文章编号:1671-4598(2019)01-0263-04 DOI:10.16526/j.cnki.11—4762/tp.2019.01.055 中图分类号:U675.7 TP391.4 文献标识码:B

基于网格化压缩挖掘船舶航道位置信息

刘亚帅，曹 伟，管志强
（南京船舶雷达研究所，南京 211106）

摘要：由于海上目标的异常多为位置异常，为了实现海上军事目标在位置上的异常检测，需要挖掘出海上正常船舶航道的位置信息；针对传统航道挖掘方法多基于单一目标的小样本进行，而无法实现海量数据挖掘航道的问题，提出了一种基于网格化压缩挖掘船舶航道位置信息的算法；该算法首先采用网格化压缩的方法提高了计算效率；之后采用九宫格矢量化方法重构了航迹方向属性；最后通过设置变阈值实现不同方向上的主航道位置信息的提取；实验结果表明网格化压缩方法有效压缩了原始数据，提高了计算效率；同时在适当的阈值下可以有效挖掘出航道的位置信息。

关键词：网格化压缩；九宫格矢量化；数据挖掘

Position information of vessel track mining based on grid compression

Liu Yashuai, Cao Wei, Guan Zhiqiang
(Nanjing Marine Radar Institute, Nanjing 211106, China)

Abstract: Since the anomalies of the maritime targets are mostly positional anomalies, in order to achieve anomaly detection of the position of the maritime military targets, it is necessary to extract the position information of the normal vessel track at sea. Aiming at the problem that the traditional channel mining method, based on a small sample of a single target, can't realize the massive vessel track data mining, The paper proposes an algorithm based on grid compression to mine the vessel track position information. Firstly, the algorithm improves the computational efficiency by using the grid compression method. Then, the trajectory direction property is reconstructed by the nine-grid vectorization method. Finally, the the main vessel track position information was extracted by using the variable threshold. The results show that the grid compression method effectively compresses the original data and improves the computational efficiency. Meanwhile, the position information of the vessel track can be effectively extract under appropriate thresholds.

Keywords: grid compression; nine-grid vectorization; data mining

0 引言

由于海上目标的异常多表现为航迹位置的异常，所以为了能够检测出这种位置异常，需要对正常船舶航道的位置信息进行提取。而船舶自动识别系统（Automatic Identification System，AIS）[1]的广泛应用生成了海量的航迹数据，这为基于AIS数据挖掘正常船舶航道位置信息提供了条件。

就现阶段而言，国内外对航路模型的构建过程多基于单一目标的小样本进行。比如国内的宁建强等[2]提出一种细粒度网格的方法，通过对航迹从不同粒度上进行处理从而提取航路模型；国外的Pallotta等[3]通过采用改良后的的DBSCAN（density-based spatial clustering of applications with noise）聚类算法构建相应的船舶航路模型；而Guillarme等[4]则采用轨迹分割的方法完成航路聚类构建；至于Osekowska等[5-7]则通过将势场的概念应用于船舶航路的构建和提取中。

这些航路模型构建的方法在海量数据背景下，都会存在样本描述整体出现偏差的问题，更有甚者发生错误与遗漏[8]。所以为了从海量AIS数据中提取出船舶的主航道位置信息，为海上目标的异常检测提供位置上的先验知识，本文提出了一种基于网格化压缩提取船舶航道位置信息的算法。首先为了提高海量数据的挖掘效率问题，该算法根据AIS航迹特点提出采用网格化压缩的方法将航迹数据进行压缩，从而在保证不丢失航迹位置信息的基础上降低数据量，从而提高了处理效率；之后由于压缩后的航迹数据缺少航向信息，所以该算法采用了九宫格矢量化的方法重新构建方向属性，既重构了方向属性，又消除了原始AIS数据中的船首向抖动的问题；最后在重构的八个方向上分别采用变阈值设置方法实现最终的密度聚类，挖掘出不同方向上航道的位置信息。

1 航道位置信息挖掘算法模型

本文基于网格化压缩挖掘船舶航道位置信息算法的总体模型架构如图1中的船舶航道位置挖掘算法流程图所示：

根据流程图可以看出，本文主要从四部分进行处理和研究的：1）数据预处理；2）网格化处理；3）九宫格矢量化；4）航道位置信息模型的挖掘与展示。

其中数据预处理过程主要是从数据去噪和航迹切割角度进行。具体过程如下所述。

收稿日期:2018-09-14； 修回日期:2018-10-22。
基金项目:海军预研课题(3020104080503)。
作者简介:刘亚帅(1993-),男(通讯作者),山西朔州人,硕士研究生,主要从事雷达数据处理,数据挖掘方向的研究。

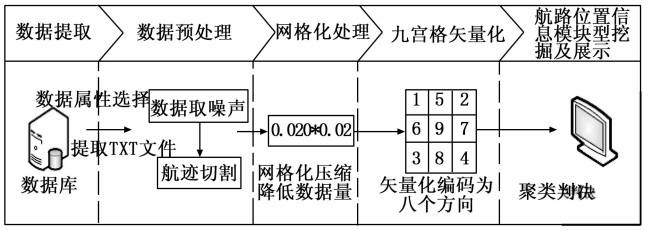

图 1 船舶航道位置挖掘算法流程

因为经由数据库中提取出的 AIS 原始数据存在大量噪声数据，例如 MMSI 噪声，航迹位置噪声及航向信息噪声等。为了能够剔除噪声数据，本文通过从 MMSI（Maritime Mobile Service Identity）号，航以及航迹经度、纬度等角度对其进行筛选去噪，即数据去噪。

因为同一个 MMSI 号对应的航迹实际上存在有多个空窗期（见定义 1），所以为了能够满足后续在网格化压缩过程中对船舶航迹数据连续性的要求，需要将同一个 MMSI 号对应的 AIS 数据依照空窗期进行分割，将其切割成若干连续的航迹，即航迹切割。

定义 1：空窗期。

由于主观或客观环境的干扰导致目标离开监控窗口，监测仪器在一段相当长的时间内丢失对这个目标的跟踪，而且在再次跟踪到该目标时，无法忽略中间缺失过程产生的误差，这段缺失过程就是该目标的一个空窗期。

2 网格化压缩方法

实际中，因为船舶的航速较低，一般大约在 30 几节以下，对于超过 30 节的船舶很少，而相较之下 AIS 信号产生的频率却很高，一般在十几秒到几十秒之间，最快的 2 到 3 秒就发射一次，这就导致在相当长的一段时间之中，AIS 系统生成了大量的航迹点，而实际上船舶的位置和航向等航行状态却并未发生太大的实质性的变化，所以这就使得船舶的航迹数据里包含了大量的冗余状态信息。为了去除这些冗余的数据，本节提出了一种网格化压缩的方法，通过对每条连续的航迹数据进行压缩处理，从而实现了冗余数据的去除，同时也有效提高了后续处理过程的计算效率。

网格化压缩方法主要是通过以下四个步骤进行处理的：

第一步：首先对待研究的区域 W 按照经纬度的值划分成大小为 step 的网格，如图 2 的网格化过程图所示（图中 A'是 A 的放大），这里的 step 就是网格的最小粒度大小，该值的大小选择决定着网格化压缩的效果；

第二步：然后将落在网格中的所有航迹点的经纬度坐标都改成网格的中心点的经纬度坐标，而对于落在网格边界线上的点统一移到左侧或上侧的格子中，即将图 2 中所有落到 A'格子中的点都压缩到点 1 位置。

第三步：之后将每一条连续的航迹按照时间属性（UDT _ TM）进行排序，并对经纬度采用一阶差分的算法进行计算；

第四步：最后将经纬度差分为零的点去掉。至此实现了网格化压缩。

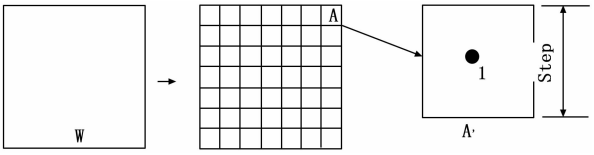

图 2 网格化过程图

3 九宫格矢量化及航道位置挖掘

而压缩后的航迹数据失去了航向信息，所以本节提出了一种九宫格矢量化的方法为压缩后的航迹重构方向属性。而为了能并行化的挖掘出不同航向上船舶航道的位置信息模型，需要从不同航向上分别进行航道位置挖掘，但由于矢量化后的不同航向上的航迹数据的分布存在差异，所以本节采用变阈值的方法分别设置密度阈值从而实现在不同航向上航道位置的提取。

3.1 九宫格矢量化

经过上一节中的网格化压缩之后，船舶的航迹数据只剩下位置信息，而缺失航向信息，所以为了能够使网格化之后的航迹数据准确反映航向信息，同时去除原始航迹中航向抖动的问题，本节提出了一种九宫格矢量化的方法重构航迹的航向属性。

定义 2：矢量化规则。

假设航迹的初始航迹点的经纬度坐标为 P0（LON0，LAT0），其下一航迹点的经纬度坐标为 P1（LON1，LAT1），则根据代数矢量的定义法则可以得出 P0 点的矢量为（LON1－LON0，LAT1－LAT0）。

假设 1：九宫格编码假设。

如图 3 所示，首先对九宫格进行编码形成编码格，然后假设 9 号格为初始航迹点的位置 P0，则下一航迹点的位置 P1 定会落到剩余的八个格中的任意一个格中，则此时 P0 的方向编码即为 P1 点落到的格子对应的编码号。

1	5	2
6	9	7
3	8	4

图 3 九宫格编码图

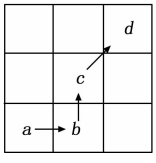

图 4 假设轨迹图

如图 4 是一个假设的航迹，根据矢量化规则与九宫格编码假设可以得出轨迹点 a 的轨迹方向编码为 7，轨迹点 b 的轨迹方向编码为 5，轨迹点 c 的轨迹方向编码为 2。这样这条轨迹方向的整体编码形式就是 7－5－2。如此这般即可将所有的压缩后的航迹点重构出其航向属性。

但是该矢量化方法需要满足假设 1 的条件，否则编码不成立。所以为了确定压缩后的航迹数据是否满足假设 1，需要对网格化压缩后的航迹数据进行统计分析，具体过程可见第四章节实验中的概率密度统计分析。

3.2 变阈值设置挖掘航道

本文实现航道位置信息挖掘的过程基于的是网格密度

聚类思想，网格密度聚类需要设置密度阈值 T_D。而为了从不同航向角度分别实现航路位置模型的提取，同时考虑到九宫格矢量化的 8 个方向的航迹数存在差异的问题，如果设置固定的密度阈值 T_D，会导致一些方向的航迹数据被其他方向的航迹数据冲淡，而误判为噪声网格，为此本节提出采用一种变阈值设置的方法（见定义 3）对不同航向的航迹数据分别设置密度阈值 T_D。

定义 3：轨迹格点阈值。

设网格密度值为 D（D_1，D_2，…，D_n），则如公式 1 所示，其中 μ 为网格密度值的均值，σ 为网格密度值的标准差，T_D 为密度阈值，m 为一个系数（$m>0$）。

$$\mu = (D_1 + D_2 + \ldots + D_n)/n$$
$$\sigma = \sqrt{((D_1-\mu)^2 + \ldots (D_n-\mu)^2)/n}$$
$$T_D = \mu + m \cdot \sigma \tag{1}$$

之后通过阈值判决，将超过阈值 T_D 的网格判为正常航道网格，将低于阈值 T_D 的网格设置为噪声网格，并将其直接去除，最后通过可视化的方法对航道网格进行展示和选择。

4　实验过程及结果分析

4.1　实验环境及数据

本实验的硬件平台为 Window7 x64 位系统，内存为 24.0 GB，处理器为 Intel（R）Core（TM）i7－4770K CPU @ 3.50 GH，软件平台为 R－3.4.3 版本。本实验使用的是 XX 雷达基站采集的厦门港口海域的近一个月的 AIS 数据，其数据总量达到 28.8 GB。

4.2　实验结果分析

4.2.1　网格化压缩效果分析

根据文献［2］介绍，图 2 中的 Step 大小是网格划分的关键。Step 太大，容易丢失船舶轨迹中的一些转向信息；Step 太小，压缩不充分。所以根据两个先验知识：1）在船舶的正常航行中，正常转弯过程都是在 4 到 5 倍的船长为半径的区域内进行；2）地球的经纬度每 0.01 度对应实际的地理距离约是 1.1 公里左右（见公式 2）。而由于正常船舶最长 400 多米，所以本文采用 0.02°×0.02°网格进行划分，这样既可以充分压缩，又不会丢失转向信息。

$$L_{1^\circ} = 2\pi R/360^\circ \approx 111.139\text{km}$$
$$L_{0.01^\circ} = L_{1^\circ}/100 \approx 1.1\text{km} \tag{2}$$

实验中，由于数据量太大，R 语言对于硬件平台的内存要求较高，在对原始数据进行处理时，无法一次性将所有数据进行展开，所以本文通过将一个月的数据依据时间属性分割成为三个数据集分别进行网格化压缩处理。

表 1　数据压缩效果表

时间/日	压缩前/GB	压缩/MB	压缩倍数
1－10	9.1	47	193.6
11－20	11	53	207.5
21－30	8.7	39.7	219.1

如表 1 所示是原始数据集分割成的三个数据集的三次数据压缩的效果，由此可见网格化压缩的倍数都在 200 倍左右，所以可见网格化压缩方法的压缩效果显著，可以在很大程度上去除冗余航迹信息，提高后续航迹位置信息提取过程中的计算效率。

为了确定压缩后的数据分布情况，从而确定压缩后的航迹数据是否满足九宫格矢量化假设要求，本实验对差分后数据的经度差分值属性（DLON）和纬度差分值属性（DLAT）进行了概率密度值统计。如图 5 所示，分别对应的是差分后的经度差分值属性（DLON）和纬度差分值属性（DLAT），其二者的分布主要集中在 0.02°，0°和－0.02°，其中中心最高峰值对应的是 0，左侧次高峰对应的是－0.02，右侧次高峰对应的是 0.02。这就表明差分后的航迹数据依然是连续的，压缩并没有失真。由此也可以证明，压缩后的航迹数据是满足九宫格编码假设的，因此压缩后的航迹数据是可以通过九宫格矢量化规则进行航向属性的重构。

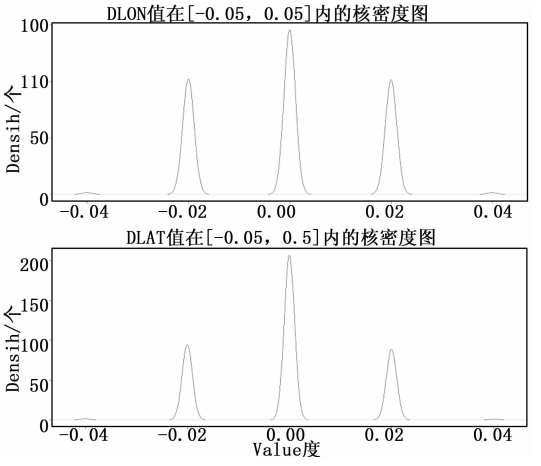

图 5　差分后概率密度统计

4.2.2　聚类效果展示

本实验采用并行化的思想，从重构的八个航向向上分别进行提取船舶航路位置模型，其网格密度阈值的选取是根据公式 1 进行的。实验中分别选取取 m 值为 0，1，1.5 和 2 进行了研究。实验结果显示，不同航向上航路的密度阈值中的 m 值的选择影响航路位置模型提取的效果。其中网格密度值较大的航向上，m 值选择较大时效果显著，而网格密度值较小是，m 值的选择较小时效果较好。

这是由于当网格密度较大时，该航向上航行船舶较多，出现异常的船舶也较多，只有提高 m 值，才能有效去除航路中的异常航迹；而当网格密度较小时，表示该航向上航行的船舶较少，尤其是受地理地形影响较大的航向区域，其基本上很少有异常航迹，所以只需要取较小的 m 值就可以完全实现航路位置模型的提取，反而较大的 m 值易于造成正常航路被误判而去除。

如图 6 所示是部分方向在适当的 m 下的航道位置聚类效果图，其中图 6.1 是 $m=1.5$ 时，3 号航向上的航道位置

模型提取的效果图；图 6.2 是 $m=1$ 时，5 号航向上的航道位置模型提取的效果图。由于采用了网格化压缩，所以航道点迹成为离散状态，即图中显示的航路是有一个一个的航迹点组合而成的航迹区域。这也同时指出，在后续的海上目标的异常检测过程中，也需要对船舶目标的航迹进行网格化压缩，使其成为离散转态后再进行对比判决检测。

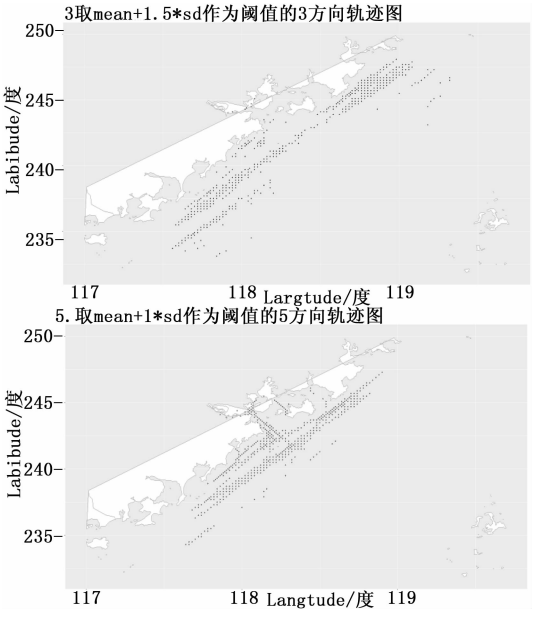

图 6 提取的航道位置信息效果

5 结束语

为了从海量 AIS 数据中挖掘出船舶航道的位置信息模型，本文根据船舶航迹存在冗余的特点，提出一种基于网格化压缩挖掘船舶航道位置信息的算法。其中针对计算效率低的问题，该算法提出采用网格化压缩方法有效去除了航迹数据中的冗余数据，提高了海量数据聚类的计算效率。同时该压缩方法也为海量路径数据的压缩提供了一种算法支持。为了解决压缩后的航迹数据缺少航向信息的问题，本文采用的九宫格矢量化方法在网格化的基础上重构了船舶航向属性，为压缩后的航迹增加方向属性，同时也有效去除了原始航迹中航向抖动的问题。最后为了避免不同方向上数据相互干扰，易将核心网格误判成噪声网格的问题，本文在矢量化的八个方向上，分别采用不同的阈值实现航道位置信息模型的提取。

本实验中可视化的结果表明在选择合适的阈值下，该算法可以有效提取船舶的主航道位置信息。但是该算法提取的船舶航道只有一个位置信息，缺少航迹中的时序信息，需要后续从时序角度进一步挖掘提取。

参考文献：

[1] Tang Cun－bao，Shao Zhe－ping，Wu Jian－sheng，et al. Based on the research of AIS information density distribution algorithm and implementation [J]. Journal of Guangzhou Maritime College，2012，20 (3)：7－10.

[2] 宁建强，黄 涛，刁博宇，等. 一种基于海量船舶轨迹数据的细粒度网格海上交通密度计算方法 [J]. 计算机工程与科学，2015，37 (12)：2242－2249.

[3] Pallotta G，Vespe M，Bryan K. Vessel Pattern Knowledge Discovery from AIS Data：A Framework for Anomaly Detection and Route Prediction [J]. Entropy，2013，15 (6)：2218－2245.

[4] Guillarme N L，Lerouvreur X. Unsupervised extraction of knowledge from S－AIS data for maritime situational awareness [A]. International Conference on Information Fusion [C]. IEEE，2013：2025－2032.

[5] Osekowska E，Carlsson B. Learning Maritime Traffic Rules Using Potential Fields [A]. International Conference on Computational Logistics [C]. Springer，Cham，2015：298－312.

[6] Osekowska E，Axelsson S，Carlsson B. Potential Fields in Modeling Transport over Water [J]. Operations Research/ Computer Science Interfaces，2015，58 (2)：259－280.

[7] Osekowska E，Johnson H，Carlsson B. Grid Size Optimization for Potential Field based Maritime Anomaly Detection [J]. Transportation Research Procedia，2014，3 (2)：720－729.

[8] 高 曙，刘甜甜，初秀民，等. 船舶异常行为研究进展及发展趋势 [J]. 中国航海，2017，40 (2)：38－43.

(上接第 262 页)

[3] Barnard A，Nwosa C. COTS Based On－Board－Computer on South Africa's Sumbandilasat：A Radiation and In－Orbit Performance Analysis [A]. IEEE Radiation Effects Data Workshop [C]. IEEE，2011：1－4.

[4] 王 峰，郭金生，李 晖. 商用现货器件在卫星中的应用 [J]. 航天器工程，2013，22 (4)：87－94.

[5] 姜秀杰，孙辉先，王志华. 商用器件的空间应用需求、现状及发展前景 [J]. 空间科学学报，2005，25 (1)：77－80

[6] 谢 钢. GPS 原理与接收机设计：Principles of GPS and receiver sdesign [M]. 北京：电子工业出版社，2009.

[7] 王 静. GNSS 接收机 CMOS 射频前端芯片系统级设计 [D]. 上海：上海交通大学，2008.

[8] 梁淏翔. 基于 FPGA 的可重构多模导航基带处理技术研究 [D]. 成都：电子科技大学，2014.

[9] 邢克飞，何 伟，杨 俊. COTS 器件的空间应用技术研究 [J]. 计算机测量与控制，2011，19 (7)：1741－1745.

[10] 丁义刚. 空间辐射环境单粒子效应研究 [J]. 航天器环境工程，2007，24 (5)：283－290.

[11] 贾文远，安军社. COTS 器件的空间辐射效应与对策分析 [J]. 电子元件与材料，2015，34 (11)：1－4.

[12] 李 毅，李 瑞，黄 影，等. 基于 COTS 的空间信息处理系统单粒子闭锁保护技术实现 [J]. 宇航学报，2007，28 (5)：1283－1287.

[13] 张 昊，王新升，李 博，等. 微小卫星单粒子闩锁防护技术研究 [J]. 红外与激光工程，2015，44 (5)：1444－1449.

[14] 袁春柱，李志刚，李军子，等. 微纳卫星 COTS 器件应用研究 [J]. 计算机测量与控制，2017，25 (2)：156－159.

[15] 王 淼，纪文章，宁金枝，等. 星载扩频应答机抗 SEU 方法及验证 [J]. 航天器工程，2014，23 (1)：91－95.