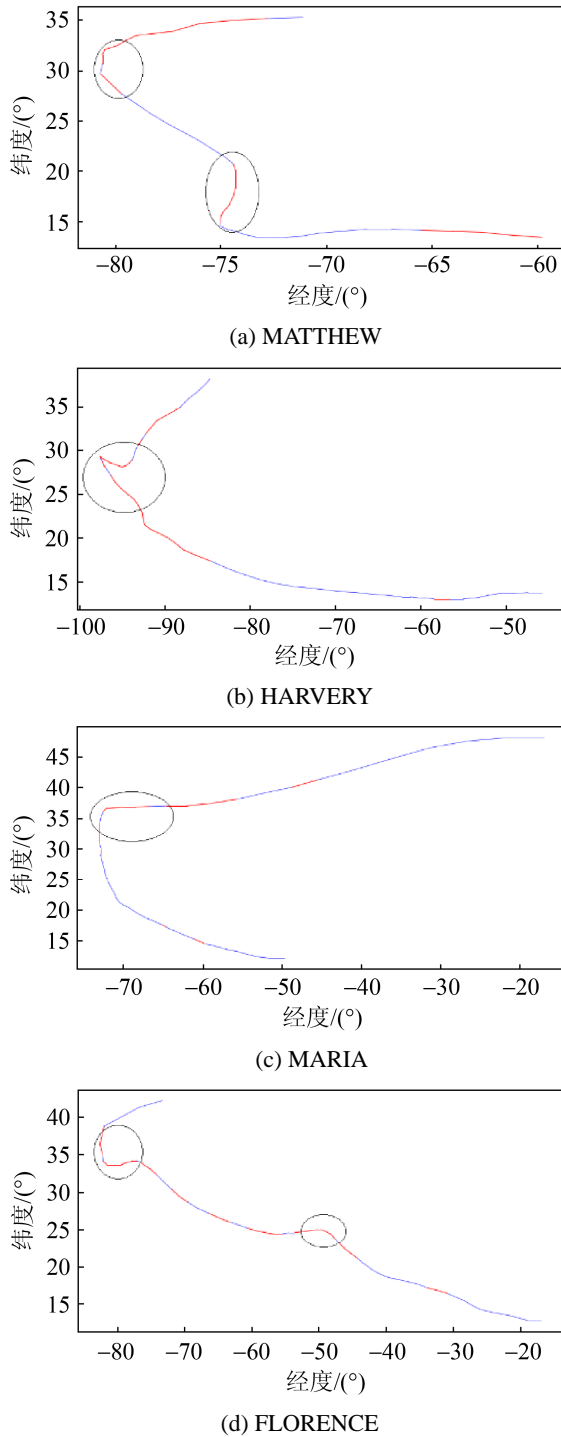


图 10 单个飓风轨迹异常检测结果图
Fig. 10 Outlier detection results of a single hurricane trajectory

3.4 与轨迹异常值检测算法对比实验

本实验将 VAEOD 方法和经典的轨迹异常值检测算法 (Trajectory Outlier Detection Algorithm, TRAOD)^[8] 方法进行对比。TRAOD 算法是一种经典的轨迹异常检测算法, 首先采用分段方法将每条轨迹分为若干轨迹线段, 然后引入调整系数的概念, 将基于距离的方法和基于密度的方法相结合实现轨迹的异常检测。本文的 VAEOD 方法将距离参数设置为 $\epsilon=3.0$, 通过 1851—1999 年的飓风轨迹作为训练集进行训练轨迹重构模型, 然后对 2000—2006 年飓风轨迹进行异常检测, 其结果如图 11(a) 所示。TRAOD 方法对 2000—2006 年轨迹异常检测结果如图 11(b) 所示。

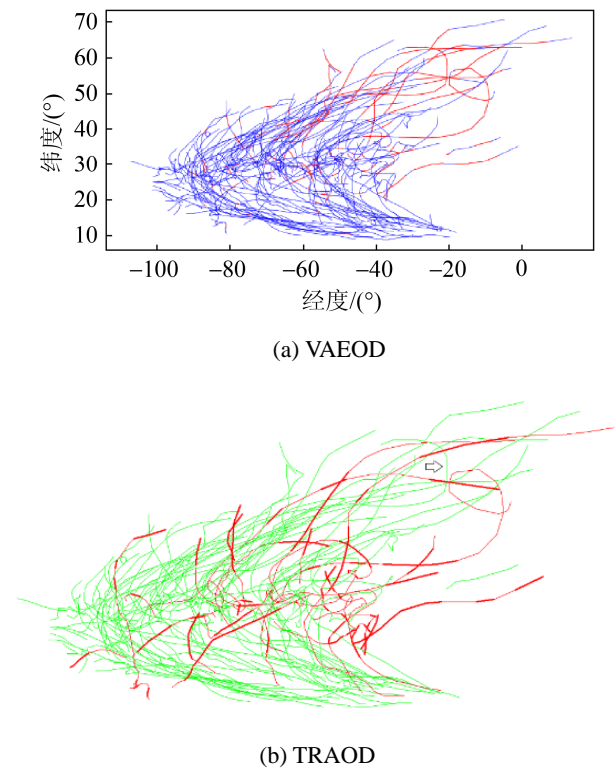


图 11 异常检测方法结果对比
Fig. 11 Results of VAEOD and TRAOD outlier detection methods are compared

VAEOD 方法检测出的异常轨迹段为 255 个, TRAOD 方法检测出的异常轨迹段 244 个, 2 种方法检测的异常轨迹段的数量相似。VAEOD 方法和 TRAOD 方法在检测结果上具有相似之处, 同时也

存在一定的差异。如图 11(b)中箭头指出的轨迹，TRAOD 对于明显异常的轨迹没有检测出来，但是 VAEOD 却检测出来了。同时 VAEOD 自适应调节能力强，人工干预少。TRAOD 的异常检测方法参数较多，有 3 种参数需要人为设置，距离阈值、邻域数量、异常分数，这些参数的设置对最终的检测结果是有影响的。因为 TRAOD 方法应用中人为设置的参数比较多，所以该方法的自适应性比较差。VAEOD 方法只有距离阈值需要根据经验预估，轨迹重构模型作为一种神经网络模型可以自动训练，因此具有较好的自适应性。TRAOD 方法只能通过大量的轨迹及其邻域的距离对比找到异常的轨迹段，因此计算比较复杂同时无法实现单个轨迹的异常检测。VAEOD 轨迹异常检测方法只要训练好轨迹重构模型，便可以实现单个轨迹的异常检测，相较于 TRAOD 方法，VAEOD 的实用性更强。

4 结论

本文的主要贡献有以下几方面：

(1) 变分自编码器需要等长的序列作为输入，因此基于滑动窗口将长度不等的轨迹序列划分成长度相等的子轨迹序列作为变分编码器的输入。

(2) 轨迹数据是一种无异常标签的序列数据，于是以变分自编码器为基础进行无监督学习，训练一个轨迹重构模型。

(3) 为了检测异常的轨迹段，从空间距离的 3 个方面即平行距离、垂直距离和角度距离来综合计算轨迹重构模型重构的轨迹和原始轨迹之间的距离，如果空间距离大于给定的阈值则定义为异常。

(4) 利用 1851—2018 年真实的大西洋飓风数据进行仿真实验，发现 VAEOD 轨迹异常检测方法同经典的 TRAOD 方法相比检测的结果更合理，自适应更强同时更具有实用性。

本文所提出 VAEOD 轨迹异常检测方法仅从空间距离的角度衡量轨迹的异常性，未考虑时间状态变化的异常，因此在后续的研究当中将结合轨迹时间状态的变化完善轨迹的异常检测。

参考文献：

- [1] Meuel T, Prado G, Seychelles F, et al. Hurricane Track Forecast Cones from Fluctuations[J]. *Scientific Reports* (S2045-2322), 2012, 2: No.446.
- [2] Mann M E, Woodruff J D, Donnelly J P, et al. Atlantic Hurricanes and Climate Over the Past 1,500 Years[J]. *Nature* (S0028-0836), 2009, 460(7257): 880-883.
- [3] Zheng Y. Trajectory Data Mining: An Overview[J]. *ACM Transactions on Intelligent Systems and Technology* (S2157-6904), 2015, 6(3): 1-41.
- [4] Tang J. Conflict Detection and Resolution for Civil Aviation: A Literature Survey[J]. *IEEE Aerospace and Electronic Systems Magazine* (S0018-9251), 2019, 34(10): 20-35.
- [5] Meng F, Yuan G, Lv S, et al. An Overview on Trajectory Outlier Detection[J]. *Artificial Intelligence Review* (S0269-2821), 2019, 52(4): 2473-2456.
- [6] Tang J, Piera M A, Guasch T. Coloured Petri Net-based Traffic Collision Avoidance System Encounter Model for the Analysis of Potential Induced Collisions[J]. *Transportation Research Part C: Emerging Technologies* (S0968-090X), 2016, 67: 357-377.
- [7] Tang J, Zhu F, Piera M A. A Causal Encounter Model of Traffic Collision Avoidance System Operations for Safety Assessment and Advisory Optimization in High-density Airspace[J]. *Transportation Research Part C: Emerging Technologies* (S0968-090X), 2018, 96: 347-365.
- [8] Lee J G, Han J, Li X. Trajectory outlier detection: A partition-and-detect framework[C]//2008 IEEE 24th International Conference on Data Engineering. Cancun, Mexico: IEEE, 2008: 140-149.
- [9] Lee J G, Han J W, Whang K Y. Trajectory Clustering: A Partition-and-group Framework[C]//2007 ACM SIGMOD International Conference on Management of Data. Beijing: ACM, 2007: 593-604.
- [10] Liu Z, Pi D, Jiang J. Density-based Trajectory Outlier Detection Algorithm[J]. *Journal of Systems Engineering & Electronics* (S1004-4132), 2013, 24(2): 335-340.
- [11] Ge Y, Xiong H, Zhou Z H, et al. Top-Eye: Top-k Evolving Trajectory Outlier Detection[C]//19th ACM Conference on Information and Knowledge Management. Toronto: ACM, 2010: 1733-1736.
- [12] Pang L X, Chawla S, Liu W, et al. On Mining Anomalous Patterns in Road Traffic Streams[C]//Advanced Data Mining and Applications - 7th International Conference. Beijing: Springer-Verlag, 2011: 237-251.
- [13] Pang L X, Chawla S, Liu W, et al. On Detection of

- Emerging Anomalous Traffic Patterns Using GPS Data[J]. *Data & Knowledge Engineering (S0169-023X)*, 2013, 87(9): 357-373.
- [14] Li X, Han J, Kim S, et al. ROAM: Rule and Motif-based Anomaly Detection in Massive Moving Object Data Sets[C]//7th SIAM International Conference on Data Mining. New Orleans: SIAM, 2007: 296-307.
- [15] Piciarelli C, Foresti G L. Anomalous Trajectory Detection Using Support Vector Machines[C]//2007 IEEE Conference on Advanced Video and Signal Based Surveillance. London: IEEE, 2007: 153-158.
- [16] Alahi A, Goel K, Ramanathan V, et al. Social LSTM: Human Trajectory Prediction in Crowded Spaces[C]//IEEE conference on computer vision and pattern recognition. Las Vegas: IEEE, 2016: 961-971.
- [17] Hinton G E. Reducing the Dimensionality of Data with Neural Networks[J]. *Science (S0036-8075)*, 2006, 313(5786): 504-507.
- [18] Kingma D P, Welling M. Auto-Encoding Variational Bayes[EB/OL]. (2013-12-20) [2020-05-24]. <https://arxiv.org/abs/1312.6114>.
- [19] Rezende D J, Mohamed S, Wierstra D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models[EB/OL]. (2014-02-16) [2020-05-24]. <https://arxiv.org/abs/1401.4082>.
- [20] 方华强, 颜寒祺, 陈波, 等. 基于自编码网络的移动轨迹异常检测[J]. *地理信息世界*, 2019, 26(5): 41-44.
- Fang Huaqiang, Yan Hanqi, Chen Bo, et al. Anomaly Detection in Mobile Trajectory Using Auto-encoder Network[J]. *Geomatics World*, 2019, 26(5): 41-44.
- [21] 霍纬纲, 王慧芳. 基于自编码器和隐马尔可夫模型的时间序列异常检测方法[J]. *计算机应用*, 2020, 40(5):1329-1334.
- Huo Weigang, Wang Huifang. Time Series Anomaly Detection Method Based on Autoencoder and HMM[J]. *Journal of Computer Applications*, 2020, 40(5): 1329-1334.
- [22] Vincent P, Larochelle H, Bengio Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders [C]//25th International Conference on Machine Learning. Helsinki, Finland: ACM, 2008: 1096-1103.