基于Bi-LSTM模型的轨迹异常点检测算法

韩昭蓉[①②③] 黄廷磊[*②③] 任文娟[②③] 许光銮[②③]

[①](中国科学院大学 北京 100049)

[②](中国科学院电子学研究所 北京 100190)

[③](中国科学院空间信息处理与应用系统技术重点实验室 北京 100190)

摘 要：定位技术的飞速发展催生了时空轨迹大数据，轨迹数据中往往存在着明显偏离轨迹的异常点。检测出轨迹中的异常点对提高数据质量和后续轨迹数据挖掘精度至关重要。该文提出了一种基于双向长短时记忆网络(Bidirectional Long Short-Term Memory, Bi-LSTM)模型的轨迹异常点检测算法。首先对每个轨迹点提取一个6维的运动特征向量，然后构建了一个Bi-LSTM模型，模型输入为一定序列长度的轨迹数据特征向量，输出为轨迹点的类型结果。同时，算法采用了欠采样和过采样的组合方法缓解类别不平衡对检测性能的影响。融合了长短时记忆网络单元和双向网络，Bi-LSTM模型能够自动学习正常点和邻近异常点在运动特征上的差异。基于真实船舶轨迹标注数据的实验结果表明，该文算法的检测性能显著优于恒定速度阈值法、不考虑数据时序性的经典机器学习分类算法和卷积神经网络模型，尤其是召回率达到了0.902，验证了该文算法的有效性。

关键词：轨迹数据；异常检测；特征提取；双向长短时记忆网络

中图分类号：TP391 **文献标识码**：A **文章编号**：2095-283X(2019)01-0036-08

DOI: 10.12000/JR18039

引用格式：韩昭蓉, 黄廷磊, 任文娟, 等. 基于Bi-LSTM模型的轨迹异常点检测算法[J]. 雷达学报, 2019, 8(1): 36–43. doi: 10.12000/JR18039.

Reference format: HAN Zhaorong, HUANG Tinglei, REN Wenjuan, *et al*. Trajectory outlier detection algorithm based on Bi-LSTM model[J]. *Journal of Radars*, 2019, 8(1): 36–43. doi: 10.12000/JR18039.

Trajectory Outlier Detection Algorithm Based on Bi-LSTM Model

HAN Zhaorong[①②③] HUANG Tinglei[*②③] REN Wenjuan[②③] XU Guangluan[②③]

[①](*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

[②](*Institute of Electronics*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

[③](*Key Laboratory of Technology in Geo-Spatial Information Processing and Application System*, *Chinese Academy of Sciences*, *Beijing* 100190, *China*)

Abstract: The rapid advances in positioning technology have created huge spatio-temporal trajectory data, and there are always obvious aberrant outliers in trajectory data. Detecting outliers in the trajectory is critical to improving data quality and the accuracy of subsequent trajectory data mining tasks. In this paper, we propose a trajectory outlier detection algorithm based on a Bidirectional Long Short-Term Memory (Bi-LSTM) model. First, a six-dimensional motion feature vector is extracted for each trajectory point, and then we construct a Bi-LSTM model. The model input is the trajectory data feature vector of a certain sequence length, and its output is the class type of the current track point. In addition, a combination method of undersampling and oversampling is applied to mitigate the effect of data distribution imbalance on detection performance. The Bi-LSTM model can automatically learn the difference between the normal points and adjacent abnormal points in the motion characteristics by combining the LSTM unit and the bidirectional network. Experimental results based on a real ship trajectory annotation data show that the detection performance of our proposed algorithm

收稿日期：2018-05-14；改回日期：2018-05-30；网络出版：2018-07-09

*通信作者： 黄廷磊 tlhuang@mail.ie.ac.cn *Corresponding Author: HUANG Tinglei, tlhuang@mail.ie.ac.cn

基金项目：国家自然科学基金(61725105, 61331017)

Foundation Items: The National Natural Science Foundation of China (61725105, 61331017)

significantly exceeds those of the constant velocity threshold algorithm, non-sequential classical machine learning classification algorithms, and convolutional neural network model. Especially, the recall value of the proposed algorithm reaches 0.902, which verifies its effectiveness.

Key words: Trajectory data; Outlier detection; Feature extraction; Bidirectional Long Short-Term Memory (Bi-LSTM) networks

1　引言

随着定位技术、无线通信技术和存储计算能力的飞速发展，大量移动目标的轨迹数据呈爆炸式的增长，包括人类活动轨迹、交通轨迹数据和动物迁徙数据等。轨迹大数据中蕴藏着丰富的有价值的目标活动信息，因此国内外学者对轨迹数据挖掘任务进行了大量的探索和研究，如轨迹聚类[1]、轨迹关联分析[2,3]、目标运动模式识别[4]、路径规划[5]和异常模式检测[6]等。分析挖掘时空轨迹数据在科研领域和人类生活应用方面都具有重大意义。但是，轨迹数据在现实环境下从来都不是完全准确的[7]，数据中存在许多与其邻近大部分轨迹点在运动特征上有显著差异的不合理的采样点，这些点称为轨迹数据中的异常点。异常点的存在严重降低了轨迹数据的质量，同时会引起后续轨迹知识发现结果的不准确甚至错误。因此，轨迹异常点检测是轨迹数据挖掘前至关重要的一步。

本文以实验室现有的船舶轨迹数据为研究对象，数据由多种不同数据源侦测获得，各个数据源之间定位精度差异较大，由于不同数据源定位误差的差异、环境干扰、人为操作失误或是目标刻意伪装欺骗，船舶轨迹数据中存在着大量不符合目标运动规律的异常点。Hawkins在1980年对异常点提出的定义为：异常点是指在数据集中显著偏离其它绝大部分数据的那些数据对象，以至于引起人们怀疑它们是由完全不同的机制产生的[8]。传统异常点检测算法主要分为基于统计的方法、基于距离的方法、基于密度的方法和基于分类的方法等[9]。在基于分类的检测方法中，首先训练一个可以区分正常数据和异常点的分类模型，然后用预先训练好的模型来判断新的观测点是否异常。方法学到的模型能够更加接近数据的实际规律，具有良好的可扩展性。

在轨迹的异常点检测问题上，由于轨迹数据是基于时间和空间的位置序列，相邻点有着上下文关系，这使得传统的异常点检测方法不能直接用于检测轨迹序列数据中的异常点。Alvares等人[10]、Chen等人[11]和Zheng等人[12]均采用恒定速度阈值法来检测出轨迹数据中的异常点，依次选取每个轨迹点与其前一个点计算即时速度，速度超出设定的恒定阈值即判定为异常点，去除检测出的异常点后再进行轨迹数据挖掘任务。Chen等人[13]提出了一种基于模型的GPS轨迹清理算法，采用三次平滑样条和时间序列方法分别对轨迹的趋势和残差进行自适应建模，可有效检测出低精度GPS轨迹上的异常点。Hu[14]结合规则轨迹集对监控区域内的船舶实时轨迹数据进行异常检测，船舶航速超出区域内正常最高航速则为异常的船舶轨迹点，进而识别异常船舶并进行预警。Wu等人[15]通过分析大量AIS (Automatic Identification System，船舶自动识别)轨迹数据，归纳出几种类型的轨迹异常点，并设计了对应的规则来实现检测，包括有：根据两点间的经纬度差值所对应的逻辑距离来判断；不在正常的AIS通信范围内为异常点；两点间实际距离与理论距离(由航速及时间计算出)的差值超过一定阈值则为异常点，最后采用拟合插值法对轨迹进行修复。上述方法均根据特定的轨迹数据人工设置了相应的参数和规则来完成检测，其存在的主要问题是：(1)参数设置过程需要大量的人工分析和测试调整，较为繁琐；(2)由于移动目标会有不同的运动状态，人工难以设定出契合数据的准确的参数阈值或模型，方法容易出现漏检和错检的情况；(3)方法通常只适用于特定类型的轨迹数据，扩展性不强。总地来说，这些方法的检测性能高度依赖于参数是否准确，不能自动学习轨迹异常点和正常数据的差异，对复杂类型数据适用性较差。

目前，深度学习方法已被证实可以直接从大数据中自动学习特征，在异常检测任务中也具有较大的潜力。Bessa等人[16]提出了一种交互可视化和检测异常轨迹段的工具RioBusData，工具基于一个多层的卷积神经网络(Convolutional Neural Network, CNN)，在公交车轨迹数据上的实验结果表明，方法可以有效检测出异常线路的公交车、时间异常轨迹段和空间异常轨迹段。Fernando等人[17]提出了一种新颖的基于长短时记忆网络(Long Short-Term Memory, LSTM)的编码器-解码器框架，引入了注意力机制，能够根据行人和其周围邻居的历史轨迹预测出感兴趣行人的未来位置，方法在两个具有挑战性的数据集上均表现出了出色的性能，同时通过对比预测路径和行人的真实路径可以检测出异常的行为。

在实际问题的驱动和上述深度学习应用的启发下，考虑到长短时记忆网络在序列处理任务中优异的特征学习能力，本文提出了一种基于双向长短时记忆网络(Bidirectional LSTM, Bi-LSTM)的轨迹异常点检测算法。本文设计了一个Bi-LSTM模型，对每个轨迹点构建一个6维的运动特征向量，选取一段时间的轨迹数据特征向量作为模型的输入，模型输出为轨迹点的分类结果(1为异常点，0为正常点)。同时，算法采用了欠采样和过采样的组合方法缓解类别不平衡对检测性能的影响。最后通过实验验证了基于Bi-LSTM模型的轨迹异常点检测算法的有效性。

2 算法描述

本文算法的核心思想是将轨迹异常点检测问题转化为有监督的分类问题，然后构建能够有效处理时序数据的长短时记忆网络模型予以解决。以下本文将首先介绍轨迹点的特征向量提取过程，然后对长短时记忆网络的相关理论知识进行描述，最后介绍本文模型的网络结构和整个算法的流程。

2.1 特征提取

特征提取是指应用专业领域知识从原始数据中找出一些具有物理意义的特征，是机器学习算法能够有效工作的重要过程。好的特征可以极大提高学习系统的性能。轨迹数据是由一系列随时间变化的时空数据点组成，即$\mathrm{TR}: P_1 \to P_2 \to \cdots \to P_{\mathrm{len}}$，这里第$i$个轨迹点可以表示为$P_i=(\mathrm{lon}_i, \mathrm{lat}_i, t_i)$，$\mathrm{lon}_i$, lat_i为轨迹点的经度和纬度值，t_i为该点的时间戳信息。对每个轨迹点，本文提取了一个6维的运动特征向量，包括航速(speed)、加速度(acceleration)、航向(course)、转角(turning angle)、转角率(turning rate)和曲率(sinuosity)。这些参数特征均为移动目标运动特性的重要表征。

以图1为例，本文对每个特征进行介绍。航速是移动目标位置变化的速率，用于表示运动快慢的程度。加速度是目标航速对于时间的变化率，描述航速变化的快慢。轨迹中异常点通常比其邻近点有更大的航速或加速度值。对轨迹点P_i，其航速和加速度计算公式分别如下：

$$v_i=\frac{\mathrm{dist}(P_i, P_{i-1})}{t_i-t_{i-1}} \tag{1}$$

$$a_i=\frac{v_i-v_{i-1}}{t_i-t_{i-1}} \tag{2}$$

其中，$\mathrm{dist}(P_i, P_{i-1})$表示点$P_i$和其前一个点$P_{i-1}$间的距离。

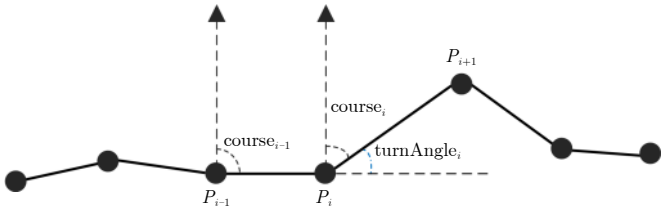


图 1 轨迹示意图

Fig. 1 A diagram of trajectory segment

航向定义为轨迹中连续点之间的移动朝向，本文取当前轨迹点与后一时刻轨迹点的连线与正北方向的夹角来表示，而转角表示两个连续轨迹点航向间的变化。与周围轨迹点相比，那些航向值和转角值明显不同的点为异常点的可能性更大。转角率表示连续轨迹点转角变化量与时间的比值。航向、转角、转角率的计算公式分别如下所示：

$$\begin{cases}\mathrm{course}_i=\arctan(X, Y)\\ X=\cos(\mathrm{lat}_i)\cdot\sin(\mathrm{lon}_i-\mathrm{lon}_{i-1})\\ Y=\cos(\mathrm{lat}_{i-1})\cdot\sin(\mathrm{lat}_i)-\sin(\mathrm{lat}_{i-1})\\ \qquad\cdot\cos(\mathrm{lat}_i)\cdot\cos(\mathrm{lon}_i-\mathrm{lon}_{i-1})\end{cases} \tag{3}$$

$$\mathrm{turnAngle}_i=\mathrm{course}_{i-1}-\mathrm{course}_i \tag{4}$$

$$\omega_i=\frac{\mathrm{turnAngle}_i-\mathrm{turnAngle}_{i-1}}{t_i-t_{i-1}} \tag{5}$$

曲率定义为两点之间移动距离与直线距离的比值。如式(6)所示，本文计算轨迹点的曲率为该点与前后两个时刻轨迹点的距离和与前后时刻轨迹点直接距离的比值。从轨迹曲线可以看出来，正常点的曲率值要远小于异常点的曲率值。

$$s_i=\frac{\mathrm{dist}(P_{i-1}, P_i)+\mathrm{dist}(P_i, P_{i+1})}{\mathrm{dist}(P_{i-1}, P_{i+1})} \tag{6}$$

本文提取的6个特征值都能真实反映目标在时空上的运动状态，可为轨迹点的分类提供有用信息，从而有助于提高检测精度。

2.2 长短时记忆网络

长短时记忆网络(Long Short-Term Memory, LSTM)[18]是循环神经网络(Recurrent Neural Network, RNN)的一种特殊形式，通过引入记忆单元

和门限机制的巧妙构思，能够学习长期依赖关系，缓解RNN存在的梯度消失和梯度爆炸问题，已广泛应用在序列处理任务中。

如图2所示，LSTM单元主要由4个部分组成：记忆单元(Memory cell)，输入门(Input gate)，输出门(Output gate)及遗忘门(Forget gate)。记忆单元之间彼此循环连接，3个非线性门控单元可以调节流入和流出记忆单元的信息。LSTM的前向计算公式如下所示：

$$\begin{cases} \boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \\ \boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \\ \boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \\ \boldsymbol{c}_t = \boldsymbol{f}_t \circ \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \\ \qquad \circ \tanh(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \\ \boldsymbol{h}_t = \boldsymbol{o}_t \circ \tanh(\boldsymbol{c}_t) \end{cases} \tag{7}$$

其中，$\boldsymbol{x}_t$是当前时刻输入向量，$\boldsymbol{f}$, $\boldsymbol{i}$, $\boldsymbol{o}$分别为遗忘门、输入门、输出门的激活向量，$\boldsymbol{c}$为记忆单元向量，$\boldsymbol{h}$是LSTM单元的输出向量，$\boldsymbol{W}$, $\boldsymbol{b}$分别为权重矩阵和偏置向量，σ为激活函数，本文选用sigmod函数，符号“$\circ$”为哈达玛积(矩阵对应元素相乘)。

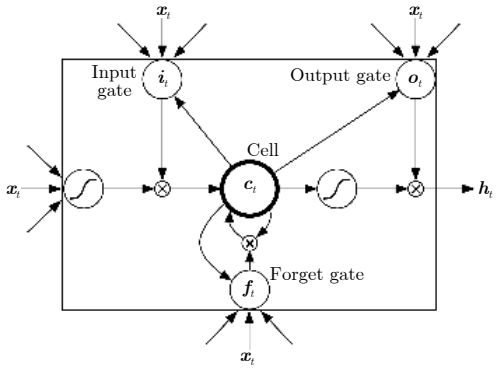


图 2 LSTM模块单元

Fig. 2　LSTM model unit

LSTM在预测当前时刻输出时，只利用前面时刻的历史序列信息，但往往输出同样取决于后续时刻的信息。为了充分利用上下文信息，Graves提出了Bi-LSTM模型[19]，Bi-LSTM网络结合时间上从序列起点开始移动的LSTM和另一时间上从序列末尾开始移动的LSTM，其输出单元由正向LSTM和反向LSTM的状态连接得到。融合了LSTM单元和双向网络，双向LSTM模型在语音识别、手写体识别、序列标注等学习任务中性能均有一定的提升。

2.3　模型构建

本文设计了一个双向LSTM (Bi-LSTM)模型，结构如图3所示。Bi-LSTM模型由两层Bi-LSTM网络和两层全连接层组成，用到了当前时刻和前后t时刻的上下文信息，通过正向、反向LSTM分别提取轨迹特征信息，适合离线检测轨迹数据中的异常点。

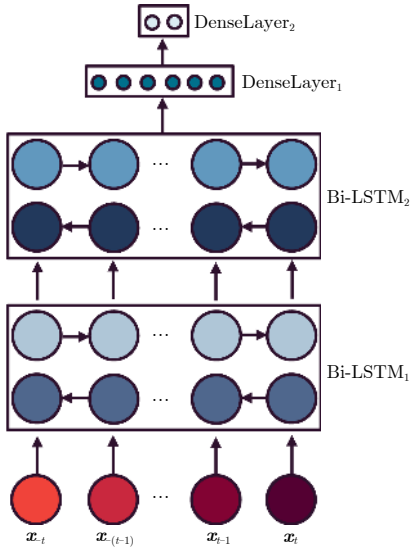


图 3 双向LSTM模型

Fig. 3　A bidirectional LSTM network

在特征构建过程后，每一个轨迹点均由一个6维向量来表示，即$\boldsymbol{x}_i = \{v_i, a_i, \text{course}_i, \text{turnAngle}_i, \omega_i, s_i\}$。模型结构图中，$\boldsymbol{x}_0$表示当前时刻的轨迹点输入信息，$\boldsymbol{x}_{-i}$和$\boldsymbol{x}_i$分别表示$\boldsymbol{x}_0$前后$i$时刻的轨迹点信息。模型的输入为一定序列长度的轨迹点向量，由Bi-LSTM自动提取序列间的特征，最后再通过两层全连接层对轨迹点进行分类。其中，异常点类标号为1(正类)，正常点类标号为0(负类)，Bi-LSTM模型序列长度为$2t+1$。由于充分考虑到一段时间内轨迹数据点的运动特征信息，Bi-LSTM模型在训练过程中可以自动学习到复杂轨迹序列中异常点和正常点的差异，模型一经训练好即可以用于轨迹异常点的检测，具有较高的实用性。为了提高模型的泛化能力，本文在Bi-LSTM层和第1层全连接层之间添加了dropout机制[20]，dropout通过部分连接来防止模型过拟合。

对每个轨迹点提取一定长度轨迹序列的运动特征后，算法采用训练好的Bi-LSTM模型来进行预测判断。算法流程如图4所示。

3　实验验证

3.1　数据集

本文采用了一个真实的船舶轨迹数据集来验证算法的有效性。船舶轨迹数据来源于实际项目，由多种数据源探测获得。本文随机抽取了船舶的300条轨迹数据，由多个有经验的判读员根据整条

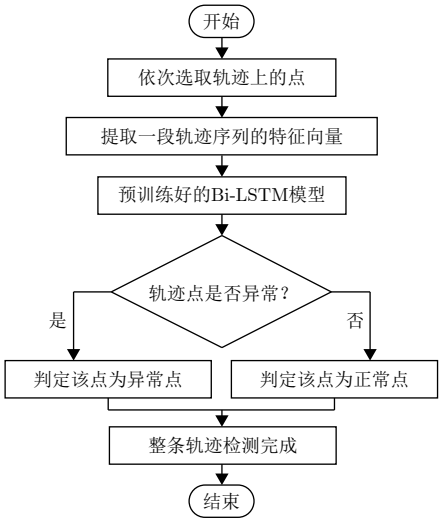
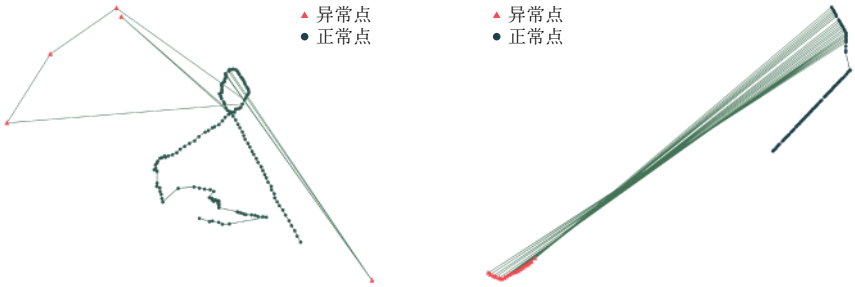


图 4 算法流程图

Fig. 4 The flowchart of our proposed algorithm

轨迹的统计分析和每个轨迹点相对于其邻近点的位置、航速和加速度信息对轨迹数据进行检验标注，数据集的标注具有较高的可靠度，由此，本文建立了一个船舶轨迹标记数据集，其中，异常点标记为1(正类)，正常数据标记为0(负类)。在具体的数据分析标注过程中，本文发现船舶轨迹数据中的异常点主要分为以下几种情形：(1)船舶运动平稳，异常点速度超出其周围点的速度但是没有超出目标的运动能力；(2)由于某些数据源定位精度较低或存在干扰情况，异常点严重偏离船舶航迹，其速度值超出目标最高航速；(3)目标轨迹点来回跳跃，航迹明显由两段不同的轨迹组成，这是由于在数据收集过程中错误地将两艘船的轨迹判识为同一个运动对象的。异常点通常为单个孤立的，同时也会有几个连续异常点的情况出现，轨迹异常点与其邻近点均有着较大的差异。图5为两段轨迹的标注结果，异常点用红色三角形表示。从图中可以看出，去除异常点的轨迹更为平滑，符合船舶的运动情况。

数据集中，正常点个数为60872，异常点个数为1456。负样本和正样本的比例约为42:1，可见存在数据类别分布不平衡的问题，这通常会影响学习系统的性能。本文将数据集随机划分为训练集(70%)和测试集(30%)。本文采用欠采样和过采样的组合方法SMOTE+ENN[21]来平衡训练数据的类别分布。SMOTE (Synthetic Minority Oversampling Technique)是一种过采样方法，其主要思想是通过在几个少数类样本间插值来形成新的少数类样本。但是，在插值过程中，SMOTE会产生噪声样本。这个问题可以通过使用欠采样方法ENN (Edited Nearest Neighbor)对插值结果进行清理来解决，任何与其k个最近邻居类别不同的样本都会被移除，从而产生一个类别平衡的训练集。本文最终在测试集上评估各个模型的检测性能。



(a) 轨迹1
(a) Track 1

(b) 轨迹2
(b) Track 2

图 5 两段轨迹的标记结果

Fig. 5 Tagging results for two track segments

3.2 评价指标

本文采用多种常用的机器学习评价指标[22]，包括精度(Accuracy)、准确率(Precision)、召回率(Recall)、F1值(F1-score)、ROC曲线(Receiver Operating Characteristic Curve)和AUC值(Area Under ROC Curve)。对于二分类问题，可将样例根据真实情况和模型预测类别的组合划分为真正例、假正例、真反例和假反例，分类结果的混淆矩阵如表1所示。

精度定义为正确预测样本数占总样本数的比率。准确率定义为正确预测的正样本数占总的预测为正样本数的比率，召回率则定义为正确预测的正样本数占实际正样本总数的比率，在异常检测应用

表 1 分类结果混淆矩阵

Tab. 1 Confusion matrix of classification results

真实情况	预测结果	
	异常点	正常点
异常点	真正例(TP)	假反例(FN)
正常点	假正例(FP)	真反例(TN)

中，相比于不将正常值错判为异常值，尽可能多地捕捉到所有的异常值更为重要，本文将更加关注召回率指标。F1值是准确率和召回率的调和均值，值较高时说明分类器性能好。这4种指标的公式分别如下所示：

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{8}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{9}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{10}$$

$$\text{F1-score}=\frac{2\cdot\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}} \tag{11}$$

ROC曲线表示真正例率(True Positive Rate, TPR)和假正例率(False Positive Rate, FPR)的关系。TPR为成功检测到的异常点的比率，FPR为错判为异常的正常点的比率，定义如式(12)所示。显然，异常检测算法应该具有较高的TPR值和较低的FPR值。当无法从ROC曲线直观比较出不同模型性能时，可以通过比较ROC曲线下的面积，即AUC值，AUC值越接近于1，表示算法的性能越好。而且，ROC曲线和AUC值对类别分布不平衡的数据不敏感[23]。

$$\left.\begin{aligned}\text{TPR}&=\frac{\text{TP}}{\text{TP}+\text{FN}}\\ \text{FPR}&=\frac{\text{FP}}{\text{FP}+\text{TN}}\end{aligned}\right\} \tag{12}$$

3.3 实验设置

本文采用Pytorch框架构建Bi-LSTM模型，在NVIDIA GTX 1080显卡上训练模型。每个轨迹点的输入向量维度为6, Bi-LSTM单元的隐藏层维度设置为3, dropout比例设置为0.5，两层全连接层节点数分别设置为6和2，最终模型输出为0(正常)或1(异常)。本文尝试采用单层和多层Bi-LSTM实现轨迹异常点检测，双层网络性能较优异，所以在实验中采用图3所示模型。

轨迹序列长度并不确定，本文测试验证了截取时间长度t 在1～25之间的模型性能，采用较短的时间长度会造成信息丢失，而采用较长的时间长度造成较长的训练时间同时由于目标运动状态变化影响检测性能，经过实验验证，本文时间序列长度t设置为10可以取得较好的实验结果。

3.4 实验结果与分析

为了验证本文Bi-LSTM模型的有效性和创新性，将网络模型与恒定速度阈值法、4种经典的机器学习分类方法和卷积神经网络模型做了对比实验。在恒定速度阈值方法(Constant Velocity Threshold Algorithm, CVTA)中，通过对数据集上的速度分析，设定恒定速度阈值为20 m/s，依次计算每个轨迹点的速度值，速度值超过阈值则判定为异常点。4种分类算法在工业界都得到了广泛使用，分别为：逻辑回归(Logistic Regression, LR)，决策树(Decision Tree, DT)，随机森林(Random Forest, RF), XGBoost (eXtreme Gradient Boosting)。本文采用scikit-learn包[24]实现了上述4种分类算法，其输入为每个轨迹点的6维运动特征向量。CNN网络结构参考文献[16]，修改模型输入为轨迹点的6维向量，输出为2维向量。

本文采用5折交叉验证法对4种分类模型分别调整了参数，通过统计算法分类精度随每个参数的变化，选择分类精度达到最高值时的参数值，从而使得这些分类算法均能获得较好性能。表2是不同方法在测试集上的检测性能指标及所用时间的比较。图6描述了不同分类模型的ROC曲线图。从表2和图6可以看出，本文提出的算法各项指标均高于其他对比方法，尤其是召回率和AUC值，Bi-LSTM模型远高于其他机器学习分类算法和速度阈值方法。高召回率意味着模型检测到更多真实的异常点，这在异常检测问题中是非常重要的。

表 2 不同方法指标对比

Tab. 2　The performance of different models

模型	分类精度	准确率	召回率	F1值	AUC值	测试时长(s)
CVTA	0.966	0.398	0.804	0.532	无	0.084
LR	0.981	0.857	0.101	0.180	0.550	1.603
DT	0.972	0.397	0.612	0.481	0.796	0.139
RF	0.989	0.811	0.623	0.705	0.810	0.317
XGBoost	0.980	0.517	0.640	0.572	0.813	0.398
CNN	0.983	0.828	0.268	0.405	0.633	4.069
Bi-LSTM	**0.995**	**0.873**	**0.902**	**0.887**	**0.950**	0.948

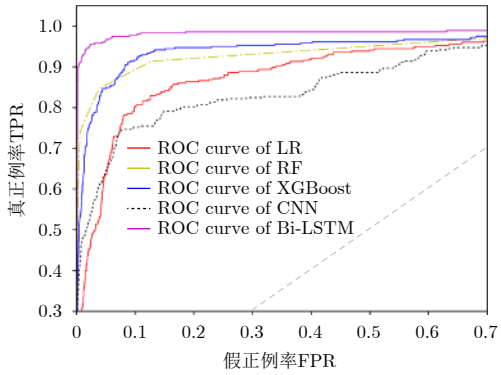


图 6 不同模型的ROC曲线图

Fig. 6 ROC curves of different models

在轨迹异常点检测问题中，一个异常轨迹点通常跟其周围正常的邻近点在运动特征上有着较大的差异。由于恒定速度阈值法未考虑每个轨迹段上目标运动状态，LR, DT, RF, XGBoost算法和CNN模型仅使用了当前检测点的特征信息，没有考虑到当前点相邻时间内的目标运动信息，这些算法检测性能均较低。特别是LR和CNN模型召回率较低，这意味着算法漏检了许多轨迹异常点。从测试时长对比可以发现，本文Bi-LSTM模型速度快于同样为深度学习的CNN网络，可以用于轨迹异常点的快速检测。

Bi-LSTM模型利用了当前点和其前后t时刻内的轨迹上下文信息。由于LSTM单元和双向网络的独特设计，双向LSTM能够自动学习正常点和异常点在序列运动特征上的差异性，免除了与数据时序性相关繁重的特征工程，为异常检测提供有效决策支持。本文提出的Bi-LSTM模型还可以方便地移植到其他不同轨迹数据或类似的处理任务上。在类别已标记好的轨迹数据集上，Bi-LSTM模型可以预先训练好。在对轨迹数据进行后续挖掘任务之前，可以采用预训练好Bi-LSTM模型准确地去除异常点，从而大幅度提高数据质量，提高挖掘任务精度。

4 结论

本文提出了一种基于双向长短时记忆网络模型的轨迹异常点检测算法。首先对每个轨迹点提取了6维的运动特征向量来表示轨迹点，然后将轨迹异常点检测问题转化为有监督的分类问题，通过对轨迹异常点检测问题的分析，构建了Bi-LSTM模型来自动学习一段长度轨迹数据中的抽象特征。同时，采取了过采样和欠采样的组合方法缓解类别不平衡对算法性能的影响。Bi-LSTM考虑了轨迹点的历史和未来信息，适用于离线处理时准确地检测异常点，模型训练好后检测过程非常快速，而且扩展性强。在真实的船舶轨迹标注数据集上，实验结果表明算法相对于不考虑时序特征的机器学习经典分类算法和卷积神经网络的有效性。在下一步工作中，本文将研究采用LSTM的变体单元(如QRNN, SRU)，通过提高网络的计算速度来加快轨迹异常点的检测速度，同时，将算法扩展至更多不同目标的轨迹数据上也是其中的一个重要研究方向。

参 考 文 献

[1] LEE J G, HAN J W, and WHANG K Y. Trajectory clustering: A partition-and-group framework[C]. Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, Beijing, China, 2007: 593–604.

[2] 李万春, 黄成峰. 基于角度和多普勒频率的外辐射源定位系统的接收器最优航迹分析[J]. 雷达学报, 2014, 3(6): 660–665. doi: 10.12000/JR14118.
LI Wan-chun and HUANG Cheng-feng. Optimal trajectory analysis for the receiver of passive location systems using direction of arrival and Doppler measurements[J]. *Journal of Radars*, 2014, 3(6): 660–665. doi: 10.12000/JR14118.

[3] 齐林, 王海鹏, 刘瑜. 基于统计双门限的中断航迹配对关联算法[J]. 雷达学报, 2015, 4(3): 301–308. doi: 10.12000/JR14077.
QI Lin, WANG Hai-peng, and LIU Yu. Track segment association algorithm based on statistical binary thresholds[J]. *Journal of Radars*, 2015, 4(3): 301–308. doi: 10.12000/JR14077.

[4] ZHENG Y, LIU L K, WANG L H, *et al*. Learning transportation mode from raw GPS data for geographic applications on the web[C]. Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 2008: 247–256. doi: 10.1145/1367497.1367532.

[5] BAO J, ZHENG Y, and MOKBEL M F. Location-based and preference-aware recommendation using sparse geo-social networking data[C]. Proceedings of the 20th International Conference on Advances in Geographic Information Systems, Redondo Beach, California, 2012: 199–208. doi: 10.1145/2424321.2424348.

[6] ZHANG D Q, LI N, ZHOU Z H, *et al*. iBAT: Detecting anomalous taxi trajectories from GPS traces[C]. Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 2011: 99–108. doi: 10.1145/2030112.2030127.

[7] ZHENG Y and ZHOU X F. Computing with Spatial Trajectories[M]. New York: Springer Science & Business Media, 2011.

[8] HAWKINS D M. Identification of Outliers[M]. Dordrecht: Springer, 1980.

[9] HAN J W, KAMBER M, and PEI J. Data Mining: Concepts and Techniques[M]. Amsterdam: Elsevier, 2011.

[10] ALVARES L O, OLIVEIRA G, and BOGORNY V. A

framework for trajectory data preprocessing for data mining[C]. Proceedings of the 21st Interantional Conference on Software Engineering & Knowledge Engineering, SEKE, Boston, USA, 2009, 21: 698–702.

[11] CHEN L, LV M Q, YE Q, *et al*. A personal route prediction system based on trajectory data mining[J]. *Information Sciences*, 2011, 181(7): 1264–1284. doi: 10.1016/j.ins.2010.11.035.

[12] ZHENG Y, XIE X, and MA W Y. Geolife: A collaborative social networking service among user, location and trajectory[J]. *IEEE Data Engineering Bulletin*, 2010, 33(2): 32–39.

[13] CHEN X J, CUI T T, FU J H, *et al*. Trend-residual dual modeling for detection of outliers in low-cost GPS trajectories[J]. *Sensors*, 2016, 16(12): 2036. doi: 10.3390/s16122036.

[14] 胡晶. 基于AIS的船舶轨迹分析与应用系统的设计与实现[D]. [硕士论文], 华中师范大学, 2015.
HU Jing. Design and implementation of vessel trajectory analysis and application system based on AIS[D]. [Master dissertation], Central China Normal University, 2015.

[15] 吴建华, 吴琛, 刘文, 等. 舶舶AIS轨迹异常的自动检测与修复算法[J]. 中国航海, 2017, 40(1): 8–12, 101. doi: 10.3969/j.issn.1000-4653.2017.01.003.
WU Jian-hua, WU Chen, LIU Wen, *et al*. Automatic detection and restoration algorithm for trajectory anomalies of ship AIS[J]. *Navigation of China*, 2017, 40(1): 8–12, 101. doi: 10.3969/j.issn.1000-4653.2017.01.003.

[16] BESSA A, DE MESENTIER SILVA F, NOGUEIRA R F, *et al*. RioBusData: Outlier detection in bus routes of rio de janeiro[OL]. arXiv: 160106128, 2016.

[17] FERNANDO T, DENMAN S, SRIDHARAN S, *et al*. Soft+hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection[OL]. arXiv: 1702.05552, 2017.

[18] HOCHREITER S and SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

[19] GRAVES A and SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. *Neural Networks*, 2005, 18(5/6): 602–610. doi: 10.1016/j.neunet.2005.06.042.

[20] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, *et al*. Dropout: A simple way to prevent neural networks from overfitting[J]. *Journal of Machine Learning Research*, 2014, 15: 1929–1958.

[21] BATISTA G E A P A, PRATI R C, and MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. *ACM SIGKDD Explorations Newsletter*, 2004, 6(1): 20–29. doi: 10.1145/1007730.1007735.

[22] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Zhi-hua. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.

[23] FAWCETT T. An introduction to ROC analysis[J]. *Pattern Recognition Letters*, 2006, 27(8): 861–874. doi: 10.1016/j.patrec.2005.10.010.

[24] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, *et al*. Scikit-learn: Machine learning in python[J]. *Journal of Machine Learning Research*, 2011, 12: 2825–2830.

作 者 简 介

韩昭蓉(1992–)，女，山西运城人，2015年在西安电子科技大学获得工学学士学位，现为中国科学院大学、中国科学院电子学研究所硕士研究生，研究方向为轨迹数据异常点检测、机器学习。
E-mail: hanzhaorong15@mails.ucas.ac.cn

黄廷磊(1971–)，男，安徽肥东人，博士后，研究员，博士生导师，入选中国科学院“百人计划”并获择优支持。2000年在上海理工大学获得博士学位，现为中国科学院电子学研究所研究员，中国科学院电子所空间智能处理系统研究室主任，主要研究方向为数据挖掘、空间大数据组织管理与可视化。
E-mail: tlhuang@mail.ie.ac.cn

任文娟(1982–)，女，河南焦作人，副研究员，博士，2011年在中国科学院电子学研究所获得博士学位，现为中国科学院电子学研究所中国科学院空间信息处理与应用系统技术重点实验室副研究员，主要研究方向为多源遥感信息融合处理与应用技术。
E-mail: wjren@mail.ie.ac.cn

许光銮(1978–)，男，浙江天台人，研究员，博士生导师，2005年在中国科学院电子学研究所获得博士学位，现为中国科学院电子学研究所研究员，中国科学院空间信息处理与应用系统技术重点实验室主任，主要研究方向为地理空间信息挖掘与应用技术。
E-mail: gluanxu@mail.ie.ac.cn