A Cluster-Based Approach to Maritime Data Analysis: The Case of SAT-AIS Data Analysis

Ireneusz Czarnowski[a,*] [0000-0003-0867-3114]

[a] Gdynia Maritime University, Morska Str. 81-87, 81-225 Gdynia, Poland

[*]Corresponding author. E-mail: i.czarnowski@umg.edu.pl

Keywords: machine learning, clustering, SAT-AIS, data analysis, maritime data analytics

ABSTRACT

The paper deals with a problem of determining packages in SAT-AIS dataset and dealing with a AIS packet collision problem. Transponders from different terrestrial AIS service areas, being in the satellite, are not synchronized between themselves or with the satellite. The AIS satellite receiver records signals transmitted by ships located in different areas, although in the satellite field-of-view. Lack of synchronization causes packet collisions and results in difficulties in identifying packages received by a satellite. In the paper a clustering procedure, i.e. machine learning technique, is proposed for packages recovering. Aim of the carried out computational experiment was to answer the question whether clustering can be helpful in a recovery of data when they are lost as is the case with a SAT-AIS system. The obtained results are interesting and gave a possibility to formulate future direction research.

1. INTRODUCTION

Modern and intelligent vessels, including autonomous ships, can contain thousands of different sensors. These sensors can generate huge data every day, so consequently, thousands of data can be collected. Use of these data requires looking for and implementing new technologies and techniques, including new approaches to data processing. Data analysis is opening up to new opportunity to create values from data. Data analysis is the process of discovery and drawing conclusions from information - from technical point of view, from data collected in different repositories. Data analysis supplies different tools for data processing, among them based on machine learning algorithms, and becomes increasingly important also for ship control. This importance should be merged with different critical decisions and ship's automation. Data analysis is also becoming crucial for shipping, where greater volumes of data are today more natural than ever before. In this case, data analysis is needed to improve performance of monitoring and optimization of operations, condition of monitoring the maintenance of equipment and assets, as well as safety improvements. DNV underlines that data analysis will not only be beneficial in the current environment, but will prepare the organization for the future. So, it is the reason why data analysis is pointed as a one of six key technologies that will transform the marine and shipping operations [2]. In [2] there are also pointed out four primary technology enablers for the collection and use of data analysis tools in shipping, i.e.:

- availability of affordable, reliable and accurate sensors for data collection,

- high data transfer speeds between ship and shore,
- continuously increasing computational power and development of IT platform solutions,
- development of analytical methods and algorithms for creating value from collected data.

An example where machine learning tools can be successfully implemented is the AIS data analysis. The Automatic Identification System (AIS) is an automatic tracking system based on transponders located on ships and used by vessel traffic services (VTS) (see, for example [4]). Based on AIS statistics and vessel data, machine learning tools can relatively accurately predict future vessel movements, several port steps ahead [2], [8]. Due to increasing ship traffic in ports, the maritime traffic safety needs more attention, and analysis of data collected from AIS can help to control and prediction of ship behaviors. Based on deep and wide analysis of data from the AIS system, a big set of different parameters and factors can be monitored [3]. Combining different data, which are available in a public domain, like for example vessel position and speed derived from AIS data, as well as weather data and others different reports for own or market needs can be created for optimization of different operations or for managing of vessels and transportation operations.

Analysis of AIS data can bring new outputs and provide concrete safety threshold for components under scrutiny [2]. AIS data are also the object of different analysis concerning behaviors of vessels which are or can be carried-out with respect to the vessels security or to monitoring and predicting of future behaviors of vessels on open sea or restricted waters. Examples of such analysis can also concern detection of anomalies in behavior of vessels, unreasonable approaching of ships, illegal activities, pollution of the environment, detection and warning against danger of collision or other dangerous situations. Vessel behaviors' analysis are now important from global scale point of view. The rationale for such analyzes are maritime terrorism and growing number of acts of piracy [6].

Thinking about increasing the effectiveness and effective of different processes, from monitoring of the environment pollution to illegal activities and behaviors of the vessels, it is reasonable to collect and analyze of AIS data from a global perspective. It is a reason of introducing of the satellite Automatic Identification System [5]. "*Space-based AIS (SAT-AIS) will make it possible to track seafaring vessels beyond coastal areas that are equipped with AIS tracking devices. SAT-AIS is a promising solution to overcome terrestrial coverage limitations with the potential to provide AIS service for any given area on Earth*" [9].

However, from a technology point of view, the SAT-AIS is more complex. Data collected by the satellite component of SAT-AIS are incomplete and contain noise. So, it is difficult to determine the AIS data packages in their basis. It means, that numerous incomplete packages result in decreasing volume of useful data. It also means, that the satellite transponder can't provide in a regular way full and complete data on vessel traffic to the ground component. Thus, numerous incomplete packages result in decreasing volume of useful data and result, that the accuracy of ongoing different analysis business can be poor or incomplete.

Machine learning tools can be very helpful in data processing for the needs of different systems, including marine systems, in case when the data is incomplete. In this paper the clustering-based approach, i.e. unsupervised machine learning, for analyzing of AIS data and for determining packages are considered. In general, aim of the research was to answer the

question, whether clustering algorithms can be helpful in a recovery of data when they are lost as is the case with the SAT-AIS system.

2. PROBLEM FORMULATION

2.1. SAT-AIS FOR MARINE TRAFIC MONITORING

Based on the International Maritime Organization regulations, ships of 300 tons or more in international voyages, cargo ships of 500 tons or more in local waters and all passenger ships, irrespective of size, are required to be equipped with Automatic Identification System (AIS). In result, the Automatic Identification System (AIS) has been introduced in late 90s., and according to the IMO regulations, it has been prepared for identifying ships and support the navigation procedures. Since then, AIS is one of the basic tools for maritime traffic monitoring. AIS is also included in Aids-To-Navigation and Search and Rescue transponders.

AIS as a radio-based communication system was originally developed to prevent collisions of large vessels. It transmits the course and speed as well as identification and position information to other vessels and shore stations. While AIS has been deployed for global operations, it has a major limitation. The limitation follows from the Earth's curvature which limits its horizontal range to about 74 km from shore. This means that AIS traffic information is available only around coastal zones or on a ship-to-ship basis. Thus, AIS has been classified as a system for exchange of information at a local scale [9].

SAT-AIS has been introduced as a system for exchange of information at a global scale. The ship's identity is recorded and decoded by satellite then sent to ground stations for further processing and distribution [10]. Such system is promoted, among others by European Space Agency (ESA) [9].

2.2. SAT-AIS PACKETS COLLISION

SAT-AIS has been introduced based on generic assumptions of AIS, especially considering transmission of packages directly from mobile components (located onboard of vessels) to the receivers. Consequently, it generates a number of problems for a satellite system (component of the satellite-based AIS), which is between the mobile components and the receivers and which aim is to detect signals from vessels and then retransmit them to the receivers in a global scale.

Main problem for a satellite system is AIS packet collisions. Collisions are observed onboard of a satellite. The AIS receiver, installed onboard of a satellite records signals transmitted by ships located in different areas, but being in the satellite field-of-view (FOV) - for example, Fig. 1 shows terrestrial AIS service areas. Transponders from different terrestrial AIS service areas, being in the satellite FOV, are not synchronized between themselves or with the satellite [10].

In other words, signals transmitted by ships from different areas but within the same FOV and within the same time slot result in the collision of AIS messages. The signals transmitted within each slot are received by the satellite with different amplitudes, time delays, and different Doppler frequency shifts due to the spatial separation of the ships [11]. It results in an interference of different signals in time and difficulties with decoding the packets, so, it is difficult to establish current information concerning the ships. It means that the AIS receiver

would not have a problem with receiving packets from only one area, for example from N_0 (see fig. 1), or if the transponders from different terrestrial AIS service areas were synchronized.

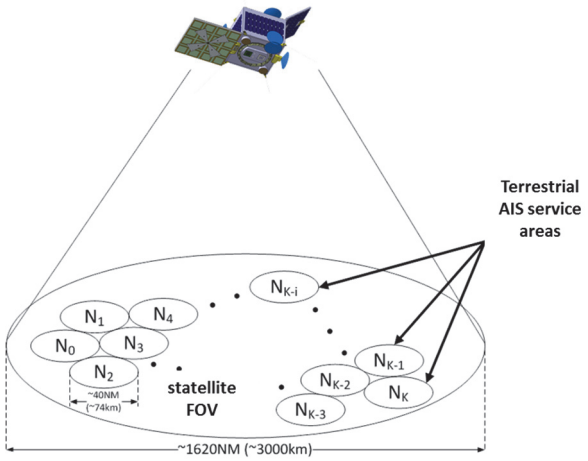

Fig. 1. AIS service areas in the satellite field-of-view [11]

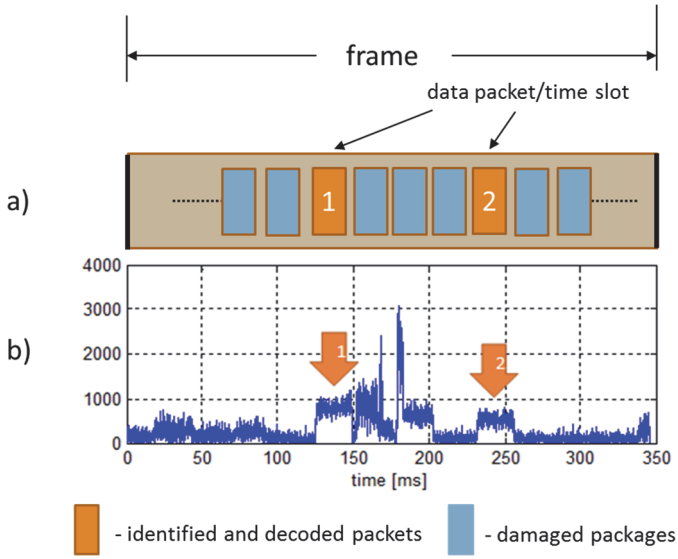

Fig. 2. A time frame presenting of location of received AIS packages - data packages (a) and a fragment of the AIS signal received by the satellite (see in [10]) (b).

Fig. 2 shows an example of data received by a satellite and data packages decoded from these data. On the picture we can observe that only two packages have been correctly identified. The rest of packages have not been decoded. We can only assume where in the frame there should be time slots.

As per statistical observations, the probability of receiving more than three ship correct packages in each frame at the satellite is quite low (see [11]). Thus, it is a reason for looking for solutions for eliminating of AIS packet collisions. The existing research directions concentrate on packet recovery by analyzing of received and sampled AIS signal. Among different approaches and methods we can find decollision methods of the packages by recognition of shape of the received sequence, decoding methods using the Viterbi algorithm, methods for progressive, reverse or hybrid estimation of signal samples or blind source separation algorithms (see, for example, [11], [12], [13] and [14]).

In this paper a clustering approach has been proposed for packages recovering. The approach is discussed in the next section.

3. A CLUSTER-BASED APPROACH TO SAT-AIS DATA ANALYSIS

3.1. UNSUPERVISED LEARNING FOR DATA ANALYSIS

The key objective of machine learning is to design algorithms that are able to improve performance at some task through experience. Learning from examples is the most important paradigm of machine learning. Learning from examples is an example of so-called supervised learning, where the data used in learning process has previously known labels. In other words, our data has some target variables with specific values that is used in the process of producing knowledge model, which next can be implemented for support in different decision aided making process. In such case we consider models and algorithms for solving classification problems, as well as regression problems or time series prediction problems.

However, in case of most problems in real-world, data do not come with predefined labels. In such case the aim of learning process is to develop machine learning models that can classify correctly this data, by finding by themselves some commonality in the features, that will be used to predict the classes on new data [14]. Such approach belongs to so-called unsupervised learning paradigm.

Unsupervised learning can be applied to: segmentation of datasets by some shared attributes, detection of anomalies that do not fit to any group or aggregate variables with similar attributes, thus to simplify datasets [14]. An example of unsupervised learning is clustering. The main aim of clustering is to find different groups within the elements in the data. It means that clustering algorithms find the structure in the data so that elements of the same cluster (or group) are more similar to each other than to those from different clusters. In this case, the strength of machine learning model results from the fact that it is able to infer that there are two different classes without knowing anything else from the data. The unsupervised learning algorithms have a wide range of applications and are useful to solve real-world problems such as anomaly detection, recommending systems, documents grouping, etc. Among the most known clustering algorithms there are: k-means, hierarchical clustering and Density Based Scan Clustering (DBSCAN).

Hierarchical clustering is one of alternative approaches for clustering. The main idea of hierarchical clustering bases on preparing dendrograms which are visualizations of a binary hierarchical clustering. An agglomerative method for preparing of dendrogram starts with each sample being a different cluster, and then in following iterations merge these that are closer to each other until there is only one cluster. However, different criteria of defining closeness can be applied (single linkage, complete linkage, average linkage, ward's linkage-method, etc.), as well as different clustering distance measures (Euclidean, Manhattan, etc.). The obtained dendrograms provide information on relationships within data. Using hierarchical clustering we do not need to initially specify the number of clusters, but they can be found after the analysis on the dendrogram.

3.2. UNSUPERVISED LEARNING FOR RECOVERING PACKAGES IN SAT-AIS

The main goal of the paper is to propose a procedure which may be utilized on board a satellite for identification and decoding of AIS messages and especially for recovering the damaged packages.

The pseudo-code explaining how the damaged packages can be recovered is shown as Algorithm 1. The proposed algorithm bases on a hierarchical clustering of packages, which at

the beginning of analysis are identified in data stream received by a satellite. Assuming that within data stream there can be correct packages, after clustering analysis the induced clusters are compared with the correct packages. Based on identified similarities between the correct packages and the obtained clusters, we can start to predict of damaged parts of packages.

Algorithm 1 Recovering the damaged packages in SAT-AIS data stream

Input: Z_t – set of received data in time t
Output: P_t – set of packages identified and decoded in time t
Begin
$P_t := \emptyset$ - set of identified packages
$P'_t := \emptyset$ - set of damaged packages
Run a decoding procedure of the dataset Z_t and identify time frame Z_{t1}, …, Z_{tK}, where K is a number of time frame
For i:=1 to K do
Identify in Z_{ti} correct P_t^{ij} and damaged P'^{ij}_t packages, where j is a number of identify package in frame i
$P_t = P_t \cup P_t^{ij}$
$P'_t = P'_t \cup P'^{ij}_t$
End For
Run the clustering procedure on data from $P_t \cup P'_t$, where G is a set of obtained group of packages
Evaluate of the similarity between packages in G and correct identified packages in P_t and update P_t
Return P_t
End

4. COMPUTATIONAL EXPERIMENT

The proposed approach has been validated experimentally. The main research question was whether the clustering can be helpful in a recovery of data when they are lost as is the case with the SAT-AIS system.

Datasets used in the reported experiment have been obtained from AIS receiver and include data from the Gulf of Gdansk. The data has been recorded in 35 minute time period and include 1077 different packages, which identify 36 vessels. Then only part of the packages describing each of the 36 ships have been preserved in the original. The remaining packages were damaged. Errors were introduced based on the mechanism of random disturbance of the package structure. It means, that dataset in the experiment consisted of original and damaged messages (packages). Finally, three sets of data were used, consisting of 500, 600 and 700 numbers of damaged packages, respectively. Aim of the experiment was to assess the possibility of identifying packets in group of damaged packages.

The clustering procedure within Algorithm 1 has been implemented basing on a hierarchical approach and using the Ward's method. Table 1 includes results obtained for the proposed approach. Fig. 3 shows an example of dendrogram obtained for considered problem.

Table 1. Results obtained for the proposed approach

Number of damaged packages	The percentage of correctly identified packages	The number of correctly identified packages
500	73%	365
600	57%	342
700	55%	385

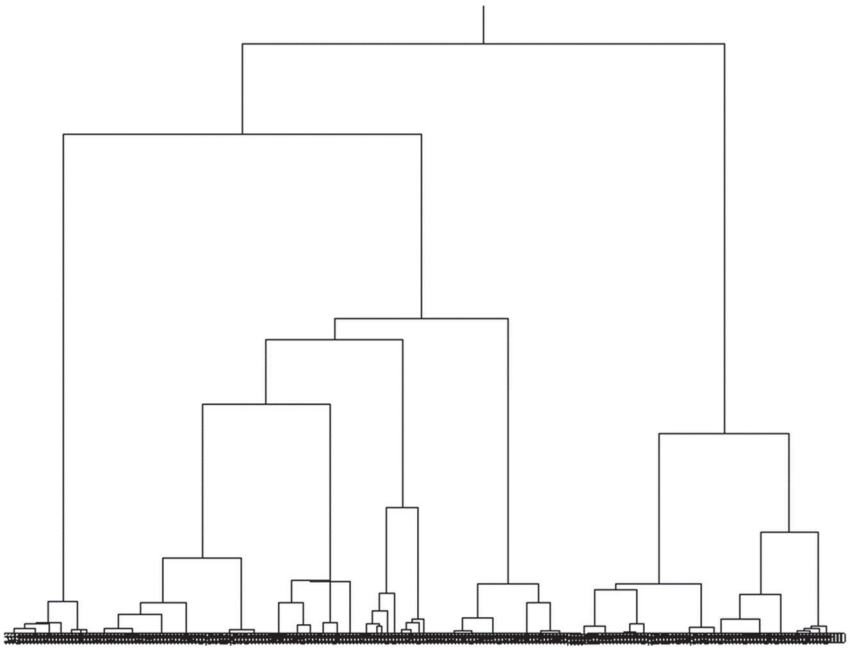

Fig. 3. Dendrogram obtained for considered problem

Based on the results, it can be observed that clustering approach can help to identify packets in the group of damaged packages. Of course we observe that the number of identified packages depend on the number of damaged packages in data. In our experiment, when the number of damaged packages was smaller, the number of identified packages has been bigger.

5. CONCLUSIONS

The paper presents experiments' results where the clustering approach has been used for recovering of damaged packages from data series received by a satellite. Although, the discussed experiment has been carried out basing on artificial data, we can conclude that the proposed approach can be considered as a promising tool for package recovering from data stream onboard of a satellite of SAT-AIS system. Of course, the paper shows initial results, but future research can result in fruitful and valuable conclusions.

Future research will focus on studying the influence of different criteria of defining closeness between packages, as well as different clustering distance measures, on performance and quality of the proposed approach.

ACKNOWLEGEMENTS

The research was founded by Gdynia Maritime University, Poland. Special thanks for Mr Marcin Waraksa from Gdynia Maritime University for sharing raw data, which have been

used in the experiment, and for his valuable consultations in the SAT-AIS domain.

REFERENCES

[1] How to solve your TMI problem: Data science analytics to the rescue, https://searchdatamanagement.techtarget.com/definition/data-analytics [Accessed in March 2019]

[2] Creating value from data in shipping. Practical Guide. DNV, https://www.dnvgl.com/maritime/Creating-Value-from-Data-in-Shipping/index.html [Accessed in March 2019]

[3] Zhou Y., Daamen W., Vellinga T., Hoogendoorn S., AIS data analysis for the impacts of wind and current on ship behavior in straight waterways, Proceedings of the 17th International Congress of the International Maritime Association of the Mediterranean (IMAM 2017), Lisbon, Portugal, 2017, 265-272

[4] Automatic Identification Systems (AIS), IMO, http://www.imo.org/en/OurWork/Safety/Navigation/Pages/AIS.aspx [Accessed in March 2019]

[5] Satellite – Automatic Identification System (SAT-AIS) Overview, https://artes.esa.int/sat-ais/overview [Accessed in March 2019]

[6] Vessel Tracking Pioneer Recalls System's Post-9/11 Origins, https://portal.midatlanticocean.org/ocean-stories/automatic-identification-system-tracking/ [Accessed in June 2019]

[7] Fournier M., Hilliard R.C., Rezaee, S., Pelot R., Past, present, and future of the satellite-based automatic identification system: areas of applications (2004–2016). WMU Journal Maritime Affairs 17(3) 311–345, 2018, DOI: https://doi.org/10.1007/s13437-018-0151-6

[8] Tsou M.C, Discovering Knowledge from AIS database for application in VTS, The Journal of Navigation 63(3) 449–469, 2010, DOI: https://doi.org/10.1017/S0373463310000135

[9] Satellite – Automatic Identification System (SAT-AIS) Overview, https://artes.esa.int/sat-ais/overview [Accessed in June 2019]

[10] Wawrzaszek R., Waraksa M., Kalarus M., Juchnikowski G., Górski T. Detection and Decoding of AIS Navigation Messages by a Low Earth Orbit Satellite. In: Sasiadek J. (eds) Aerospace Robotics III. GeoPlanet: Earth and Planetary Sciences. Springer, Cham, 2019, DOI: https://doi.org/10.1007/978-3-319-94517-0_4

[11] Swetha G.M., Hemavathy K., Natarajan S., Overcome Message Collisions in Satellite Automatic ID Systems, 2019, https://www.mwrf.com/systems/overcome-message-collisions-satellite-automatic-id-systems [Accessed in June 2019]

[12] Prévost R., Coulon M., Bonacci D., LeMaitre J., Millerioux J., Tourneret J., Extended constrained Viterbi algorithm for AIS signals received by satellite, 2012 IEEE First AESS European Conference on Satellite Telecommunications (ESTEL), Rome, 2012, pp. 1-6. DOI: https://doi.org/10.1109/ESTEL.2012.6400111

[13] Seta T., Matsukura H., Aratani T., Tamura K, An estimation method of message receiving probability for a satellite automatic identification system using a binomial distribution model. Scientific Journals of the Maritime University of Szczecin 46(118)

101-107, 2016
[14] Waraksa M, Żurek J., SAT-AIS receiver radio interface load analizies from the satellite side. Przegląd telekomunikacyjny - Wiadomości telekomunikacyjne 6:384-387, 2017, DOI: https://doi.org/10.15199/59.2017.6.52
[15] Wierzchoń S., Kłopotek M.A., Modern Algorithms of Cluster Analysis. Springer International Publishing, Cham, Switzerland, 2018, DOI: https://doi.org/10.1007/978-3-319-69308-8