

基于 1DCNN-LSTM 的船舶轨迹分类方法

崔彤彤 王桂玲 高 晶

大规模流数据集成与分析技术北京市重点实验室(北方工业大学) 北京 100144

北方工业大学信息学院 北京 100144

(ctt930814@163.com)



摘 要 由于监控设备视野有限、代价昂贵等问题,导致基于船舶图像或视频的船舶分类效果欠佳,改进船舶分类方法、提高船舶分类的准确率迫在眉睫。近几年,随着各类轨迹数据采集系统的兴起,通过船舶航行轨迹数据实现船舶类型的分类逐渐成为可能。针对使用传统二维卷积神经网络(Convolutional Neural Network,CNN)对船舶轨迹分类存在特征压缩和时序特征表达能力匮乏的问题,文中提出了一种一维 CNN(One-Dimensional CNN,1DCNN)与长短期记忆网络(Long Short-Term Memory,LSTM)相结合的混合分类模型,对采集到的船舶自动识别系统(Automatic Identification System,AIS)数据进行船舶类型识别。首先,针对 AIS 采集到的船舶轨迹数据进行预处理,过滤噪声数据;然后,针对隐含在船舶轨迹信息中的特征对于 1DCNN 而言过于隐晦的问题,提出了一种针对大规模船舶轨迹数据的,且 1DCNN 能够识别的轨迹分布特征向量的构建算法,同时在此基础上提取了 LSTM 能够识别的时序特征向量;最后,将训练后的 1DCNN 模型与 LSTM 模型进行数据融合得到混合船舶分类模型。以渤海区域 2016 年 6 月的船舶 AIS 数据为基础,使用 1DCNN 与 LSTM 相结合的混合模型对渔船、客船、油船、集装箱船和散货船 5 类典型船舶的轨迹数据进行分类,并将其与使用一种神经网络如 LSTM 作为分类器的方法进行对比,结果表明所提方法具有明显的有效性,是一种有效的船舶轨迹分类方法。

关键词: 船舶轨迹分类;一维卷积神经网络;长短期记忆网络;分布特征向量;时序特征向量

中图法分类号 TP183

Ship Trajectory Classification Method Based on 1DCNN-LSTM

CUI Tong-tong, WANG Gui-ling and GAO Jing

Beijing Key Laboratory on Integration and Analysis of Large-scale Stream Data (North China University of Technology), Beijing 100144, China

School of Information, North China University of Technology, Beijing 100144, China

Abstract Due to the limited vision and cost of the monitoring equipment, the classification methods of ships based on images or videos are not very effective. So it is urgent to improve classification methods of the ships and the accuracy of those methods. In recent years, with the widely used of various trajectory data acquisition systems, it has become possible to classify ship types through ship trajectory data. Based on the problem that the traditional two-dimensional convolutional neural network is lacking the ability of feature compression and temporal feature expression in ship trajectory recognition, this paper proposes a hybrid model which combines one-dimensional convolutional neural network (1DCNN) with long short-term memory (LSTM). This model can identify ship types by using the data collected from the automatic identification system (AIS). Firstly, this paper pre-processes the ship trajectory data collected by AIS to filter the noise data. Secondly, to solve the problem that the features hidden in the original ship trajectory information are over obscurity for 1DCNN, this paper proposes an algorithm for constructing the trajectory distribution feature vectors which can be accepted by 1DCNN for a large number of ship trajectory data. On this basis, the algorithm extracts the time series feature vectors which can be accepted by LSTM. Finally, this paper combines the trained 1DCNN model and LSTM model to get a hybrid ship classification model. Based on the AIS data of Bohai area on June 2016, the hybrid model combining 1DCNN with LSTM is used to classify five different typical ships including fishing ships, passenger ships, tanker ships, container ships and bulk-cargo ships. The experimental results show that compared with the method of using

到稿日期:2019-10-24 返修日期:2020-03-28 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:北京市自然科学基金(4172018);国家自然科学基金(61832004,61672042);中电科海洋信息技术研究院有限公司高校合作课题(402054841879);北方工业大学毓优团队培养计划项目(107051360018XN012/020)

This work was supported by Beijing Natural Science Foundation (4172018), National Natural Science Foundation of China (61832004, 61672042), University of China Electric Ocean Information Technology Research Institute Co., Ltd(402054841879) and YuYou Team Training Project of North China University of Technology (107051360018XN012/020).

通信作者:王桂玲(wangguiling@ict.ac.cn)

a neural network such as LSTM as classifier, the proposed method is obviously effective, and is an effective ship trajectory classification method.

Keywords Ship trajectory classification, One-dimensional convolutional neural network (1DCNN), Long short-term memory (LSTM), Distribution feature vector, Time series feature vector

1 引言

分类是人工智能、模式识别和数据挖掘领域中重要的数据处理方法,其目的是通过样本数据建立分类模型,将实体属性值分类到给定的数据类别中。轨迹分类是根据训练轨迹的特征来判断未知轨迹的种类。时空轨迹是移动对象的位置和时间的记录序列,例如飞机、车辆和船舶的时空轨迹等。轨迹分类一直是交通运输工程、交通地理学的重要研究内容^[1]。

目前,随着智能交通的迅猛发展,国内外对船舶类型的分类研究主要是基于监控设备采集到的船舶图片或视频的检测技术来实现。船舶分类在众多行业均有重要的作用,例如渔业控制部门检测非法渔船,海上安全部门预防海上恐怖主义、走私等^[2]。但是,由于监控设备视野有限,有时不能准确地采集到需要识别海域的船舶图片和视频信息,即使采集到也代价昂贵,这对于海上航道安全和航道监管有一定的影响。随着各类轨迹数据采集系统的兴起,通过船舶航行轨迹实现船舶类型的分类逐渐成为可能。船舶自动识别系统是一种开放式数据传输系统,它的广泛应用有力地推动了航运智能化发展,并在船舶交通信息采集与分析、船舶航行监测以及海事空间规划中发挥了重要作用^[3]。通过 AIS 收集到的一系列船舶位置数据,具有海量性、地理范围大等特点,但是数据质量不高,给船舶分类带来了极大的挑战。

近年来,将船舶 AIS 数据与神经网络相结合来实现船舶分类逐渐备受关注。一维卷积神经网络和长短期记忆网络是两种常见的处理文本数据的多层神经网络。1DCNN 主要针对数据的整体特征分块进行卷积,实现文本数据的特征学习,同时解决了传统的二维 CNN 对输入数据进行维度压缩而损失大量的特征信息的问题。LSTM 主要针对数据的时序特征进行学习,从而解决了传统方法对于数据的时序特征学习能力匮乏的问题。由于 1DCNN 和 LSTM 模型常见的应用是以文本数据向量化后的数值向量为输入,而不是直接接受原始数据作为输入,而船舶轨迹数据通过适当处理后可生成满足神经网络模型输入要求的数值向量,因此,本文提出了一种基于 1DCNN 和 LSTM 的混合模型与船舶轨迹相结合来实现船舶轨迹分类的方法。本文的具体内容和贡献如下:

- (1)针对 AIS 船舶轨迹数据具有海量性、质量不高的特点,本文采用 MapReduce 框架对船舶轨迹数据进行并行化预处理,以过滤噪声数据;
- (2)提出了基于 MapReduce 并行化计算框架船舶分布特征向量提取算法和时序特征向量提取算法,从预处理 AIS 数据中提取向量,并将其作为最终分类模型的输入;
- (3)建立 1DCNN-LSTM 的混合分类模型,分别对提取到的分布特征向量和时序特征向量进行学习,将得到的结果进行加权数据融合,最终实现渔船、客船、油船、集装箱船和散货船 5 种船舶轨迹的有效分类。

2 相关工作

由于船舶轨迹数据采集的局限性,针对船舶轨迹数据的

分类研究并不是很多,已有研究方法主要分为基于轨迹间距离的聚类方法、人工提取特征和机器学习方法相结合的分类方法和神经网络分类方法。神经网络分类方法主要包括船舶图片或视频与神经网络相结合以及基于船舶轨迹的时序特性与神经网络相结合来实现分类。

Cao 等^[4]基于 AIS 数据绘制船舶分布的密度图,根据稠密区域来识别航线。识别主要航线的方法只考虑了不同船舶轨迹的行驶位置的差异,未考虑其他因素,在判断船舶类型时容易混淆。因此有研究学者在船舶轨迹的数据分析与研究中,加入船舶轨迹聚类这一方法,来找出具有相似船舶运动演化方式的轨迹簇。Liu 等^[5]提出了基于 KNN (K-Nearest Neighbor)的船舶轨迹分类算法,通过船舶轨迹的初步聚类,得到 KNN 分类样本轨迹,然后将综合距离作为 KNN 分类中轨迹间的距离来实现船舶轨迹分类。但结论得出该类方法更适用于内河船舶轨迹分类。Zhao 等^[6]针对数据的特征,利用改进支持向量机(Support Vector Machine, SVM)的方法实现分类。由于船舶分类的最佳特征选取往往取决于特定的数据或应用,而针对船舶轨迹数据人工提取到的特征往往具有局限性,提取的有限特征不能全面地表达各类船舶的特点。近年来,相比人工特征提取与机器学习相结合的方法,深度神经网络逐渐成为研究热点之一, CNN 和循环神经网络(Recurrent Neural Network, RNN)在各个研究方向被广泛使用。前者通过识别待分类对象整体更深层次的特征在图片领域取得了巨大成功,而后者通过结合 LSTM^[7]模型能够胜任时间序列的分类研究,但在航海领域的研究较少。Endo 等^[8]在使用 CNN 进行图像分类的基础上,提出了将原始轨迹数据结构转换成图像数据结构,同时保留有效的时空信息,然后使用 CNN 自动提取特征的方法来实现船舶行为模式的识别。该方法同样对航行距离较短或数据量较小的船舶轨迹有着较好的识别效果,针对海洋上船舶的长时间航行,在数据转换成图片的过程中,由于距离过长而在规定的范围内不能得到明显的特征,因此对最终的分类结果有一定的负面影响。Ljunggren^[9]对船舶轨迹进行特征提取后将其作为输入,训练了不同大小的卷积神经网络来对正在移动的渔船、货船、油船和客船进行分类,但通过人工方式来提取船舶的速度和航首向的均值、中位数等特征,使网络提取到的更深层次的特征具有局限性,并且没有考虑到船舶轨迹数据中时序特征的作用,因此其效果在传统机器学习方法上有所提高但仍有所不足。因为船舶轨迹数据属于时序数据,本文在研究工作中又参考了部分针对时序数据实现类型识别的方法, Zhang 等^[10]提出了一种分类模型将 CNN 和 RNN 相结合,实现了在大量的时间序列数据上的分类,但结论认为学习特征不够全面,结果还有待提高。

综上,本文考虑到船舶数据的海量性和时序性以及在整个范围内不同类型的船舶轨迹有各自的特点,采用将 1DCNN 和 LSTM 相结合的混合模型来实现船舶轨迹的分类。

3 并行化轨迹的预处理和特征向量提取算法

3.1 船舶 AIS 轨迹数据的预处理

由于船舶轨迹数据的采集间隔比较稀疏、测量误差等原因,船舶 AIS 数据会产生噪声数据,且采集到的船舶数据量非常庞大^[11],这样的数据会对最终的分类准确率有一定的负面影响。因此,实验中首先需要对采集到的原始 AIS 数据进行预处理,本文对船舶轨迹的预处理主要为过滤噪声数据。考虑到船舶在正常航行过程中不会频繁转向,产生的轨迹在一定范围内近似于直线,因此采用基于时间和空间阈值的过滤算法。该算法的原理是计算点到轨迹的距离和时间间隔,如果点到轨迹的距离和时间间隔均大于或其中之一大于预定阈值,则删除该点;若划分出的轨迹段内轨迹点的数目少于预定阈值,则删除该轨迹段,以消除噪声点。

本文采用的数据来源于 2016 年中国渤海湾区域内渔船、客船、油船、集装箱船和散货船的轨迹数据,其中一个月的数据大小约为 60GB,包括 2 亿余条记录,保存在 Hadoop 分布式文件系统(Hadoop Distributed File System, HDFS)中。每一条记录包括 27 个字段,本文在使用数据时对其进行筛选,将每条记录处理为由表 1 所列的 4 个字段组成。

表 1 筛选数据字段
Table 1 Fields description of filtered data

Name	Meaning	Sample	Explain
utc	Timestamp	1 483 200 258	Collection time
mmsi	Maritime mobile service identify	205 306 000	Ship unique communication code
lon	Longitude	122. 121 405	Longitude of the ship collection point
lat	Latitude	38. 618 043	Latitude of the ship collection point

本文使用 MapReduce 并行化计算框架来实现预处理算法,以提升效率。具体步骤如下。

(1)数据分割:将 HDFS 中保存原始数据的文件进行分割,划分成 $n(n \gg m, m$ 表示数据节点数)个数据块,每个块交由一个数据节点 *Datanode* 处理。

(2)Map 阶段:输入为原始的船舶 AIS 数据;Map 函数从数据中提取 *utc, mmsi, lon, lat* 这 4 个字段;输出为以 *mmsi* 为 *key*,以提取到的属性信息(*utc, lon, lat*)为 *value* 的 *key-value* 对,其形式为 $\langle mmsi, (utc, lon, lat) \rangle$ 。即按照 *mmsi* 号来保存每条船舶轨迹的数据信息。

(3)Reduce 阶段:输入为 Map 阶段的输出,每个 *reduce* 处理相同 *key* 的数据。

第 1 步 将相同 *mmsi* 的数据按照 *utc* 排序,设定时间阈值 *t* 和距离阈值 *d*,以及轨迹点数目阈值 n_z ;

第 2 步 将第 *i* 个轨迹点保存在集合 *list1* 中,计算第 *i+1* 个点和第 *i* 个点之间的距离和时间间隔并与阈值 *d* 和 *t* 进行比较,如果同时满足,则保存第 *i+1* 个点,然后 $i=i+1$,继续执行第 2 步,否则执行第 3 步;

第 3 步 计算集合 *list1* 中点的个数 *N*,如果 $N < n_z$,则将 *list1* 中的点视为噪声点,并清空 *list1*, $i=i+1$,执行第 2 步;若 $N \geq n_z$,则将集合 *list1* 中的点视为一段轨迹,并执行第 4 步;

第 4 步 保存 *list1* 中的点,并以 *mmsi* 为 *key*,以元组 (*utc, lon, lat*)为 *value* 输出,其形式为 $\langle mmsi, (utc, lon, lat) \rangle$ 。然后清空 *list1*, $i=i+1$,执行第 2 步,直到遍历完相同 *mmsi* 的数据。

(4)将预处理后的数据按照 *mmsi* 保存在不同的文件中,包括 *mmsi, utc, lon, lat* 4 个字段。

基于 MapReduce 的轨迹预处理算法流程如图 1 所示。

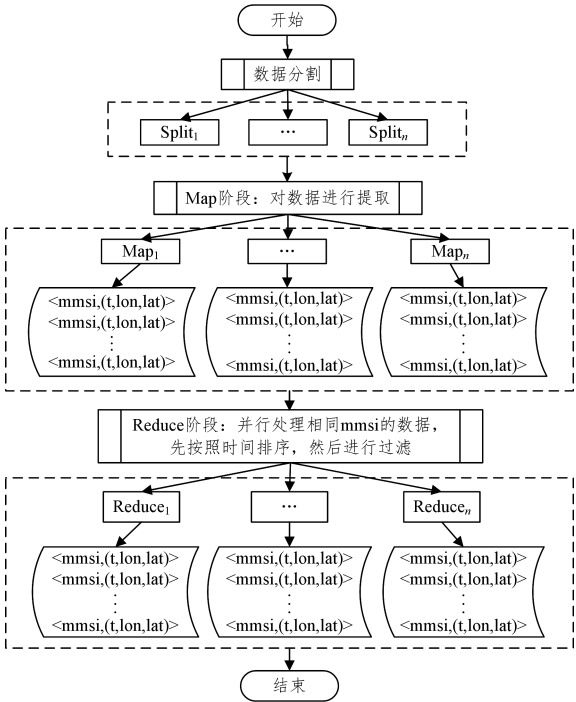


图 1 基于 MapReduce 的轨迹预处理流程图

Fig. 1 Flow chart of trajectory preprocessing based on MapReduce

3.2 船舶轨迹数据的分布特征向量及其构建

经过预处理后的船舶 AIS 数据包含时间戳、经度和纬度 3 种基本时空信息。对于分类模型来说,基本时空信息不能直观地表达船舶类型信息,分类模型难以建立船舶时空信息到船舶轨迹类型的有效映射。只有将这些时空信息转化成分类模型可以理解的数据后,才能实现有效的船舶轨迹分类。因此,在分类模型对船舶轨迹类型进行识别前,一个重要的步骤就是构建特征向量^[12]。特征向量是指分类模型能接受的输入数据。已有分类方法针对船舶轨迹的特征分析主要是从提取到的各属性特征的概率统计的角度进行,如计算属性的均值、方差和中位数等统计量^[13]。使用统计量有如下缺点: 1)统计量需要人工计算求得,若选取的特征种类过多,会增加人工计算量;2)一些深层次的特征对于使用人工提取和计算也有很大的难度。因此,本文使用分布特征向量作为分类模型的输入数据^[14]。分布特征向量是指在一段轨迹中,提取到的特征在特定范围内占特征总数的比例所组成的向量。通过使用分布特征向量,分类模型能够分析轨迹段整体的特点,同时也能分析轨迹段各局部之间的联系,自主学习船舶轨迹中的基本特征和深层次特征,实现船舶轨迹的分类。

本文通过对已有传统分类算法进行研究以及比较得出,航速、加速度、航首向和曲率这 4 个属性是区分船舶类型的最有效属性。其中,航首向是指以地理北为基准,顺时针到船舶

在某一时刻船头正对方向之间的夹角,范围为 $[0^\circ, 360^\circ]$ 。曲率是指船舶在航行过程中在某一时刻轨迹偏离直线行驶的程度。本文选择这4个属性值在规定区间内出现的次数占总体次数的比例作为分布特征向量的组成元素。分布特征向量的具体构建方法如下。

第1步 根据统计学的方法^[14]计算轨迹数据预处理后划分的每一段轨迹的航速、加速度、航首向与曲率,将包含 n 个数据点的每一段轨迹按照式(1)所示矩阵的形式表示出来,矩阵中 $v_i, a_i, h_i, s_i (i=1, 2, \dots, n)$ 分别表示轨迹段中第 i 个轨迹点的航速、加速度、航首向和曲率。

$$D = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \\ a_1 & a_2 & \cdots & a_n \\ h_1 & h_2 & \cdots & h_n \\ s_1 & s_2 & \cdots & s_n \end{bmatrix} \quad (1)$$

第2步 根据划分后针对每一段轨迹计算得到的航速、加速度、航首向和曲率4个属性的最大值和最小值来确定各属性取值区间,并分别对各属性的取值区间进行归一化处理。最终,航速、加速度、航首向和曲率的取值区间分别为: $[0, 15]$, $[-36, 36]$, $[0, 360]$ 和 $[0, 10]$ 。同时,规定分布特征向量的维度为60维,即4个属性的取值区间平均分为60个子区间。加速度的正负表示相对于船舶航行前进的方向,与前进方向相同为正,反之为负。然后统计各子区间内每个属性值出现的次数,并计算其占总次数的比例,按照如式(2)所示矩阵表示出来,其中第 $i (i=1, 2, \dots, 60)$ 列表示第 i 个子区间内该属性的出现次数占总体次数的比例。

$$D' = \begin{bmatrix} v_1' & v_2' & \cdots & v_{60}' \\ a_1' & a_2' & \cdots & a_{60}' \\ h_1' & h_2' & \cdots & h_{60}' \\ s_1' & s_2' & \cdots & s_{60}' \end{bmatrix} \quad (2)$$

分布矩阵 D' 不仅可以捕获到航速、加速度、航首向和曲率的最大值、最小值和均值等基本统计量,还可以捕获到需要进一步计算得到的中位数等其他统计量,根据速度的分布还可以获取船舶的停顿率。与单独统计属性的统计量相比,该方法可以包含人工提取有困难的各属性完整的统计量信息。

第3步 将分布矩阵 D' 平铺排列,可以得到最终作为输入的 1×240 的分布特征向量,向量从左到右每一个元素分别表示每段轨迹中航速、加速度、航首向和曲率在当前区间的出现次数占总范围内次数的比例,如式(3)所示:

$$d' = [v_1' \cdots v_{60}' a_1' \cdots a_{60}' h_1' \cdots h_{60}' s_1' \cdots s_{60}'] \quad (3)$$

该向量最终展现了一条轨迹的特征分布,某特征的值越大表示该特征在此区间内出现的频率越高。本文选取了某一时间段内的渔船、客船、油船、集装箱船和散货船的典型轨迹,提取了分布特征向量,结果如图2所示。图2中,区间 $[0, 60]$ 表示航速, $(60, 120]$ 表示加速度, $(120, 180]$ 表示航首向, $(180, 240]$ 表示曲率,图2中值越高的地方表示对应特征在此区间内的可能性越大。可以看到,在航速方面,客船航速较高,渔船较低,且分布较为集中;同属于货船类的油船、集装箱船和散货船的速度分布相对而言略微分散。在加速度方面,因为船舶在海上的运动趋于平稳状态,所以5类船舶加速度主要分布在中间数值较小的区域,但集装箱船和散货船的分

布明显略微分散。同理,曲率也相对集中分布在较小的区域,但油船和集装箱船的曲率和其他船舶类型相比有明显差异。因为船舶在海上运动,航首向与航程的变化有关,不同航程路线对应的航行方向一般不同,所以分布会比较均匀,但集装箱船的分布明显较为集中。由此可以看出,5种船舶类型的各个属性分布在整体上有各自的特点,因此可以使用分布特征向量作为分类模型的输入,对各船舶类型的特征进行学习。

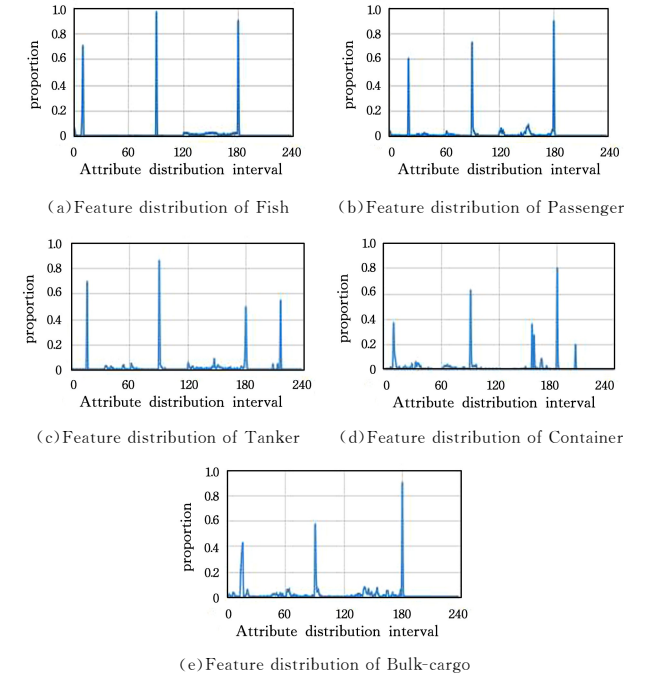


图2 5种类型船舶的4种特征分布的对比图

Fig. 2 Distribution map of four features of five types of ships

3.3 船舶轨迹数据的时序特征向量及其构建

由图2可知,分布特征向量能够捕捉到完整的轨迹属性的分布特征,因此分布特征向量更接近于对应船舶类型的一般情况。若考虑到船舶轨迹随时间的变化,则5种类型的船舶的航行状态更加随机,属性的分布将不再具有代表意义。因此,本文同时采用时序特征向量作为分类模型的输入向量。时序特征向量,是指船舶轨迹中能够随时间变化的属性特征所组成的向量。在时序上,航速的变化最明显,其能够刻画在同一时间段内同属于货船的油船、集装箱船和散货船的速度具体变化。加速度的变化与速度的变化相关。航首向作为船舶的行驶方向属性,也随着时间的推移而变化。由文献[15]可知,油船在航行过程中的航向稳定性比散货船和集装箱船更好,同时,船舶在不同航道航行过程中曲率的变化也有细微的差别,与航首向的变化相关。因此,针对轨迹分段后在同一时间段的轨迹,提取计算得到的每个采集时刻的航速和航首向两种特征,用于构建时序特征向量并将其作为分类器的输入向量,进行船舶的分类。

时序特征向量的组成元素主要为每一段轨迹中每个采集时刻的时间戳,以及通过计算得到的该时刻的航速和航首向。式(4)所示的向量 T_i 表示某类船舶在 t 时刻的属性元素所组成的时序特征向量。其中, utc_i, v_i, h_i 分别表示该时刻的时间戳、航速和航首向,当使用分类模型对船舶类型分类时,将同一时间段内各船舶的时序特征向量作为分类模型的输入。

$$T_i = [utc_i \quad v_i \quad h_i] \quad (4)$$

3.4 基于 MapReduce 的分布特征向量提取算法和时序特征向量提取算法

本文针对大规模的船舶 AIS 数据,提出了基于 MapReduce 并行化框架的分布特征向量提取算法和时序特征向量提取算法。

3.4.1 分布特征向量构建算法

提取训练数据和测试数据时需添加各船舶类型标签在特征向量的最后一列。不添加标签时,针对各类型船舶的分布特征向量算法构建的核心步骤如下:

(1)Map 阶段:Map 函数从数据中提取 (utc, lon, lat) 这 3 个字段作为 $value$,将 $mmsi$ 号作为 key 值输出到 $reduce$,输出形式为 $\langle mmsi, (utc, lon, lat) \rangle$ 。

(2)Reduce 阶段:

第 1 步 首先遍历输入到 $reduce$ 阶段的 $value$ 值,将相同 $mmsi$ 号的数据按照时间戳从小到大进行排序。

第 2 步 因为每条船包含多段轨迹,所以设定轨迹分段的阈值,包括时间阈值 t 、距离阈值 d ,以及轨迹段数目阈值 n 。

第 3 步 分别计算每个轨迹点的航速、加速度航首向和曲率作为属性添加到轨迹点的属性中,最终每个轨迹点应包含 $(mmsi, utc, lon, lat, v, a, h, s)$ 7 个属性,其中 v 表示航速, a 表示加速度, h 表示航首向, s 表示曲率。

第 4 步 设定各个属性的阈值取值区间 $[v_{min}, v_{max}]$, $[a_{min}, a_{max}]$, $[h_{min}, h_{max}]$, $[s_{min}, s_{max}]$,并按照设定的维度 m 进行划分,并设定每个子区间中出现属性值的个数 $x=0$,并将其对应保存在数组 v, a, h, s 中,即 v, a, h, s 4 个数组中每个元素为航速、加速度、航首向和曲率的取值子区间内属性值出现的个数。在本阶段中,数组内的元素均为 0。

第 5 步 遍历轨迹数据,将第 i 个点放入序列 $list1$ 中,判断第 $i+1$ 个点与第 i 个点的距离和时间间隔,如果计算得到的距离或时间间隔中有一项小于设定的阈值,则将第 $i+1$ 个点加入到该段轨迹的序列 $list1$ 中, $i=i+1$,继续执行第 5 步;否则计算该段轨迹中轨迹点的个数 N ,与轨迹段数目阈值进行比较,当 $N \geq n$ 时,执行第 6 步,否则清空 $list1$, $i=i+1$,继续执行第 5 步。

第 6 步 遍历 $list1$,判断第 i 个轨迹点 v 属于的 $v[p]$, $v[p] = v[p] + 1$ (p 表示阈值范围切分后子范围的序号);同理,对加速度、曲率和航首向也进行上述计算。 $i=i+1$,继续执行第 6 步,直到遍历完 $list1$ 中的轨迹点。

第 7 步 求数组 v, a, h, s 中各数据占轨迹段数据总数的比例,并分别保存在数组 $V_{mmsi}, A_{mmsi}, H_{mmsi}, S_{mmsi}$ 中,清空数组 v, a, h, s ;同时将数组 $V_{mmsi}, A_{mmsi}, H_{mmsi}, S_{mmsi}$ 保存在链表 $list2$ 中,清空 $list1$ 。继续执行第 5 步,直到遍历完相同 $mmsi$ 的数据。

第 8 步 以 $mmsi$ 为 key ,以数组 $V_{mmsi}, A_{mmsi}, H_{mmsi}, S_{mmsi}$ 为 $value$ 输出。

(3)文件输出:以 $mmsi$ 号为文件名,得到每条船中每段轨迹的分布特征向量,以多文件的形式输出。

基于 MapReduce 的分布特征向量提取算法的流程图如图 3 所示。

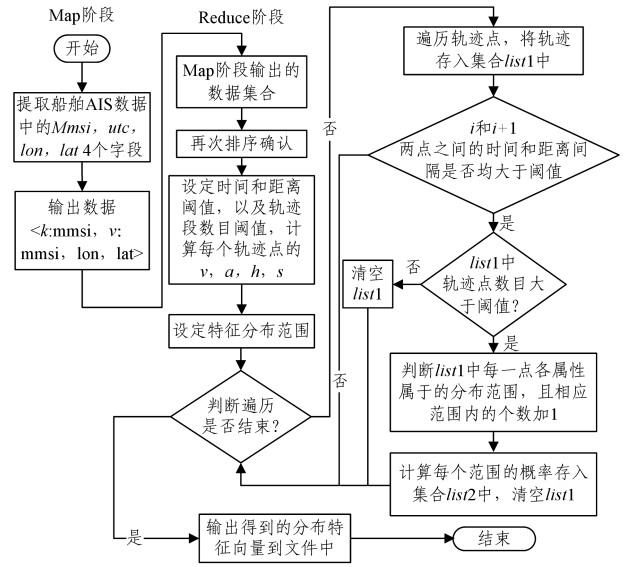


图 3 基于 MapReduce 的分布特征向量提取算法的流程图

Fig. 3 Flow chart of distribution feature vector extraction algorithm based on MapReduce

3.4.2 时序特征向量构建算法

基于 MapReduce 的时序特征向量提取算法的主要部分的流程如图 4 所示。

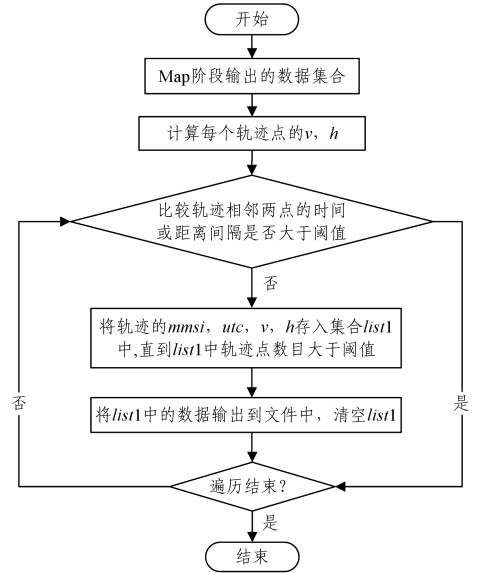


图 4 基于 MapReduce 的时序特征向量提取算法的流程图

Fig. 4 Flow chart of sequence feature vector extraction algorithm based on MapReduce

时序特征向量提取算法是对分布特征向量提取算法进行修改得到,在 $Reduce$ 阶段的主要步骤如下:

第 1 步 首先遍历输入到 $Reduce$ 阶段的 $value$ 值,将相同 $mmsi$ 号的数据按照时间戳从小到大进行排序。

第 2 步 因每条船包含多段轨迹,首先设定轨迹分段的阈值,包括时间阈值 t 、距离阈值 d ,以及轨迹段数目阈值 n 。

第 3 步 分别计算每个轨迹点的速度、航首向作为属性添加到轨迹点的属性中,最终每条轨迹应包含 $(mmsi, utc, lon, lat, v, h)$,其中 v 表示航速, h 表示航首向。

第 4 步 遍历轨迹数据,判断第 $i+1$ 个点与第 i 个点的距离和时间间隔,如果其中一个满足阈值,则加入到该段轨迹

的序列 $list1$ 中, $i=i+1$; 否则计算该段轨迹中轨迹点的个数 N , 与轨迹段数目阈值进行比较, 当 $N \geq n$ 时, 执行第 5 步, 否则, 清空 $list1$, $i=i+1$, 继续执行第 4 步。

第 5 步 遍历 $list1$, 将每个轨迹点的 utc, v, h 存入到 $list2$ 中。

第 6 步 以 $mmsi$ 为 key , 以 utc, v, h 为 $value$ 输出。

4 基于 1DCNN-LSTM 的船舶分类模型

4.1 1DCNN 和 LSTM 网络适用性分析

本文提出了一种将一维卷积神经网络(1DCNN)与长期记忆网络(LSTM)相结合的混合模型来实现船舶的分类, 其主要思路有两方面。首先, CNN 作为一种神经网络模型, 可以学习数据在整体结构上的特征, 且 CNN 模型在训练过程中能够减少神经网络参数的数量, 并且通过卷积核的权值共享使得运算量大大降低, 减轻过拟合^[16]。而 1DCNN 不会因为对文本数据的维度压缩而导致大量特征信息损失, 近年来在众多数据分类领域都有相当出色的效果。考虑到 1DCNN 模型的输入数据是文本数据向量化后的数值向量, 而船舶轨迹数据通过适当处理后可生成满足神经网络模型输入要求的数值向量。因此, 本文选用 1DCNN 作为以分布特征向量为输入数据的分类模型。其次, 船舶轨迹是基于时间的序列数据, 时序特征也是区分船舶类型的关键信息。LSTM 在 RNN 的基础上用 LSTM 层替换隐藏层, 可以保存历史信息并确定其是否可以重新被使用, 从而解决了长期依赖的问题, 近年来逐渐成为了一种可用来处理时间序列向量的典型模型^[17]。同时考虑到 LSTM 的输入数据也为数值向量, 因此, 本文将 LSTM 网络作为以时序特征向量为输入数据的分类模型。因为每种输入向量在经过特定的神经网络后都能检测到不同的特征, 对不同特征的学习可能会得出不同的结果, 但不是所有的特征都对船舶类型的识别起到作用。因此, 为了让网络更有效地进行输出判断, 在每个网络输出的船舶类型向量后增加权重作为参数参与到最终模型的训练和优化的过程, 从而对两个单独网络的分类结果作出选择。模型最终的输出结果为 1DCNN 和 LSTM 的输出结果的融合。

综上所述, 1DCNN 和 LSTM 网络分别以分布特征向量和时序特征向量为输入数据对船舶轨迹进行学习, 可以发掘不同船舶类型之间的差别。因此, 本文使用 1DCNN-LSTM 的混合模型来实现船舶轨迹的分类。

4.2 网络结构

4.2.1 1DCNN 网络结构及其参数设计

本文提出的 1DCNN 网络结构如图 5 所示。

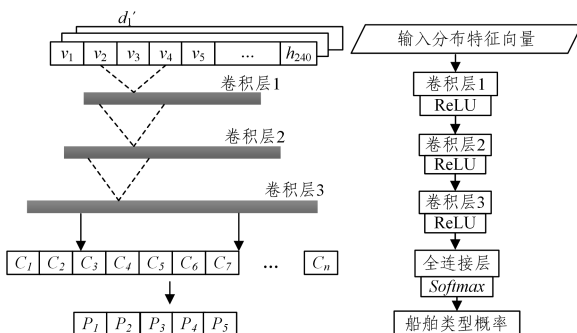


图 5 1DCNN 的设计流程图

Fig. 5 Design flow chart of 1DCNN

(1)输入层:将分布特征向量 d' 输入至 1DCNN 网络中, 利用 1DCNN 网络对 d' 的每个子集进行特征提取。

(2)卷积层:本文使用的 1DCNN 网络采用 3 个卷积层级联, 每个卷积层的输出经过非线性函数 ReLU 激活后作为下一个卷积层的输入, 为避免关键信息的损失, 未在卷积层之间加入池化层。

(3)全连接层:在最后增加一层全连接层用于分类, 激活函数为 softmax, 输出当前分布特征向量为 5 种船舶类型的概率。

输入的分布特征向量按照式(5)进行计算。

$$h_i^k = f_a((W_k * d')_i + b_k) \quad (5)$$

其中, h_i^k 是指第 i 个分布特征向量在第 k 层进行卷积运算后的值, f_a 是激活函数, W_k 是该层的权重值, b_k 是对应的偏置值。最终, 1DCNN 输出的是网络识别的 5 种船舶类型的概率向量, 如式(6)所示:

$$P = \{P_1, P_2, P_3, P_4, P_5\} \quad (6)$$

其中, $P_i (i=1, 2, 3, 4, 5)$ 分别表示渔船、客船、油船、集装箱船和散货船的概率。

4.2.2 LSTM 网络结构及其参数设计

本文提出的 LSTM 网络结构如图 6 所示。

(1)输入层:将时序特征向量 T_i (见式(8))输入至 LSTM 网络。

(2)LSTM 层:LSTM 层由一系列记忆单元 (Memory Cell) 组成, 能够保存历史信息, 如图 6 中左图所示。历史信息由 3 个门控制, 即输入门 (Input Gate)、遗忘门 (Forget Gate) 和输出门 (Output Gate), 它们能够决定历史信息是否重要到能被记住以及能不能被输出^[7]。在时间序列 $\{t | t=1, 2, \dots, n\}$ 下, 输入门在某一时刻 t 的输出 a_t^i 如式(7)所示^[18]:

$$a_t^i = \sum_{i=1}^I w_{it} T_i + \sum_{h=1}^H w_{ht} b_h^{t-1} + \sum_{c=1}^C w_{ct} s_c^{t-1} \quad (7)$$

其中, w_{ij} 表示从记忆单元 i 到 j 的连接权重, 输入门、遗忘门和输出门由下标 τ, φ 和 ω 表示。 I 和 H 分别表示输入层和 LSTM 层, C 为记忆单元的个数。对于图 6 所示的 LSTM 网络, 输入门接收到的值受到上一时刻 LSTM 输出值和所有记忆单元状态值的影响。 b_h^{t-1} 和 s_c^{t-1} 分别为上一时刻中所有不同记忆单元的输出和同一模块所有记忆单元的状态。 T_i 表示 LSTM 模型的输入, 组成某一段轨迹中各采集时刻时序特征向量, 如式(8)所示:

$$T_i = \begin{bmatrix} utc_1 & v_1 & h_1 \\ utc_2 & v_2 & h_2 \\ \vdots & \vdots & \vdots \\ utc_n & v_n & h_n \end{bmatrix} \quad (8)$$

a_t^i 经过激活函数 f 激活后输入给记忆单元。LSTM 记忆单元的遗忘门用来控制上一时刻的隐藏层状态 s_c^{t-1} , 类似输入门, 其输出如式(9)所示^[18]:

$$a_{\varphi}^t = \sum_{i=1}^I w_{i\varphi} T_i + \sum_{h=1}^H w_{h\varphi} b_h^{t-1} + \sum_{c=1}^C w_{c\varphi} s_c^{t-1} \quad (9)$$

同样, a_{φ}^t 经过激活函数 f 激活。输入门控制的本次特征和上一时刻不同记忆单元的输出 b_h^{t-1} 两部分组成记忆单元的输入, 如式(10)所示^[18]:

$$a_c^t = \sum_{i=1}^I w_{ic} T_i + \sum_{h=1}^H w_{hc} b_h^{t-1} \quad (10)$$

记忆单元的输出受遗忘门的控制, 如果有不被遗忘的上一时刻的状态 s_c^{t-1} , 则接收该状态同时用于计算。因此, 记忆

单元的输出如式(11)所示^[18]:

$$s_c^t = b_c^t s_c^{t-1} + b_c^t g(a_c^t) \quad (11)$$

其中, g 为激活函数。最终,输出门的输入如式(12)所示^[18]:

$$a_o^t = \sum_{i=1}^I \omega_{oi} T_i^t + \sum_{h=1}^H \omega_{oh} b_h^{t-1} + \sum_{c=1}^C \omega_{oc} s_c^{t-1} \quad (12)$$

输出门对应的激活公式如式(13)所示^[18]:

$$b_o^t = f(a_o^t) \quad (13)$$

输出门的输出同时也是下一个记忆单元的输入。

为了减少过拟合问题,本文在 LSTM 层后加上了 Dropout 层用于正则化,其中 dropout 比率为 0.5。因为在训练中

LSTM 层记忆单元的个数影响最终分类的结果,所以本文针对该层的记忆单元个数进行尝试,选取训练之后船舶类型识别误差最小的值作为该层的节点数。其中经验公式^[18]为:

$$M = \sqrt{(m+n)} + x \quad (14)$$

其中, m 和 n 分别为输入层和输出层的节点数, x 为 $[0,10]$ 内的常数。

(3)全连接层:在最后增加一层全连接层用于分类,激活函数为 softmax,输出当前时序特征向量为 5 种船舶类型的概率。最终,LSTM 网络输出的船舶分类结果如式(6)所示。

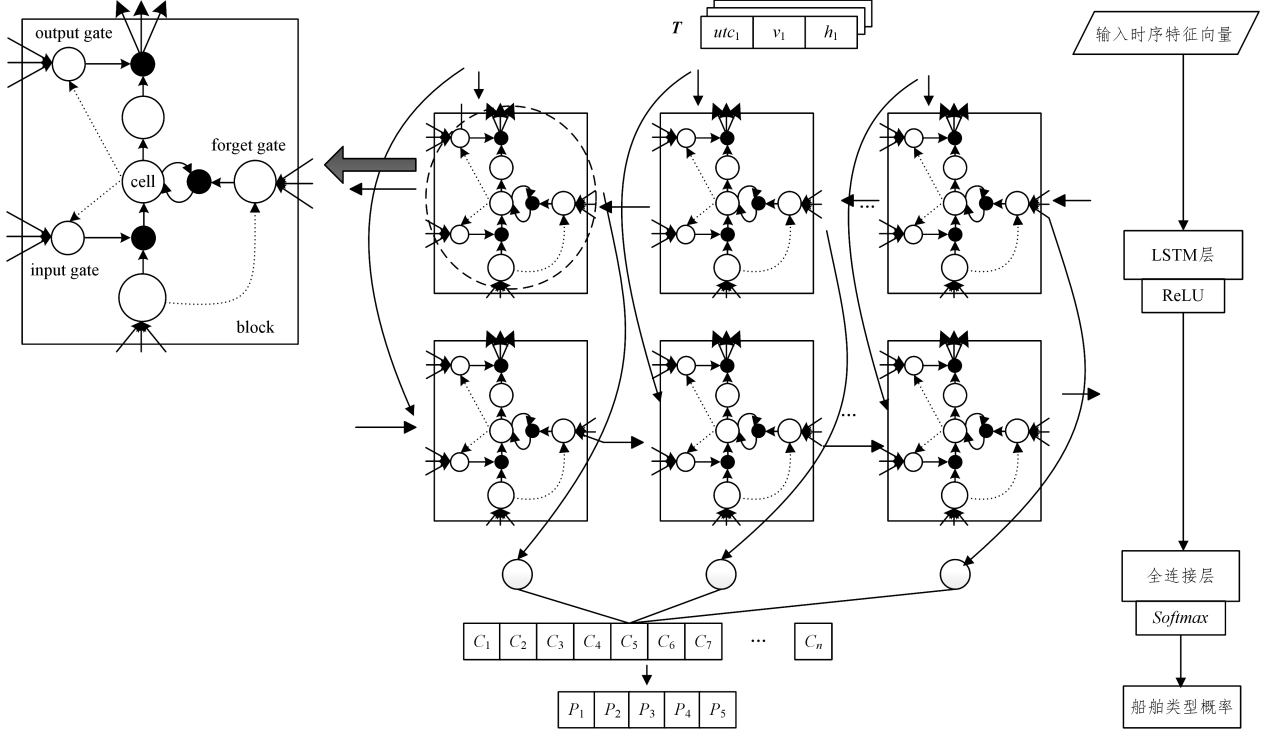


图6 LSTM的设计流程图

Fig. 6 Design flow chart of LSTM

4.2.3 基于1DCNN-LSTM的网络结构设计和数据融合

本文使用1DCNN-LSTM分类模型,具体网络结构如图7所示。网络的输入分为两部分:分布特征向量和时序特征向量。这两类特征向量分别输入1DCNN和LSTM中,经过训

练后分别得到对应的输出结果,然后使用基于权重分配的方式融合两个神经网络的输出,得到识别出的船舶类型结果向量。数据融合指对输入数据是分布特征向量和时序特征向量的两个模型进行分类结果进行加权求和,得到最终的分类结果^[14]。

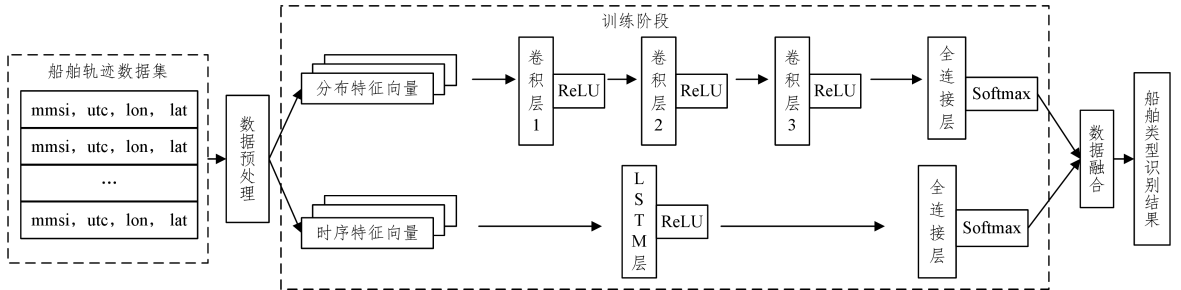


图7 基于1DCNN-LSTM的网络结构设计

Fig. 7 Design of network structure based on 1DCNN-LSTM

本文在每个网络映射出船舶类型的向量之后增加权重,通过将权重与输出向量进行矩阵点积,来将两个网络的结果进行相加,取求得的向量中概率最大的类型为最终结果。假设分布特征识别网络输出的船舶类型概率向量为 $\mathbf{P}_F$, 时序识别网络输出的概率向量为 $\mathbf{P}_S$, 且均为一个 1×5 的向量, 可分

别表示为 $(\mathbf{P}_{f1}, \mathbf{P}_{f2}, \mathbf{P}_{f3}, \mathbf{P}_{f4}, \mathbf{P}_{f5}), (\mathbf{P}_{s1}, \mathbf{P}_{s2}, \mathbf{P}_{s3}, \mathbf{P}_{s4}, \mathbf{P}_{s5})$, 其中每个向量 $\mathbf{P}_{ij}$ ($i = f, s, j = 1, 2, 3, 4, 5$, 分别表示渔船、客船、油船、集装箱船和散货船) 表示每个船舶类型的概率。则本文将两种分类网络的结果与权重向量矩阵点积后, 得到最终结果, 如式(15)所示:

$$P = P_F \cdot W_F + P_S \cdot W_S \quad (15)$$

其中, W_F 和 W_S 分别表示输出船舶类型向量对应的权重, 分别为 $(W_{f1}, W_{f2}, W_{f3}, W_{f4}, W_{f5})$ 和 $(W_{s1}, W_{s2}, W_{s3}, W_{s4}, W_{s5})$; “ $\cdot$ ”表示矩阵点积, 即矩阵中对应位置元素相乘, 且每种类型两个权重值相加为 1。根据前文所述, 油船、集装箱船和散货船更依赖于时序特征的识别结果, 而分布特征的识别结果在此基础上辅助识别, 因此 W_{s3}, W_{s4}, W_{s5} 的值分别比 W_{f3}, W_{f4}, W_{f5} 的值大, 而渔船和客船更依赖于分布特征向量的识别结果, 因此 W_{f1}, W_{f2} 的值分别比 W_{s1}, W_{s2} 的值大。

5 实验分析

5.1 实验环境和数据集

实验对全部数据进行预处理。提取船舶轨迹特征向量的大数据运行环境的 Hadoop 集群版本为 2.6.0。该集群包括 17 个节点、1 个 master 节点和 16 个 slave 节点, 内存为 640 GB。对船舶轨迹进行分类的 1DCNN-LSTM 模型使用 Windows 7 操作系统, 以 python3.6 编程语言实现, 使用的学习框架为 Keras, 后端为 Tensorflow, 模型在单台物理机运行, 内存为 16 GB。使用的轨迹数据集为 AIS 采集到的 2016 年 6 月份渤海湾包括渔船、客船、油船、集装箱船和散货船等船舶类型在内的真实船舶轨迹数据, 共 252191708 条。从中提取 5 类船舶轨迹点, 每条数据包含该船舶在采集时刻的轨迹数据信息。

分布特征向量以时间段为单位进行划分, 经过预处理后包括渔船、客船、油船、集装箱船和散货船在内总共有 154958 个时间段, 其中渔船 35143 段、客船 30453 段、油船 28750 段、集装箱船 30476 段、散货船 30136 段, 即总共有分布特征向量数据集 154958 条。时序特征向量按照时间点划分, 本次实验采用预处理后的时序特征向量, 共有 154958 个轨迹段。在实验阶段, 将所有数据混合打乱, 选取全部数据的 70% 为训练集、全部数据的 30% 为测试集对分类模型进行训练。

5.2 实验设置

5.2.1 评价标准

本文首先对采集到的 2016 年 6 月份中国渤海湾区域船舶数据进行预处理和特征向量的提取, 采用 MapReduce 并行化计算框架实现, 数据全部处理完成所用时间较单机处理有所提升。

本文采用分类准确率 (Accuracy)、精确率 (Precision)、召回率 (Recall) 和 F_1 值来评价模型。其中, 准确率为分类正确的样本数占样本总数的比例; 精确率为被正确分类的样本数与被分类到样本的总数的比例, 用于衡量分类的正确程度; 召回率为被正确分类的样本数与应当被分类到样本的总数的比例, 用于衡量实验结果的完整程度; F_1 值为召回率和精确率的调和平均值。

因为模型不会严格输出标准的单位向量表示识别的船舶类型, 而是输出每个元素都接近 0 或 1 的一个向量, 且元素值越高表示模型认为此方式的可能性最大, 所以采用以下公式来衡量模型输出的准确率:

$$acc = \frac{\sum_{i=0}^n l}{N} \quad (16)$$

$$l = \begin{cases} 0, & \arg \max(X_{i1}) \neq \arg \max(X_{i2}) \\ 1, & \arg \max(X_{i1}) = \arg \max(X_{i2}) \end{cases} \quad (17)$$

其中, acc 表示 Accuracy, N 为测试集中向量的总个数。 X_{i1} 表示测试集的船舶类型向量, X_{i2} 表示经过模型识别后输出的船舶类型向量, 函数 $\arg \max$ 表示输出向量中概率最大值对应的船舶类型所在的位置。

精确率、召回率、 F_1 值的计算公式如下:

$$Precision = \frac{correct}{Num_{pre}} \quad (18)$$

$$Recall = \frac{correct}{Num_{true}} \quad (19)$$

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \quad (20)$$

其中, $correct$ 表示真实类别为该船舶类型同时识别结果也为该船舶类型的个数, Num_{pre} 表示识别为该类船舶的总数, Num_{true} 表示该类船舶的总数。

5.2.2 网络参数选择

实验构建 1DCNN-LSTM 的混合模型来实现船舶类型的识别。本实验中, 每次模型训练的批次 (batch size) 大小为 128, 选择 Adam 作为优化器 (optimizer), 损失函数使用分类交叉熵 (catagorical_crossentropy), 采用早停止准则 (early stopping criterion) 设定训练次数 (epoch), 当准确率没有提升时停止训练, 并设置首次出现最终准确率的训练次数为最大训练次数 (max_Epochs)。首先, 本实验分别对 1DCNN 和 LSTM 两个网络模型进行参数的选择。

(1) 1DCNN 参数的选择

本文中基于 keras 的 1DCNN 需要优化的参数为卷积核大小。在本实验中, 采用三层卷积层, 每一层的卷积核大小相等, 并对卷积核大小进行实验。卷积核数量依靠经验尝试确定。图 8 给出了在通道数量为 48 的条件下, 针对不同卷积核大小得到的模型准确率。

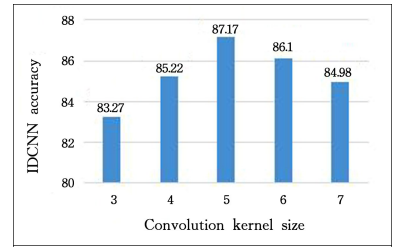


图 8 基于不同卷积核大小的模型准确率

Fig. 8 Model accuracy based on different convolution kernel sizes

从图 8 可以得出, 当通道数量不变时, 随着卷积核大小的增加, 网络的识别能力在总体上呈下降趋势。因为当卷积核增大时, 导致网络中的权重数目过多, 训练网络花费的时间也会增加, 这与文献[16]中关于使用小卷积核代替大卷积核的描述一致。当卷积核大小为 5 时, 1DCNN 的识别能力最佳, 因此选择卷积核尺寸大小为 5。

(2) LSTM 参数的选择

对于 LSTM 模型, 主要参数为时间步 (timestep) 和记忆单元的个数。LSTM 采用时序特征向量作为输入, 维度的大小依赖于 timestep, 即输入多少时刻的时序特征向量。记忆单元个数主要由式 (14) 决定。本实验选择时间步长为 6, LSTM 层记忆单元数目为 13, 输出层节点数目为 5 (5 种船舶

类型)。LSTM 层后设置 dropout 比率为 0.5,模型效果最好,最终达到 84.36%的准确率。

5.2.3 对比方法

LSTM 模型是近年来处理时间序列数据的主流神经网络模型^[17],船舶轨迹数据是时间序列数据。因此,为测试 1DCNN-LSTM 模型的效果,本文将单独使用基于 LSTM 的分类模型与其进行对比实验。在对比实验中,提取每一时刻的航速、加速度、航首向和曲率组成时序特征向量,并将其作为对比模型 LSTM 的输入向量。其余训练和测试流程不变。

5.3 实验结果

本文首先通过 1DCNN,使用上述的网络参数对经过预处理和分布特征向量提取后的数据集进行训练,并基于早停止准则,当训练次数为 13 时,模型最终准确率达到最高,具体结果如表 2 所列。

表 2 1DCNN 船舶类型的识别结果

Table 2 Identification results of ship type on 1DCNN

Real class	Forecast class				
	Fish	Passenger	Tanker	Container	Bulk-cargo
Fish	9 889	51	99	231	230
Passenger	58	8 757	240	43	107
Tanker	278	179	7 372	245	404
Container	367	291	598	7 387	400
Bulk-cargo	64	154	1 493	377	6 757
Precision/%	92.80	92.84	75.2	89.2	85.6
Recall/%	94.2	95.1	87.0	81.7	76.4

从表 2 可以得出以下几点结论:1)分类效果在渔船上表现效果最好,因为与货船、油船、客船相比,渔船的运行状态能更好地区分出来,例如,在捕鱼过程中,渔船的轨迹通常是曲折蜿蜒的且具有较低的速度,接近于 0;渔船从一个捕鱼区域到另一个捕鱼区域的航行过程中,具有较高的速度^[19]。2)客船的精确率相对较高,客船从一个港口航行至另一个港口的过程中基本呈匀速行驶状态,航速偏高,航行过程中一般不会停泊,因此对客船的分类效果相比油船、集装箱船和散货船较好,其召回率比渔船高。3)由于油船、集装箱船和散货船 3 类船舶都属于货船,因此航行轨迹比较相似,故该方法在区分这 3 类船舶的效果时没有渔船和客船的效果显著,精确率和召回率普遍偏低,网络对船舶轨迹的识别仍然受到特征相似性的困扰,还需要进一步的区分。4)使用式(16)可以得出模型准确率为 87.2%。

下面将时序特征向量输入到 LSTM 模型中,与 1DCNN 模型同时进行特征学习。因为最终的结果是 1DCNN 与 LSTM 两种分类模型的结果与权重矩阵点乘的融合,所以权重的选择对结果也会产生一定的影响。本文以 1DCNN 网络为主, W_F 中每个元素的取值范围为[0,1],最终得到每种船舶类型的准确率如图 9 所示。

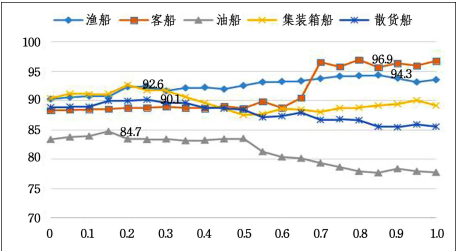


图 9 基于不同权重的 5 种船舶类型准确率

Fig. 9 Accuracy of five ship types based on different weights

由图 9 可知,对于渔船和客船,当 1DCNN 输出结果的权重大于 LSTM 输出结果的权重时,准确率更好;相反,对于油船、集装箱船和散货船而言,LSTM 的输出结果权重更高时,结果更佳。因此,本实验最终选择的权重矩阵 W_F 为(0.85, 0.8, 0.15, 0.2, 0.25), W_S 为(0.15, 0.2, 0.85, 0.8, 0.75)。此时,模型准确率达到最高,为 91.7%,5 种类型船舶各自的 precision 和 recall 如表 3 所列。

由表 3 可知,增加了 LSTM 模型对时序特征向量的学习后,可以得到以下结论:1)召回率方面,客船最高,而且集装箱船、油船和散货船也有了明显提升,比仅使用 1DCNN 进行分布特征的学习结果要好,验证了这 3 种类型的船舶轨迹在时序变化上的差异较分布特征来说更加明显,但不及渔船和客船。考虑到客船的轨迹特征最为平稳,变化最少;同时,在速度方面,渔船在捕鱼时段和非工作时段能产生明显的区分,因此网络对货船的识别仍然受到轨迹特征不够明显而导致的分类效果不好的困扰。2)精确率方面,渔船和客船的识别结果最为准确,集装箱船的精确率更好,然而模型易将油船和散货船混淆。3)因为不同船舶类型在精确率和召回率上的趋势变化不一致,例如集装箱船和散货船的精确率高而召回率低,其他 3 类船舶相反,所以采用了 F_1 值作为另一个评价标准对模型进行衡量。其中,客船的 F_1 值最高,为 96.6%,且与渔船和集装箱船的 F_1 值都超过了 90%,油船的 F_1 值最低,但也有 87.3%。4)根据式(16)可以得出模型的准确率为 91.7%,说明 1DCNN 和 LSTM 的混合模型能够对 5 种船舶轨迹进行有效的类型识别。

表 3 1DCNN-LSTM 船舶类型的识别结果

Table 3 Identification results of ship type based on 1DCNN-LSTM

Real class	Forecast class				
	Fish	Passenger	Tanker	Container	Bulk-cargo
Fish	10 089	42	57	157	198
Passenger	38	8 841	129	32	96
Tanker	258	87	7 772	191	317
Container	289	167	349	8 087	251
Bulk-cargo	20	34	867	263	7 857
Precision/%	94.3	96.4	84.7	92.6	90.1
Recall/%	95.7	96.8	90.1	88.5	86.9
F_1 /%	94.9	96.6	87.3	90.5	88.5

综上所述,1DCNN 和 LSTM 相结合对船舶轨迹进行类型识别,虽然对油船、集装箱船和散货船的识别效果不如渔船和客船,但是较仅使用 1DCNN 对分布特征进行学习,效果有了明显的提高。

5.4 对比实验

对比实验采用 LSTM 分类模型,时间步大小为 6,每条时序特征向量是长度为 5 的向量,包括时间戳、航速、加速度、航首向和曲率 5 个元素,LSTM 层经过多次实验后采用记忆单元数目为 15。LSTM 层后同样设置 dropout 为 0.5。对比实验采用准确率对模型进行评价,实验结果如表 4 所列。

表 4 不同模型的对比结果

Table 4 Comparison results of different models

Model	Accuracy/%
LSTM	83.3
1DCNN-LSTM	91.7
1DCNN	87.2

根据表 4 中 LSTM 模型和 1DCNN-LSTM 模型的准确率

对比结果可知,只考虑到船舶轨迹的时序信息对于船舶类型的识别还不够,需要考虑船舶轨迹信息的分布情况,因此从分布和时序两个方面对船舶轨迹进行分析,可以达到更好的识别效果。使用 1DCNN 模型单独考虑船舶轨迹的分布情况虽然可以达到不错的效果,但仍有欠缺。将二者结合最终达到的识别准确率比 LSTM 模型的准确率提高了 8%,比 1DCNN 模型的准确率提高了将近 5%。

综上,结合船舶轨迹的时序特征和分布特征对船舶轨迹类型进行识别,较只用单一类型的特征进行识别时准确率有了一定的提高。对于船舶轨迹数据而言,不同船舶类型的时序特征和分布特征的重要性也有一定的差距。但是二者均为不可分割的一部分,本文根据不同船舶类型中不同特征的重要性设定不同的权重,将结果融合,最终达到了更好的分类效果。

结束语 本文针对船舶轨迹类型识别任务提出了一种基于 1DCNN 和 LSTM 的混合网络模型。首先对数据进行预处理,然后基于处理后的数据使用文中提出的基于 MapReduce 框架的分布特征向量提取算法和时序特征提取算法对船舶轨迹数据进行向量提取,与单机版程序相比,效率有所提升。最后使用 1DCNN 代替传统二维 CNN,通过分布特征向量来提取船舶整体特征,使用 LSTM 通过时序特征向量来提取船舶时序特征,并将二者的识别结果融合,对船舶轨迹数据进行分类。实验结果表明,本文提出的模型有效解决了传统二维 CNN 对数据特征的压缩问题和时序特征表达能力匮乏的问题。与目前主要用于时序数据分类的主流模型如 LSTM 模型相比,本文模型对船舶轨迹分类的能力有了明显提高,可以实现船舶类型的有效分类。

然而,从实验结果可以发现,同属于货船类的油船、集装箱船和散货船的轨迹特征没有渔船和客船明显,轨迹分类的结果仍有不足,因此未来工作将会集中在以下方面:首先对油船、集装箱船和散货船这 3 种船舶做进一步的特征分析;然后对目前最新的大规模数据的分类方法进行研究,并在此基础上对本文提出的轨迹分类模型做进一步的改进,提高分类的准确率,并实现对渔船、集装箱船和散货船更准确的分类。

参 考 文 献

[1] CHEN J H, LU F, PENG G J. Research progress on trajectory analysis of marine transportation vessels [J]. China Navigation, 2012, 35(3): 53-57.

[2] XU Y P. Research on the Trajectory Analysis of Marine Transportation Ships [J]. China Water Transport (2nd Half), 2017, 17(2): 16-17.

[3] XIAO W, SHAO Z P, PAN J C, et al. Ship trajectory clustering model based on AIS information and its application [J]. China Navigation, 2015, 38(2): 82-86.

[4] CAO W Q, LI Z X, WEI Q, et al. Trajectory Classification Method Based on Probability Density Estimation of Regional Distribution [J]. Computer Engineering, 2018, 44(4): 262-267, 286.

[5] LIU L, CHU X M, JIANG Z L, et al. Ship trajectory classification algorithm based on KNN [J]. Journal of Dalian Maritime University, 2018, 44(3): 15-21.

[6] ZHAO X Q, ZHANG L. SVM-based high-dimensional unba-

lanced data set classification algorithm[J]. Journal of Nanjing University(Natural Science), 2018, 54(2): 452-461.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[8] ENDO Y, TODA H, NISHIDAK, et al. Deep feature extraction from trajectories for transportation mode estimation[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham, 2016: 54-66.

[9] LJUNGGREN H. Using Deep Learning for Classifying Ship Trajectories[C]// 2018 21st International Conference on Information Fusion (FUSION). Cambridge, 2018: 2158-2164.

[10] ZHANG G H, LIU B. Research on time series classification using CNN and bidirectional GRU [J]. Journal of Frontiers of Computer Science and Technology, 2019, 13(6): 916-927.

[11] ZHENG Z T, ZHAO Z F, WANG G L, et al. Method for ship trajectory extraction for port stop area identification [J]. Journal of Computer Applications, 2019, 39(1): 113-117.

[12] ZHENG Y, LIU L, WANG L, 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. ACM, 2008: 247-256.

[13] ZHU J, JIANG N, HU B. Application of Multiple Motion Parameters of Moving Objects in Trajectory Classification [J]. Journal of Earth Sciences, 2016, 18(2): 143-150.

[14] YE S N. Traffic pattern recognition based on deep convolutional network [D]. Chengdu: Southwest Jiaotong University, 2018.

[15] WANG J J. The safety of large tankers through the new, Ma Strait [J]. China Navigation, 2006 (3): 27-30.

[16] SSEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 2818-2826.

[17] ZHANG L. Research on LSTM-based smartphone trajectory recognition [D]. Lanzhou: Lanzhou University of Technology, 2018.

[18] QUAN B, YANG B C, HU K Q, et al. Ship track prediction model based on LSTM [J]. Computer Science, 2008, 45 (S2): 126-131.

[19] GENG J L. Research on big data storage optimization and behavior recognition technology of fishing vessel trajectory [D]. Hangzhou: Hangzhou University of Electronic Technology, 2018.



CUI Tong-tong, born in 1993, postgraduate. Her main research interests include data processing technology and software services.



WANG Gui-ling, born in 1978, Ph.D. associate professor, is a senior member of China Computer Federation. Her main research interests include large-scale streaming data integration and analysis, and service computing.