海洋大数据关键技术及在灾害天气下船舶行为预测上的应用

王冬海，卢峰，方晓蓉，郭刚
中电科海洋信息技术研究院有限公司，北京 100041

摘要
随着海洋数据量的爆炸式增长，海洋大数据受到越来越多的关注。主要分析和总结了当前海洋大数据的研究现状和关键技术，聚焦了机器学习在海洋大数据中的模型预测研究的实例，对海上船舶在灾害天气（台风）下的行为进行了回归训练和预测。通过构建和对比决策树、Bagging、随机森林等多种机器学习算法，对样本数据进行学习、预测和检验评估。最终结果表明，随机森林方法在灾害天气下船舶密度的预测应用中具有良好和稳健的效果。

关键词
海洋大数据；机器学习；船舶行为预测

中图分类号：TP181　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2017044

Ocean big data and applications in ship behavior prediction under disaster weather

WANG Donghai, LU Feng, FANG Xiaorong, GUO Gang
CETC Ocean Information Co., Ltd., Beijing 100041, China

Abstract
With the explosive growth of marine data, the ocean big data have received more attention and concern recently. The current status and key technologies of ocean big data here were summarized and analyzed. A specific case about the application of machine learning in the prediction model of ocean big data was also focused, which was a forecasting test of maritime ships behavior based on regression training in disaster weather (typhoon). The sample data for validating and evaluating three machine learning algorithms of decision tree, Bagging and random forest were trained and tested. The final results prove the best and robust effect of the random forest algorithm in the prediction of ship density under the disaster weather.

Key words
ocean big data, machine learning, ship behavior prediction

1 引言

在经济全球化的今天，全球90%的贸易都经过海洋，全球70%的经济活动都发生在沿海地区，沿海地区海洋经济发展已经成为带动我国国民经济增长的重要因素。随着信息技术的快速发展和国家海洋战略的实施，与海洋相关的科学观测/监测与数值计算、海洋经济和管理等数据日益增多，与海洋相关的音频、视频、文字和图片等数据大量涌现，数据存储量、规模、种类飞速增长，海洋大数据正成为大数据领域的重要应用之一。

海洋大数据作为全球大数据的重要组成部分，是实现海洋信息行业智能化管理和“互联网+”的基础和前提，也是实现我国“海洋强国”战略的重要支撑与保障。随着我国“空天地海潜”一体化立体监测技术的发展和数字海洋建设的全面深入，海洋信息化已经逐步从数字海洋向智慧海洋发展，海洋数据在数量、增长速度、种类扩展3个方面都有了飞跃式的进展，海洋数据蕴含的价值也越来越高。

同时，海洋大数据还面临着一些挑战：海洋相关数据体量巨大、类型多样、数据利用率较低、处理算法过于简单、远海海域数据获取不足等问题，难以满足海洋信息服务的需求。迫切需要发展海洋大数据及其应用技术，充分挖掘海洋数据价值，全面提升资源保护与开发、环境预警与预报、应急与救助、安全管控等领域的智能化、精细化能力，为实现“海洋强国”“一带一路”国家战略提供信息技术支撑。

本文针对海洋大数据技术现状，围绕国家海洋发展战略在海洋安全建设、智慧海洋建设等方面的关键技术研究与工程应用，介绍了海洋大数据研究的关键技术及海洋大数据在灾害天气下辅助决策方面的初步应用。

2 海洋大数据关键技术

海洋大数据应用技术平台基于云计算架构，搭建包括数据汇集、数据存储和数据处理的大数据业务处理系统以及运维管控、安全保障、标准规范3个支撑体系，如**图1**所示。

2.1 海洋多源信息感知探测技术

构建覆盖空、天、海、岸、潜的一体化数据采集信息网络，获取来自天基信息系统（卫星）、无人机信息系统、岸基雷达和观测站、船载探测平台、浮标、水下观测信息系统（水下滑翔机、水下潜器和海底观测网等）多源观测信息，实现海洋的全天时、全天候环境与目标观测，通过海上综合通信传输网络，对感知网络进行集成连接，形成一体化综合信息网络，获取卫星遥感影像数据、航空影像遥感数据、沿海台站观测数据、岸基雷达观测数据、海洋浮标观测数据、调查船走航断面的观测数据、海底潜标平台数据等海洋观测/监测数据以及渔业经济数据、渔业捕捞数据、渔业管理数据、海洋旅游数据、航运交通数据、海上贸易数据、全球海关数据等海洋行业数据，达到对海域安全态势、环境信息、海域资源、目标活动的全面掌控。

2.2 海洋大数据处理平台技术

海洋大数据平台基于云计算架构，解决海量数据的分布式存储、管理和分析等大数据业务，改变海洋信息资源使用的无序状态。突破海量数据存储及高效管理，

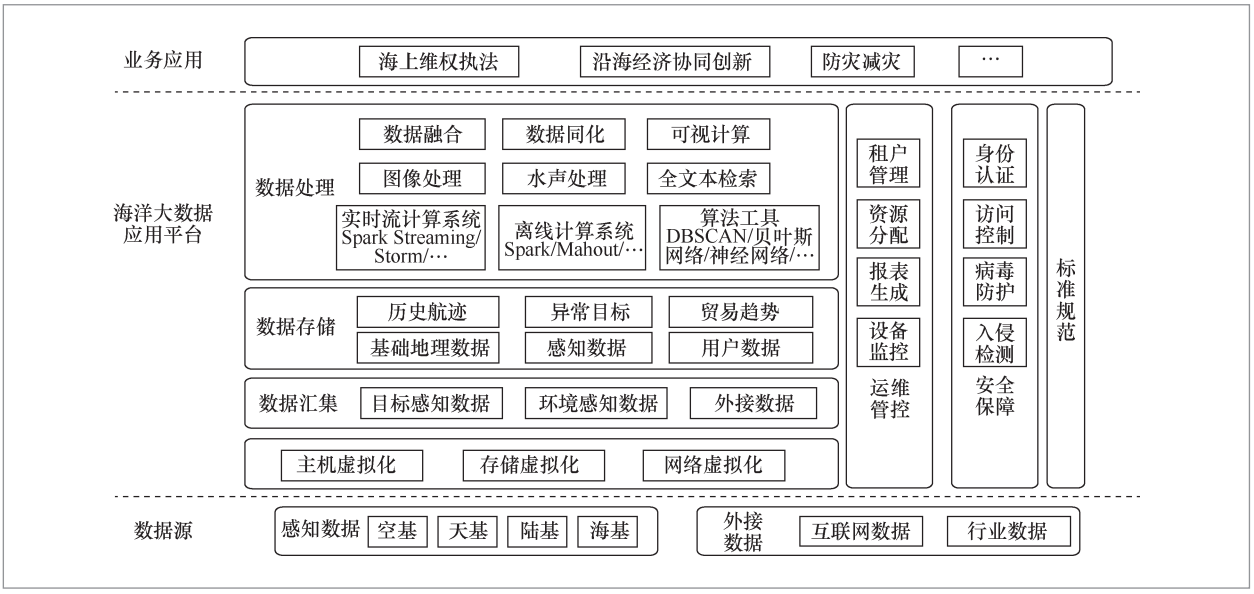

图1 海洋大数据应用技术平台架构

重点解决各类涉海信息自成体系、数据格式不统一、数据量和采样频率差异大等问题，构建统一数据提取接口，制定信息技术标准和数据转换规范，建立多源大数据存储及管理系统。数据库采用分布式非结构化数据库——HBase，数据统一采用基于Hadoop分布式文件系统（Hadoop distributed file system，HDFS）进行存储。针对海量数据的分布式存储及离线快速分析处理，采用包括实时性处理能力强的Spark计算框架以及适用于超大规模作业离线处理的基于map/reduce并行编程模型的Hadoop计算框架，对海量涉海数据进行批量自动转换，最终实现海洋数据从存储、管理到数据清洗、融合、挖掘、显示的大数据平台构建[1]。

2.3 海洋多维重建与可视技术

海洋环境要素多维重建与可视计算是在基于地球球体模型的三维可视化基础平台上，对海底、水体、海面和海岸的各种海洋自然要素以及海洋自然现象进行可视化表达、再现或预现。综合运用增强现实等技术实现海洋要素、自然要素、海上设施、目标要素等的三维可视化表达。将计算机生成的海面及海岸等虚拟图形叠加在用户看到的一个现实海岸及海面场景上，从而代替虚拟现实中完全由计算机虚拟生成的世界。海洋要素数据可视化通过海洋数值模拟，实现对海水温度、盐度、海表面高度异常、海流、密度、声、光、电、磁等参数的三维动态再现。海洋自然要素通常采用场模型来表达，实现对泥沙沉积、矿产等海底地质、地形地貌、矿产资源、海底电缆管道和毗邻区、专属经济区及大陆架区域的大陆坡线、海槽等自然要素的可视化表达。目标要素包括出现在水面及水下的船舶、无人潜航器、蛙人等目标。将不同参数的海洋状态数据叠加展示在二维、三维海洋地理信息系统（geographic information system，GIS）平台之上，实现对海洋基础数据、海洋目标数据、海洋环境数据以及衍生数据（海洋同化数据、海洋遥感反演数据、数值分析输出数据等）的管理、集成、分析以及可视化表达等功能，为研究

海洋系统的结构与功能、揭示并认识海洋现象的各种规律等活动提供通用、易用、规范的工具。

2.4 海洋大数据关联与挖掘技术

针对海上分布式多源异构性传感器间目标关联问题，利用多特征融合的目标关联方法，通过分析雷达、船舶自动识别系统（automatic identification system，AIS）、广播式自动相关监视（automatic dependent surveillance-broadcast，ADS-B）系统、电磁、光电等多传感器之间观测上提取的共有特征，计算目标间通过特征信息融合成的关联测度，形成关联判决依据，并在关联决策上采用基于有效特征数累积的全局最优关联算法，对直接的关联依据决策判决进行修正，提供海洋情报的关联挖掘和辅助决策[2]。通过采集海洋气象、海浪、洋流、海洋资源、海洋灾害等海洋环境信息以及AIS、ADS-B、雷达、光电等手段感知目标信息，再结合航运交通信息、海上贸易信息、地理信息、市场信息等，采用序列建模、聚类等无监督方法以及决策树（decision tree）、随机森林（random forest）、支持向量机、神经网络、贝叶斯等有监督方法的机器学习预测分析，得到相关关系与基本规律，预测未来的变化趋势[3]，为海洋资源利用、航运、渔业、旅游等各项海洋活动提供信息服务支撑。

3 海洋大数据在灾害天气下船舶行为预测上的应用

利用机器学习对海洋关联事件进行预测是海洋大数据应用的一个重要方向。采用机器学习中的决策树、Bagging、随机森林等算法，对海上船舶在灾害天气（台风）情况下的行为进行了预测。针对机器学习在多源异构海洋大数据的预处理、特征工程、特征选择、模型训练、模型评估等算法流程进行了介绍。

3.1 灾害天气下的船舶行为预测

海上船舶在灾害天气下需要随时掌握天气变化情况，并在台风、海啸等极端天气来临之前及时做出到就近港口避难等行为反馈。然而不同海域的船舶在何时做出何种避难行为往往受到船长的主观因素影响较大。船舶在灾害天气下的行为模式是否存在显著特征，能否得到合理的预测，该问题的解决对于灾害天气下港口应急调度与高效管理具有重要意义，可通过台风路径的预测信息精确预测船舶的行为，从而减轻灾害天气对航运业的经济损失。近年来随着大数据技术的发展，机器学习的强大学习和智能化应用在各行各业逐渐火热和成熟。机器学习主要研究计算机模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，目前已经成为多源异构大数据挖掘和处理的重要科学工具。

本文通过船舶行为与异常天气的回放来构建极端天气条件与船舶密度变化的算法预测模型，根据对大量样本的学习、预报和检验，得到灾害天气情况下的船舶行为预测，为海上防灾预警、港口泊位管理与指挥调度等应用提供信息支撑。

3.2 多源数据采集

本文主要采用中国气象局台风最佳路径数据集[4]、美国国家环境预报中心（National Centers For Environmental Prediction，NCEP）全球数值环境再分析

场[5]和全球船舶自动识别系统数据来进行分析训练研究。台风路径数据由中国气象局热带气旋资料中心提供，该中心网站提供了1949年以来西北太平洋海域热带气旋每6 h的最佳路径数据集①。该数据集参数主要包括台风路径经纬度坐标、时间、强度等级等。同时，还获取了同步的三维NCEP再分析环境场数据，该数据由美国国家海洋和大气局（National Oceanic and Atmospheric Administration，NOAA）的国家环境预报中心②开发和提供。该中心每天定时发布前一天4次的同化再分析数据，分别为00:00、06:00、12:00和18:00，数据空间分辨率是2.5° ×2.5° 经纬网格，垂直方向26层（从地面到10 hPa）。该资料集分为大气等压面资料、地面（海表）资料、通量资料等。本文主要使用地面（海表）资料作为辅助分析。AIS资料[6]主要来自船舶上配备的船舶自动识别系统，通过连接船上全球定位系统（global positioning system，GPS）定位仪、测深仪、电罗经等设备，能够自动采集并发射船舶实时的静态信息和动态信息（船舶身份、船舶位置、吃水、航速、船舶艏向、船舶类型、船舶长度、宽度等），实时反映船舶航行状态和海上交通态势。本文采用AIS船舶静态信息和动态信息进行分析，全球AIS一年的数据量约为300多亿条。此外，由于船舶空间分布密度和距沿岸各港口的距离存在一定关系，所以这里还引入了全球16 831个船舶停靠点的坐标信息。该数据主要包含了港口的地理坐标、名称、所属国家等信息。

3.3 数据分析和处理方法

3.3.1 多源异构数据预处理

预处理主要针对需要预测的船舶分布密度进行各种数据的匹配、插值处理、质量控制等步骤。这里采用的数据特征呈现多源异构性，包括从1~3维的不同领域和特征信息的数据。需要针对计算船舶分布密度问题进行多源异构数据的预处理。最终获得一套时空匹配的多源异构融合数据集，为后面的训练和预测研究奠定基础。这里的船舶密度利用AIS数据进行网格化处理，然后针对每个网格的数据进行求和统计。

台风最佳路径数据采用文本格式保存，是混合数值和字符型信息保存的一维数组。首先从台风最佳路径数据选取过境南海海域的时段，针对这些台风时段的数据，采用线性插值方法将6 h一次的定位数据插值到1 h的时间分辨率。由于地理网格化的船舶密度可能和台风中心距离密切相关，所以这里还要利用地球坐标最近距离算法求解每个网格中心点和台风中心的绝对距离。NCEP再分析资料是采用气象上标准的网络通用数据格式（network common data form，NetCDF）存储的三维资料。由于时间分辨率不高，这里采用时间权重方法进行插值处理，计算式如下：

$$\begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_n \end{bmatrix} = w_1 \begin{bmatrix} P_1^{t1} \\ P_2^{t1} \\ \vdots \\ P_n^{t1} \end{bmatrix} + w_2 \begin{bmatrix} P_1^{t2} \\ P_2^{t2} \\ \vdots \\ P_n^{t2} \end{bmatrix} \quad (1)$$

这里P_1~P_n表示需要获得的第1~n个参数（主要包括气压、气温等），w_1和w_2表示每个时刻的再分析资料的时间权重，P_n^{t1}和P_n^{t2}表示前后两个时间对应的参数。最后，将经过时间插值的三维数据插值到0.5° ×0.5°（50 km）水平分辨率进行匹配。经过特征分析结果表明，灾害天气下的气温、相对湿度等参数的变化特征不太明显，与船舶行为的关联性不大，而风

① http://tcdata.typhoon.gov.cn

② https://www.esrl.noaa.gov

场、气压和降水在灾害天气下有显著的变化响应，可以作为灾害天气（台风）的表征参数。另外，从本算例可以看出，能够影响船舶航行行为的特殊天气情况主要为台风、风暴潮（海啸）等极端天气情况。一般的天气情况对船舶航行行为影响不显著。在开展气象环境对船舶行为影响分析时，可以重点以台风、风暴潮等灾害天气情况为主要数据源，以风场、气压、降雨等数据为辅助数据进行分析。通过相关性分析进行变量筛选（过程图片太多，考虑篇幅在此省去），选取与台风最佳路径最相关的气象数据（风场、气压、降雨），删除与台风路径相关性较小的气象数据（气温、湿度）。由于以逗号分隔值（comma separated value，CSV）格式存储的AIS数据受到信息传输、错误解码等因素的影响，无法避免地会存在错误信息，因此需要对AIS数据进行清洗和插值补充，从而提高AIS数据的可用性和可靠性。这里选取106°~115° E，10.5°~20.5° N范围，按照小时分辨率对AIS全年数据进行0.5°×0.5°网格上的分布密度计算，得到需要特征库数据集。最后，基于AIS网格数据，对全球船舶停靠点进行研究区域内的快速自动筛选，确定118个停靠点及相对每个船舶密度空间网格的距离因子。

在参数选择过程中，根据一般经验、特征重要性排序和模型预测的误差结果反馈对特征参数做了筛选（删除特征重要性较低的参数）。最终选择的特征参数包括：网格距最近港口距离（*distance*）、每天时刻（*ta*，取00:00~23:00的整点）、网格距台风中心距离（*typhoon_distance*）、台风中心经度（*typhoon_lon*）、台风中心纬度（*typhoon_lat*）、台风年龄（*ddt*）、NCEP海面降雨场（*rain*）、NCEP海面风场（*wind*）、NCEP海面气压场（*pressure*）、网格船舶密度（*density*），共10个参数。

- 网格距最近港口距离（*distance*）：由于交通流（AIS）与感兴趣点（point of interest，POI）有关，其中感兴趣点是指对交通流有明显影响的地点，选取港口作为POI。
- 每天时刻（*ta*）：白天和晚上船舶的行为活动存在差异，因此加入该特征。
- 网格距台风中心距离（*typhoon_distance*）：由于缺乏台风作用距离参数，因此用台风中心距网格距离来代替。
- 台风中心经度（*typhoon_lon*）、台风中心纬度（*typhoon_lat*）：台风位置影响船舶行为。
- 台风年龄（*ddt*）：台风生成到消亡存在时间周期，船舶行为与台风生成后的时间有关。
- NCEP海面降雨场（*rain*）、NCEP海面风场（*wind*）、NCEP海面气压场（*pressure*）：通过相关性分析选取与台风相关的气象参数降雨、风场、气压场。由于台风路径信息仅包含台风位置和强度信息，缺乏空间变化，因此在此加入了气象场数据。
- 网格船舶密度（*density*）：模型的预测因子。

其中模型输出为船舶密度，其余9个与气象、时间、POI相关的变量为模型输入。从特征库资料随机选取其中80%的数据作为训练集，其余20%的数据作为测试集。

3.3.2 机器学习训练模型选择

针对需要预测的问题，选择了3种主要的机器学习模型来训练前面预处理的多源异构数据集。模型包括决策树、Bagging和随机森林方法。除了以上3种模型外，还可以选择神经网络、支持向量机（support vector machine，SVM）、XGBoost等回归模型，本文暂不做详细探索。

决策树模型是一种树型结构(可以是二叉树或非二叉树),基于特征对实例进行分类或回归的过程。其每个非叶节点表示一个特征属性上的判定,每个分支代表这个特征属性在其值域上的输出,而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始,测试待分类项中相应的特征属性,并按照其值选择输出分支,直到到达叶子节点,将叶子节点存放的类别作为决策结果。

Bagging是bootstrap aggregation的简称,它是一种有放回的抽样方法。Bagging方法是多模型融合方法,它主要是为了解决单一分类器容易产生过拟合的问题。Bagging通过重复取样,相同训练的数据多了之后,能够减少结果的方差,可以理解为综合多个弱分类器的结果得到一个强分类器。

随机森林[7]是通过构建多个弱分类器,使得最终分类效果能够超过单个分类器的一种融合算法。随机森林可理解为由很多决策树组成的森林。随机意味着每棵树之间没有任何联系,都是独立的。它也是按照Bagging的方法重复取样,抽取的数量和样本总量相等。但是在训练树的时候并不是把所有特征都用上。假设总共有M个特征。每次训练一棵树的时候,随机抽取其中的$m(m<<M)$个特征进行训练。随机森林中的树不需要进行剪枝操作。因为样本的抽取、特征的抽取已经保证了随机性,大大减少了过拟合的可能性。

分类与预测模型对训练集进行预测得到的准确率并不能很好地反映预测模型未来的预测性能,为了能够有效地判断一个预测模型的性能表现,需要一组没有参与预测模型建立的数据集(测试集),并在该数据集上评价预测模型的准确率。将数据分为训练数据集、测试数据集,然后通过训练数据集进行训练,通过测试数据集进行测试。模型预测效果的评估方法采用均方误差(MSE)、均方根误差(RMSE)、平均绝对误差(MAE)、正则均方误差(NMSE)等。

3.4 结果分析

3.4.1 模型训练

图2显示了采用决策树、Bagging、随机森林模型对特征库样本的训练结果,横坐标为训练集的船舶密度(可理解为真值),纵坐标为模型预测的船舶密度(预测值)。

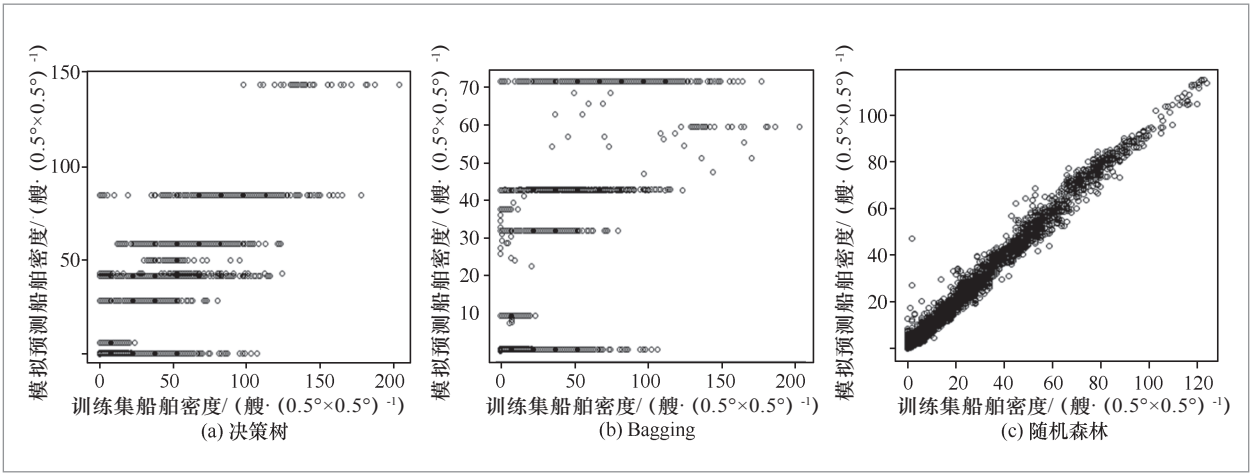

图 2 决策树、Bagging、随机森林模型训练结果

可以看出随机森林模型的预测值与真值几乎为一条直线，模拟结果远远优于决策树和Bagging方法。说明随机森林模型能够很好地预测台风天气下船舶的密度变化。

3.4.2 误差分析

采用均方误差、均方根误差、平均绝对误差、正则均方误差4项指标进行模型的误差分析。模型训练集误差（见**表1**）和测试集误差（见**表2**）显示，随机森林模型的误差远远优于决策树和Bagging法的误差。

表3和**图3**显示了随机森林模型的特征重要性降序排序结果。随机森林对连续变量设置了两种重要性，一种是平均均方误差减少百分比（%IncMSE），另一种是平均节点不纯度下降量（IncNodePurity）。变量重要性排名第1位的是台风年龄（台风生成后的时间）；排名第2位的是每天的时刻，说明白天或夜晚船舶的行为响应不同；排名第3位的是网格距最近港口距离；排名第4位的是台风中心纬度；排名第5位的是台风中心经度；排名第6位的是气压场；排名第7位的是距台风中心距离。风场和降雨场的影响较小，其原因可能是，台风登陆带来大风强降雨之前，船舶已经进入避风港，并将持续停留，直到大气和降雨天气好转。另外，两种特征重要性定义不同导致其排序的结果也不同③。这是由于预测变量船舶密度是空间变化的，而某些特征因素是纯时间（如台风年龄），因此虽然在%IncMSE重要性上影响很大（加噪声后的误差），但由于缺乏空间分

③ http://www.paper.edu.cn/releasepaper/content/201507-212

表1 训练集误差分析

	决策树	Bagging	随机森林
NMSE	0.186 732 4	0.292 541 1	0.009 209 107
MSE	8.526 040 2	13.367 176 8	0.420 479 980
RMSE	2.919 938 4	3.654 747 2	0.648 444 277
MAE	0.928 237 8	1.173 239 1	0.183 384 508

表2 测试集误差分析

	决策树	Bagging	随机森林
NMSE	0.198 112 20	0.299 288 40	0.022 468 15
MSE	8.190 381 80	12.373 221 60	0.928 881 23
RMSE	2.861 884 30	3.517 559 00	0.963 784 85
MAE	0.910 828 80	1.172 814 90	0.297 336 24

表3 随机森林模型特征重要性排序

变量	符号	%IncMSE	IncNodePurity
台风年龄	*ddt*	99.988 95	238 107.37
每天时刻	*ta*	54.869 66	46 363.45
网格中心距岸距离	*distance*	51.286 49	2 091 026.23
某时刻台风中心纬度	*typhoon_lat*	47.441 66	166 999.35
某时刻台风中心经度	*typhoon _lon*	44.589 38	174 932.34
某时刻气压场	*pressure*	36.538 57	874 671.41
网格中心距台风距离	*typhoon _distance*	33.984 29	146 821.91
某时刻风场	*wind*	31.817 96	822 703.11
某时刻降雨场	*rain*	24.381 35	525 582.96

布信息，它们在IncNodePurity的重要性排序并不高。

以上结果说明，在台风等灾害天气下，船舶行为受到天气作用的影响十分显著。

由于影响船舶航行的水文气象因素还有海浪、海冰、海流、海雾等[8]，未来可以考虑在特征数据库中加入海浪、海雾等海洋环境数据，进一步提高模型预测精度。另外，由于在台风作用半径以外，对船舶行为影响较小，因此，应当加入台风作用半径的参数来修正各网格点距离台风中心距离的参数。最后，还应当考虑加入K层交叉验证（K-fold cross-validation），将K个模型在K个测试集上的准确率（NMSE/RMSE）的平均值作为模型的综合性能评价指标，从而减少由于抽样不均匀导致的训练集和测试集的误差变化。

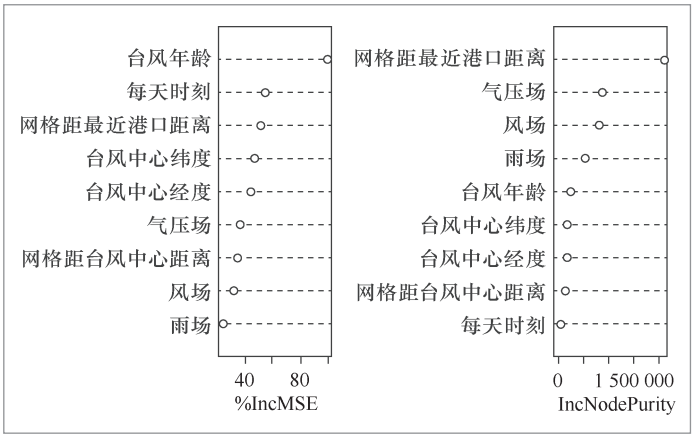

图 3 随机森林模型特征重要性排序

4 结束语

本文介绍了海洋大数据的特点与发展现状，分析了海洋大数据行业的数据来源与特点，介绍了海洋大数据的关键技术，并使用机器学习中的决策树、Bagging、随机森林模型开展了海上船舶密度分布预测的大数据应用案例研究。目前，海洋大数据仍然面临着诸多挑战，海洋数据在不同行业间难以共享，数据缺乏标准化统一管理等。然而，随着技术的发展，对海洋的认知和大数据技术的深入结合，海上信息服务应用与智能化管理必然将得到逐步提高。

参考文献:

[1] 孙朝随，刘青，胡桐，等. 海洋大数据处理软件体系结构设计[J]. 中国海洋大学学报(自然科学版)，2016，45(2)：134-137.
SUN C S, LIU Q, HU T, et al. Software architecture for oceanographic big data processing[J]. Periodical of Ocean University of China, 2016, 45(2): 134-137.

[2] 黄昌. 海洋气象导航服务信息系统的设计与实现[D]. 上海：华东师范大学，2010.
HUANG C. Designing and developing marine meteorological service operational system[D]. Shanghai: East China Normal University, 2010.

[3] KÜHNLEIN M, APPELHANS T, THIES B, et al. Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests[J]. Journal of Applied Meteorology & Climatology, 2014, 53(11): 2457-2480.

[4] YING M, ZHANG W, YU H, et al. An overview of the China meteorological administration tropical cyclone database[J]. Journal of Atmospheric & Oceanic Technology, 2014, 31(2): 287-301.

[5] KALNAY E, KANAMITSU M, KISTLER R, et al. The NCEP/NCAR 40-year reanalysis project[J]. Bulletin of the American Meteorological Society, 1996, 77(3): 437-471.

[6] 肖潇，邵哲平，潘家财，等. 基于AIS信息的船舶轨迹聚类模型及应用[J]. 中国航海，2015，38(2)：82-86.
XIAO X, SHAO Z P, PAN J C, et al. Ship trajectory clustering model based on AIS

data and its application[J]. Navigation of China, 2015, 38(2): 82-86.
[7] BREIMAN L. Random forest[J]. Machine Learning, 2001, 45: 5-32.
[8] 王辉, 刘娜, 逄仁波, 等. 全球海洋预报与科学大数据[J]. 科学通报, 2015, 60(5): 479-484.
WANG H, LIU N, PANG R B, et al. Global ocean forecasting and scientific big data[J]. Chinese Science Bulletin, 2015, 60(5): 479-484.

作者简介

王冬海(1968-),男,中电科海洋信息技术研究院有限公司研究员,中国电子科技集团公司首席专家,长期从事信息系统总体、系统仿真、信息安全等前沿技术研究工作,对信息系统仿真和软件工程有深入研究,在软件配置管理方面有丰富的实践经验。

卢峰(1972-),男,中电科海洋信息技术研究院有限公司高级工程师,长期从事信息系统总体、信息处理技术等方向的研究工作,曾在微软和联想公司长期从事国内外大型系统总体设计,熟悉大数据挖掘技术,在软件计算和服务平台方面有丰富的实践经验。现负责海洋大数据平台架构搭建及海洋信息处理技术研发。

方晓蓉(1990-),女,中电科海洋信息技术研究院有限公司助理工程师,主要研究方向为海洋大数据、海洋观测数据分析、海洋模型数值模拟。

郭刚(1983-),男,中电科海洋信息技术研究院有限公司工程师,主要研究方向为大数据分析、信息安全。

收稿日期:2017-03-21