iGroup Learning and iDetect for Dynamic Anomaly Detection with Applications in Maritime Threat Detection

Chencheng Cai, Rong Chen, Alexander D. Liu, Fred S. Roberts, Minge Xie
Department of Statistics and CCICADA Center
Rutgers University
Piscataway, New Jersey 08854
Email: froberts@dimacs.rutgers.edu

***Abstract*—The maritime transportation system is critical to the US and world economy. This paper reports on two novel statistical tools, iGroup learning for individualized grouping and baseline distribution formation, and iDetect for subsequent individualized detection of anomalous deviations from the baseline distribution. These statistical methods are being developed, tested, and implemented in the context of maritime threat detection, but can be easily applied in other areas. In the maritime domain, the tools aim to provide early warnings of anomalies and assessments of resulting risk for vessels being monitored. The paper presents some preliminary results about these tools and specifically reports on a case study aimed at finding anomalous behavior for vessels approaching a port.**

1. Introduction

Security is of paramount importance to human existence. Accurate and early detection of threat is becoming increasingly crucial to security. Novel and sophisticated statistical methods/algorithms are needed to enhance our ability to detect threats, and take advantage of the vast amount of readily available data.

The maritime transportation system is critical to the US and world economy [9]. After the 9/11 terrorist attack, air-traffic has seen significant security and threat prevention enhancement. The detection and prevention of threats from maritime traffic, however, lags behind, and it raises many challenges, due to the vast ocean space and complex river systems, long and often unmonitored shorelines, extremely busy ports, and the sheer volume of goods being transported (e.g., account for about 25% of US GDP). Threats through maritime traffic can be multi-faceted and serious, ranging from human and drug trafficking, smuggling, transport of nuclear material and dirty bombs, to garbage dumping and illegal fishing. Hence it is important to have an efficient early detection and risk assessment system for maritime traffic over space and time. Identification of unusual maritime traffic rapidly in terms of time and accurately in terms of location is critical to being able to apprehend those who are violating some law or planning some illegal or dangerous act. With today's data gathering capability and global regulation and agreements, leading to large amounts of data obtained through what is called the Automatic Identification System for vessels (see Sec. 2), it is now possible to achieve such a goal with sophisticated and advanced statistical tools.

We have started to develop two novel statistical tools, i**Group learning** for individualized grouping and baseline distribution formation and i**Detect** for subsequent individualized detection of anomalous deviations from the baseline distribution, and hence accurate risk assessment. These statistical methods are being developed, tested and

implemented in the context of maritime threat detection, although they are general statistical methods that can be easily applied in others areas such as cell phone monitoring, cyber security, anti-money laundering, and others. We are consolidating publicly available maritime traffic data sets and geological/geographical/geophysical data sets that provide the necessary information for analysis, including vessel characteristics and spatial and temporal characteristics of vessel movement. Our tools aim to provide early warnings of anomalies and assessments of resulting risk for vessels being monitored, finding ways to trace vessel movements vessels, with an emphasis on threat detection. In this paper, we present some preliminary results obtained from our iGroup learning and iDetect tools.

2. Automatic Identification System (AIS)

Since the year 2000, the International Maritime Organization (IMO) has required, via the *International Convention for the Safety of Life at Sea* [8], that all ships of 300 gross tons or more, and all passenger ships install an identification and location device on board that consistently and automatically transmits dynamic data (location, course, speed, etc.) and voyage-related information (destination, estimated time of arrival (ETA), etc.) to Vessel Traffic Services (VTS) stations as well as to other ships. The Automatic Identification System (AIS), developed by technical committees under the auspices of the IMO, uses Global Positioning Systems (GPS), gyrocompass, other shipboard sensors and digital VHF radio communication equipment. Vessel identifiers such as vessel name and VHF call sign are programmed in the device and are also included in the transmittal.

The global AIS system receives data from approximately 1,000,000 ships with updates for each ship as frequently as every two seconds while in motion and every three minutes while at anchor. AIS tracks ships automatically by electronically linking data with other ships, AIS base stations, and satellites. This system enables ships to share positional data with other ships, and while its primary initial function was for traffic management, collision avoidance, and other safety applications, it has extensive other uses. Overall, AIS offers awareness about vessels operating within the maritime transportation system [1], [6].

Specifically, AIS data include the vessel's Maritime Mobile Service Identity (MMSI), a unique identification number; navigation status (at anchor, under way using engine(s), or not under command); rate of turn; speed over ground; position accuracy; Longitude and Latitude; course over ground in degrees; true heading from gyro compass; and time stamp. Less frequent broadcasts also include IMO ship identification number (which remains unchanged upon transfer of the ship's registration to another country); international radio call sign assigned to the vessel by its country of registry; vessel name; type of ship/cargo; dimensions of ship; type of positioning system (GPS, Differential Global Positioning Systems (DGPS) or Long Range Navigation (LORAN)); draught of ship; destination; ETA at destination. Figure 1 shows a partial route of a typical vessel, using AIS data. The grey area shown, obtained from AIS data and methods we have developed, depicts the 'normal' routes taken by similar ships. A ship that goes out of the grey area may indicate a high level of abnormality - an anomaly.

The data is massive. For example, US coast and waterway related information accounts for 32 GB of AIS data each day. Discussions with the US Coast Guard (USCG) Research and Development Center and USCG Headquarters have led to the idea that AIS data could be used to flag early warning of changed shipping patterns and other anomalous behavior. Selected AIS data are available from the public domain at https://marinecadastre.gov/ais/.

Figure 1. Partial Route of a Vessel from AIS Data

3. iGroup Learning and iDetect

To deal with heterogeneity in the population, conventional methods often cluster/group individual entities into subgroups and use the data in the same subgroup for statistical analysis, as often seen in subgroup analysis, personalized medicine and fusion learning (see for example [3], [5]). The clustering and subgrouping are typically performed a priori. Such approaches have several disadvantages. First, the forming of subgroups depends on a pre-determined number of clusters, a parameter that is difficult to determine. Second, all analytical outcomes and inferences (e.g., estimated parameters, testing) are identical for all individuals in the same subgroup. More importantly, in many situations, there may not be any clear-cut and well divided subgroup structure in the population. In these situations, the conventional subgroup analysis imposes an artificial grouping structure to the population and the analysis often leads to large biases and thus invalid inference in many cases.

We have begun to develop a novel statistical method, individualized grouping (i**Group**), to be used to identify a group of vessels that behave similarly to a target individual vessel, and hence enable us to effectively establish baseline (normal) behavior, in terms of a joint distribution of features, of the target vessel. In contrast to the traditional methods, iGroup focuses on each individual and forms one individualized group for each individual, by locating individuals that share similar characteristics of the target individual. It sidesteps all the aforementioned difficulties by forming an iGroup specifically for the target individual while ignoring other entities that have little in common with the target, even if there is no clear-cut and well divided subgroup structure. In our anomaly application, such a group (formed based on standard features) allows us to establish a baseline behavior (joint distribution of risk-related features) for the individualized group with high accuracy and from which to detect any anomaly for the target individual.

While customized grouping for a specific individual is not new, our method has a solid theoretical foundation. A detailed discussion of the theoretical development of the methods is beyond the scope of this paper, but roughly speaking the iGroup methodology is similar to nonparametric estimation in spirit. There are significant differences, and nonparametric estimation can be viewed as a special case of iGroup. iGroup is different from clustering as it is localized while clustering is a global partition technique.

Our preliminary analysis suggests that iGroup Learning is robust and effective for handling heterogeneity arising from diverse sources in big data and it is ideally suited for goal-directed applications in individualized inference. Additionally, in terms of computation, by ignoring a large number of irrelevant entities and zoning directly to individuals, iGroup Learning is parallel in nature and

can scale up better for big data than any existing competitors.

We have also begun to develop an individual detection (i**Detect**) method to score the anomaly (abnormality) of the target individual against the baseline distribution of features obtained from its own iGroup. This is essentially an outlier detection problem. However, the challenge here is that the features under consideration may be complex and of mixed type, and the features' importance may not be equal. A new approach is needed to assess the deviation of an observation from its baseline distribution. Our iDetect is based on the idea of data depth [4], extended to complex space, noisy features and features of unequal importance.

4. Features for the Applications and their Use with iGroup and iDetect

We identified key feature groups to be used in iGroup Learning and iDetect through consultation with maritime traffic experts. A partial list includes the following:

- Vessel Profile: vessel type, dimensions, draught, etc.
- Risk Features: country of registration, ports visited in the near past, etc.
- At-anchor Features: AIS indicator, on-land indicator, port indicator, etc.
- In-motion Features: average speed or acceleration, maximum speed or acceleration, etc.
- Functional Voyage Features: trajectory of vessel, time series of speed or acceleration, etc.
- Dyadic Product Formula Features: dyadic parameters of time series of variety of variables.

For maritime traffic monitoring, the iGroup method works this way. Suppose A is a given vessel and $d(.,.)$ is an appropriate similarity or distance measure between the target's feature set D_A and vessel k's feature set D_k. Then we look for the iGroup (clique) consisting of all vessels k so that $d(D_A, D_k) < \tau$ where τ is a threshold determined by optimizing some criterion. Compared with clustering methods, iGroup has the following advantages: it is not necessary to specify the number of clusters, and grouping is based on an individual vessel's characteristics.

Our detection method measures the (potential) anomaly of a target individual vessel against its cohort (iGroup) in terms of feature distribution. Because the detection rule is based on an individual vessel's own baseline distribution, the procedure is called *individualize detection* (iDetect). There are many possible outlier detection rules [2], for example:

- 1.5 interquartile range rule: detect outliers based on empirical quantiles.
- two/three standard deviation rule: detect outliers lying more than two/three standard deviations away from the mean
- parametric distribution-based. quantile rules: detect outliers by estimated quantile from a parametric distribution model.
- multiple hypothesis testing procedures: detect outliers that can reject a corresponding hypothesis testing.
- data depth method: detect outliers that are far away from the data center.

5. Case Study: Finding Anomalies for Vessels Approaching a Port

To illustrate our ideas, we focused on 534 vessels/voyages (tankers, cargo vessels) arriving in the Port of Newark between July and November 2014. We investigated behaviors starting from crossing the 12 nautical mile US territorial sea (TS) boundary to arrival at the port. The trajectory, a functional feature, was used as the standard feature for iGroup. Outliers in duration (time spent from TS boundary to port) were detected based on the two standard deviation rule. See Figure 2.

The distance between trajectories was set to the dynamic time warping distance (DTW) [7] that finds an optimal match between two given

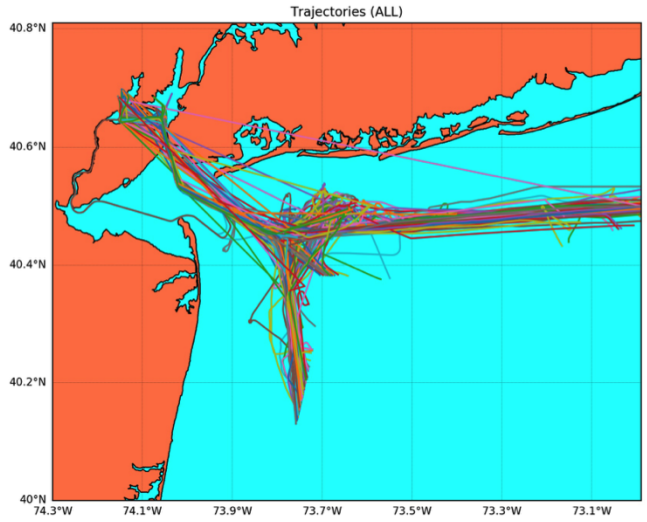

Figure 2. Trajectories of 534 Vessels Heading to Port of Newark

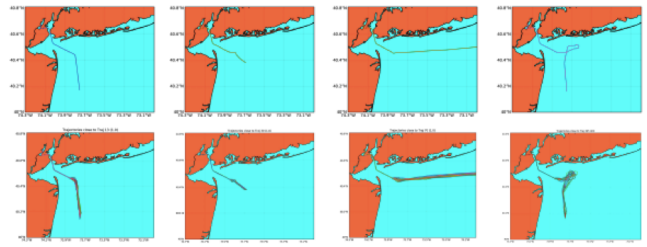

Figure 3. Cliques found by iGroup for Four Vessels/Voyages. Top is Vessel/Voyage, Bottom is its Clique.

sequences (two time series). The eight parts of Figure 3 demonstrate the cliques found by iGroup for four selected vessels/voyages. The top is the vessel trajectory and the bottom is the set of trajectories in its clique.

We then looked for outliers that have abnormal time duration (from territorial sea boundary to Port) compared to vessels in their clique (vessels with a similar trajectory). Outliers were detected by the two standard deviation rule: Vessels in a clique with time duration at least two standard deviations from the clique mean. 95 outliers were detected. These are shown in Figure 4: (a) 50 vessels had a prior dock before the Port of Newark (left); (b) 18 vessels were anchored somewhere outside the port for an extremely long time (middle); (c) the other 27 vessels were traveling too

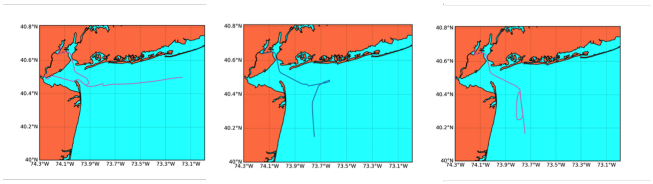

Figure 4. Outliers among Vessels/Voyages in Trajectories of Vessels Heading to Port of Newark

fast/slow compared with their cliques (right). At least (b) could reflect a vessel that was a threat because it was potentially unloading illegal goods such as weapons or narcotics.

Our algorithm can be completed in $O(K^2T)$ time, where $K = 534$ is the number of trajectories and $T \sim 200$ is the average number of time stamps in the DTW algorithm to calculate distance between trajectories. On a 6 core machine, the algorithm took about 26 minutes to do the grouping and detect the outliers. With a single thread, one could guess the time might rise to something like 2.5 hours.

6. Future Work and Research Challenges

Next steps in our research include:

- Broaden the feature sets by including more geological information, vessel history, etc.
- Extend the current results to a high-dimensional features case.
- Investigate theoretical properties of iGroup and iDetect.
- Develop an individualized feature selection scheme.
- Develop an online system that enables online-updating of features and online outlier detection.

The research challenges here include the following:

- Data quality: part of AIS data is incomplete, unreliable or deliberately misleading

- Handling functional data: Treatment of functional data is significantly different from numerical values.
- Extracting useful features from functional data is generally difficult. (E.g., trajectory is a function.)
- Dimension reduction: It is challenging to select from such a large number of features an efficient small subset that could be collected or derived from our data sets

Acknowledgements. The authors thank the National Science FoundatIon for support under grant DMS-1737857 to Rutgers University. This paper has benefitted from discussions with Captain David Nichols (US Coast Guard, retired) about vessel features and the idea of investigating anomalies in vessel arrivals at a port.

References

[1] DiRenzo III, J., Goward, D., Roberts, F.S.: The Little-known Challenge of Maritime Cyber Security. Proc. of the Int. Conf. on Information, Intelligence, Systems and Applications (IISA), IEEE, 1-5 (2015). DOI: 10.1109/IISA.2015.7388071

[2] Hodge, V., Austin, J.: A Survey of Outlier Detection Methodologies. Artificial Intelligence Review 22.2, 85-126 (2004).

[3] Lipkovich, I., and Dmitrienko, A.: Tutorial in Biostatistics: Data-driven Subgroup Identification and Analysis in Clinical Trials. Statistics in Medicine 36 (2017).

[4] Liu, R., Parelius, J.M, Singh, K.: Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference (with discussion and a rejoinder by Liu and Singh). The Annals of Statistics 27, 783-858 (1999).

[5] W. Y. Loh, W.Y., Fu, H., Man, M., Champion, V., Yu, M.: Identification of Subgroups with Differential Treatment Effects for Longitudinal and Multiresponse variables. Statistics in Medicine 35 (2016).

[6] Net Help Security: Digital Ship Pirates: Researchers Crack Vessel Tracking System, October 16, 2013, http://www.net-security.org/secworld.php?id=15781, accessed Feb. 21, 2015.

[7] Sakoe, H., Seibi C:. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing 26.1, (43-49 (1978).

[8] International Maritime Organization: AIS Transponders. http://www.imo.org/en/OurWork/safety/navigation/pages/ais.aspx, accessed Feb. 13, 2017.

[9] United States Coast Guard: Western Hemisphere Strategy, Sept. 2014.