Port type prediction based on Machine Learning and AIS data analysis

Thomas Charrot, Juliette Guégan, Aldo Napoli, Cyril Ray

Port Type Prediction Based on Machine Learning and AIS Data Analysis

Thomas CHARROT
Naval Academy Research Institute
Brest, France
charrot.eleve@ecole-navale.fr

Juliette GUEGAN
Naval Academy Research Institute
Brest, France
guegan2.eleve@ecole-navale.fr

Aldo NAPOLI
MINES ParisTech–PSL Research University
Sophia Antipolis, France
aldo.napoli@mines-paristech.fr

Cyril RAY
Naval Academy Research Institute
Brest, France
cyril.ray@ecole-navale.fr

***Abstract*—The estimated number of main ports and stationary areas in the world account for almost 25,000 and most of them are not well known. Being able to provide the navigator with information such as the type of surrounding ports in his navigation area is therefore of interest. Automatic Identification System data transmitted by ships is a valuable source of information whose potential can be exploited to give further knowledge on the maritime situation. It is also useful to extract knowledge about ports' activities and types. This paper aims at analysing AIS data using machine learning methods, and more specifically supervised classification to establish a harbour map with the objective of identify port's type especially for the less documented areas of the globe.**

***Index Terms*—AIS data analysis, machine learning, port type classification.**

I. Introduction

Positioning data analysis has proven to be a key element for better knowledge of maritime traffic. During the last decade, many studies have focused on the use of the Automatic Identification System (AIS) data for different purposes such as the understanding of maritime traffic or ships' behaviour. AIS is an automated navigation-aid system designed for the real-time exchange of ships' data through radio communications. Main information broadcast by the system concern positioning (longitude, latitude, speed, course, ...), identity (international id, name, dimension, ship type, ...) and voyage (draught, destination mainly). Some of the leading research topics based on AIS include

- *Vessel trajectory analysis and prediction* [1]
- *Detection, classification, and identification* of ships [2]
- *Traffic forecasting*, like port volume handling and cargo throughput forecast [3]
- *Collision prevention*, analysing risk of ships collisions through the study of their navigational behaviour [4]
- *Anomaly detection*, mostly adopting methods to model normal traffic against which any irregular behaviour is associated to potential threats such as fishing, illegal traffic (human beings, narcotics, goods), piracy [5]
- *Fakes and errors detection*, like ship type errors or positioning fakes [6]

II. Problem Statement

Recent research on maritime mobilities models the interaction between ships and ports in the form of an origin-destination set which aims to facilitate the understanding of mobility. However, ports characteristics do not remain static, can evolve according to the maritime traffic, type of ships, the environment, seasonality, or port structure.

Therefore, one the recurrent feedback from naval officers is that most ports in the world are not well known. The goal of our work is to provide the navigator, but also control centers or navy with valuable information and maps of ports, especially in the least documented areas of the globe where such information is not always easily accessible.

In this paper, we have used two supervised classification methods, KNN (K-Nearest Neighbours) and Random Forest classifier, which, beyond being some of the most commonly used approach, have a relatively simple behaviour to apprehend and implement [2], [7]–[10]. Although other methods have been shown to give good results, those produced by these methods are already proven on many data [7].

III. Supervised classification

Supervised classification is a group of methods from supervised machine learning, which consists in assigning a label to data on which we have measurements in order to define a model using a specific classification method and to make predictions on new data.

Two supervised classification methods have been used in this paper: K-Nearest Neighbours (KNN) and Random Forest. These two methods are presented hereafter.

A. KNN

KNN takes into account the K inputs of the training dataset closest to the new data to be predicted and retains for Y (cf. Figure 1) the most represented output. This algorithm is thus based on a notion of distance and can thus see its accuracy improved if the data are normalised.

KNN has the advantage of being a fairly simple algorithm to learn. Nevertheless, hyper-parameters play a crucial role:

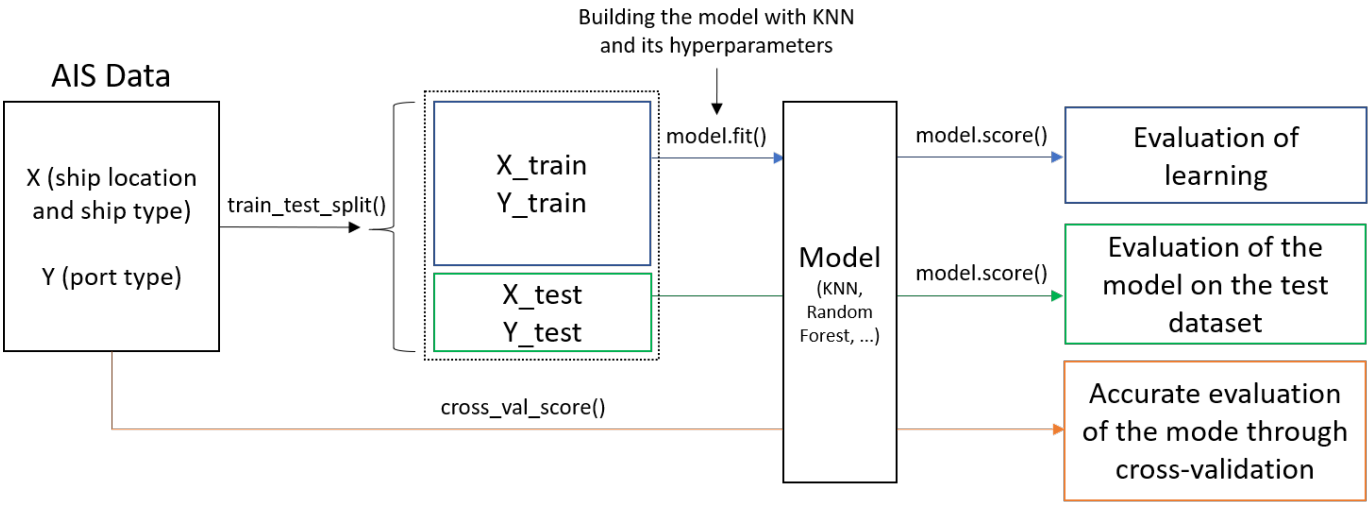

Fig. 1. Schematic representation of a typical implementation of a supervised learning algorithm.

in the case where the prediction is made between two quasi-equivalently represented classes, a change in the number of neighbours or their weights can completely change the result. We can also already highlight the importance of the notion of distance in its implementation. On the other hand, this algorithm tends to be memory intensive and may therefore be limited by the size of the training dataset.

B. Random Forest classifier

This method is based on decision trees. It is a decision process that can be represented as a tree, with branches whose ends (the "leaves") are the possible decisions. The branches are obtained by nodes, each corresponding to a decision between several parameters. The main drawback of the decision tree is that it is very dependent on the data sample used.

To overcome this problem, the Random Forest algorithm consists of the successive application of decision trees. This corresponds to the boosting technique (or bagging: boostrap aggregating): by combining the results of several independent models, the variance is reduced and therefore the prediction error as well. The final prediction of the Random Forest corresponds to the most frequent category returned by the set of decision trees. The Random Forest is a powerful model in prediction problems and is fast to train. It is thus particularly adapted in the case where there is a significant number of explanatory variables (thus of columns in X). Its main drawback is its "readability": the results are not very representative of the approach used.

IV. Learning Process

The dataset used is a public dataset that covers a period of six months and provides around 18.6 million of ship positions collected from 4842 ships over the north-west of France [11]. The aim being to establish a port cartography on which ports would be classified according to ships' locations and shiptype (obtained from message 5 of the AIS), it was necessary to retrieve the coordinates of ports covered by the geographical area of the AIS data. For this purpose, we used the reference port data of 222 Brittany ports including a unique identifier, their name and their geographical coordinates.

Ships having speed below 0.3 knots in a range of 1 nautical mile around each port have been selected for the study. This corresponds to stationary ships, either in a port, or at anchor near the coast. Let us mention that data includes vessels that have multiple shiptypes. Arbitrarily choosing a single type (e.g. the majority type) for these few ships would be a bad option because it generates a bias. Since this situation only concerns a very small number of MMSIs (less than 2% of the ships), we chose to remove the ships associated with several shiptypes.

Classification methods implemented in this work are based on the open-source library Scikit-Learn [12]. The modelling process implemented with the classification methods is summarised in Figure 1.

The classification process consists of 3 phases:

- The training phase: the algorithm is given a data set containing both the explanatory variables X_train and the variables to be predicted Y_train so that it associates the two (train to signify training);
- The test phase: the algorithm is given a set of data (X_test, Y_test) that is known, but which is not part of what the model has learned. The latter then makes its prediction Y from X_test, and then compares it to Y_test in order to evaluate its performance. This testing technique has the advantage of being simple and immediate, but let us note that there are more thorough evaluation techniques, such as cross-validation, which we will discuss later;
- The prediction phase: after having trained and tested algorithms to ensure their reliability, we can submit an input X so that it predicts output Y. Typically, in the context of our work, we give it a set of ship locations so that it can deduce a shiptype associated with each port.

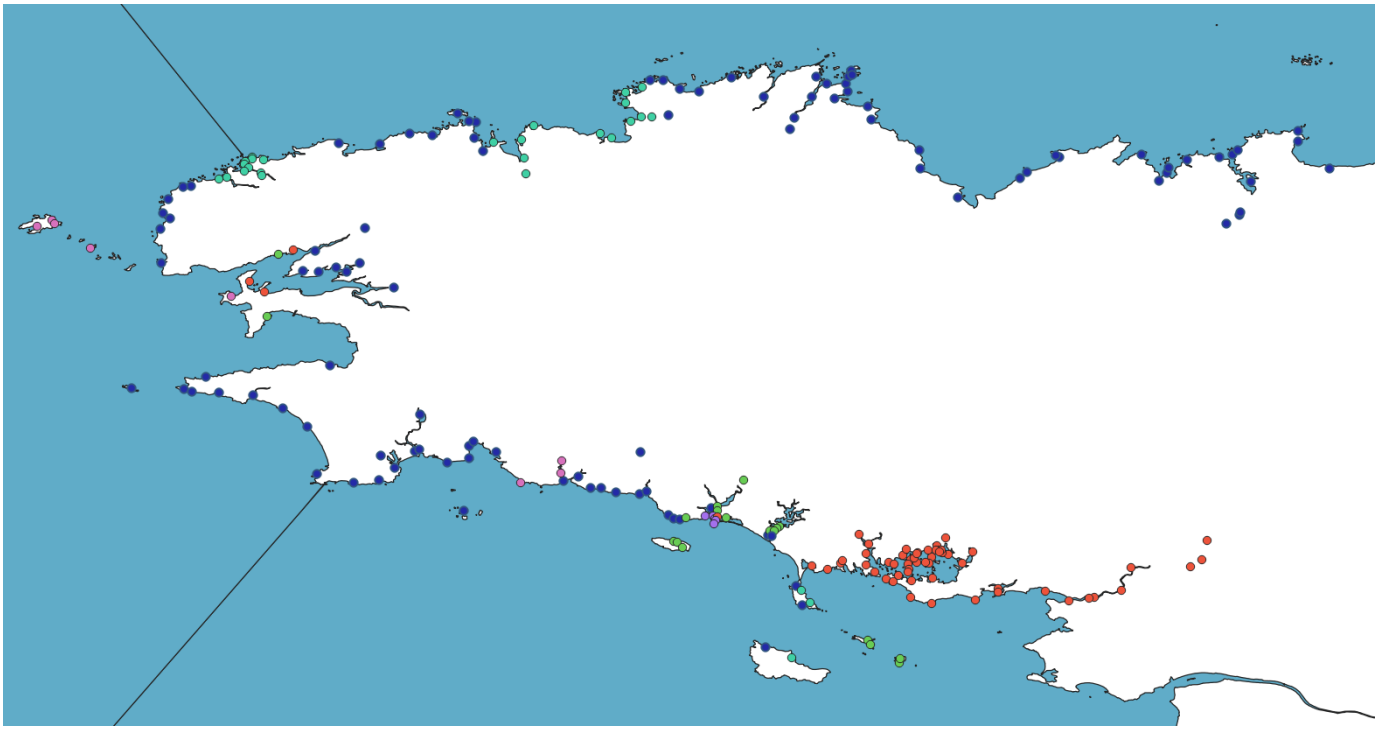

Fig. 2. KNN port type prediction. Each point represent a port. Colours correspond to the port's type prediction (Blue = Fishing; Red = Sailing Vessel; Violet = Pleasure Craft; Pink = Search and Rescue; Cyan = Passenger; Green = Cargo; Orange = Tanker).

The hyperparameters of KNN and Random Forest have been set by the GridSearchCV function of Scikit-Learn. Cross-validation was used to evaluate the models. 222 real ports were classified. An analysis of the confusion matrix was also carried out to visualise and compare the numbers of correct and erroneous predictions and thus determine the best predicted port type by the model.

V. Results and discussion

The results of the modelling are presented below. The 7-class port type prediction maps (Fishing, Sailing Vessel, Pleasure Craft, Search and Rescue, Passenger, Cargo, Tanker) of each model are proposed and compared.

A. Results with KNN

Scores obtained with KNN are greater than 0.99 in learning and testing, and greater than 0.94 in cross validation. This reflects a very efficient model with the data used. In addition, the confusion matrix shows very few prediction errors.

The main classification error concerns 42 pleasure boats which are classified as fishing vessels; this represents 1.73% of this ship type. The second most important error concerns 16 fishing vessels classified as pleasure craft, that is 0.11% of fishing vessels. These errors arise from a certain proximity between these two ship types: most of fishing vessels come from small ports located all along the coast, in which one can also find sails yachts. Finally, the results are very good for all the other ship types. This includes among others the cargo ships and tankers which, given the similarity between these vessels, showed very few prediction errors. Figure 2 shows an example of maps.

B. Results with Random Forest

The results obtained with Random Forest allow the model to be validated. Nevertheless, there is less confusion between fishermen and boaters: in our case study, 21 fishermen are classified as pleasure craft (0.14% of the total number of fishermen), and 25 for the opposite situation (i.e., 1.03% of the pleasure boaters). Although the error rate for pleasure crafts is higher than for KNN, the fact remains that slightly fewer errors are made. Finally, as for KNN, the other ship type predictions perform very well. Figure 3 shows the port type prediction map for the 222 ports in Brittany.

C. Discussion

Differences can be observed between the map established by the KNN (Figure 2) and the one given by the Random Forest (Figure 3), especially on groups of ports, such as in the North of Brittany, in the western part of Brittany near the "Phare du Four", or in the South of Brittany, above Lorient city. These can be easily explained by the low presence of data in these areas: if during the training phase, very little data is located in this area, then inevitably the prediction may vary depending on the method used. Typically, KNN will tend to classify the port as its neighbour, while Random Forest may see other criteria and predict a different classification. Variations in more usual ports are more difficult to explain.

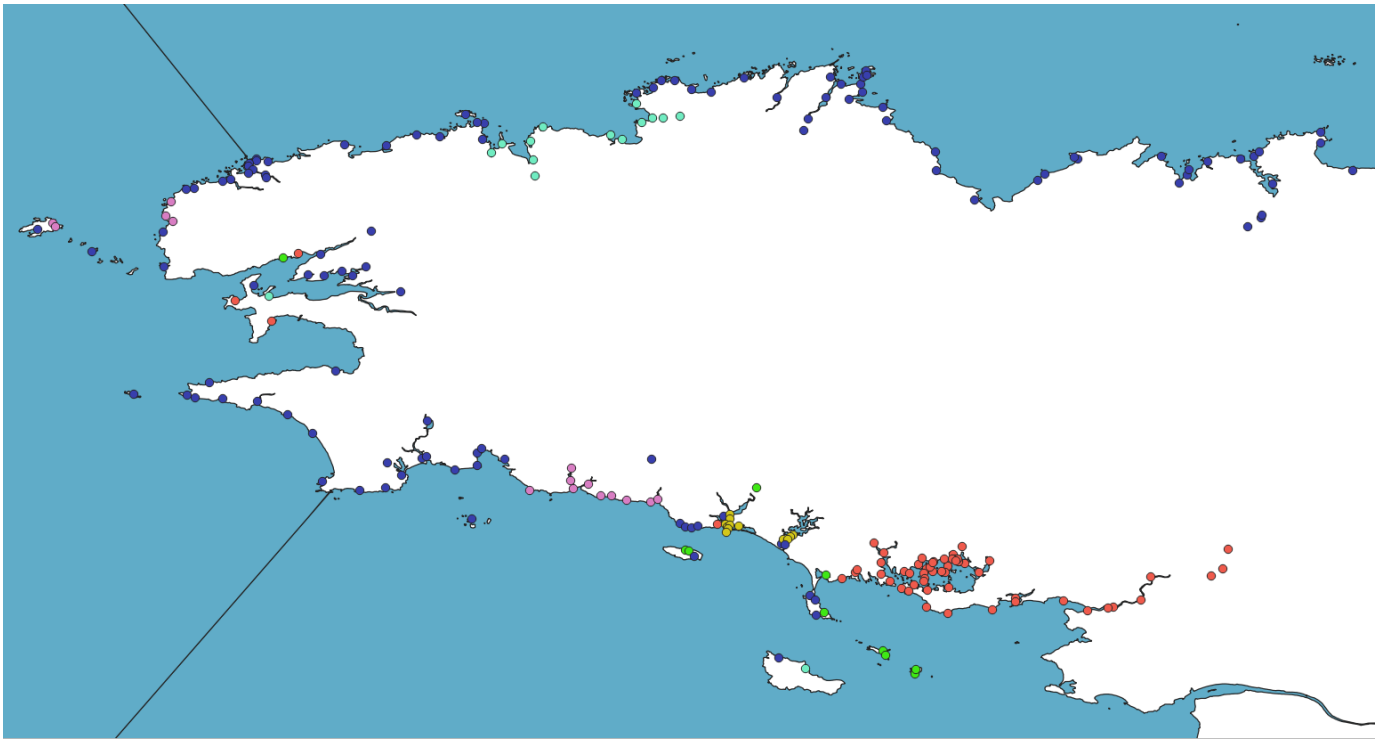

Fig. 3. Random Forest, Port type prediction. Each point represent a port. Colours correspond to the port's type prediction (Blue = Fishing; Red = Sailing Vessel; Violet = Pleasure Craft; Pink = Search and Rescue; Cyan = Passenger; Green = Cargo; Orange = Tanker).

With a good knowledge of all the ports of Brittany listed, we could correlate each result obtained to see which model is most often correct. Nevertheless, this approach requires at least a consistent documentation on the ports of the region. This seems feasible in the case of Brittany, but we should not forget that our model is intended to be used in other areas, which will not necessarily be documented.

Moreover, one of the objectives of our approach is to be free of documentation. We must therefore rely on the numbers returned by the evaluation of the models. In this case, the KNN obtains a cross-validation score that is almost 0.018% higher than that of the Random Forest. Although this difference is small, we believe that KNN is the preferred model. It also has the advantage of a much lower computation time than Random Forest with its 100 trees to run.

VI. Conclusions

This paper presents an analysis of ports based on machine learning and AIS data. The designed models based on KNN and Random Forest classification, allows for the determination of (1) areas where vessels are stationary (port, anchorage areas) and (2) the associated type of port (fishing, trade, etc.). While results, validated with marine officers, are satisfactory, current works consider other methods such as SVM or neural networks.

References

[1] P. Borkowski, "The ship movement trajectory prediction algorithm using navigational data fusion," *Sensors (Basel, Switzerland)*, vol. 17, 2017.

[2] N. Damastuti, A. Siti Aisjah, and A. A. Masroeri, "Classification of ship-based automatic identification systems using k-nearest neighbors," in *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2019, pp. 331–335.

[3] A. Jugovic, S. Hess, and T. Poletan Jugović, "Traffic demand forecasting for port services," *PROMET - Traffic and Transportation*, vol. 23, 01 2011.

[4] T. A. Johansen, A. Cristofaro, and T. Perez, "Ship collision avoidance using scenario-based model predictive control," *IFAC-PapersOnLine*, vol. 49, no. 23, pp. 14–21, 2016, 10th IFAC Conference on Control Applications in Marine SystemsCAMS 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405896316319024

[5] M. Riveiro, G. Pallotta, and M. Vespe, "Maritime anomaly detection: A review," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 5, p. e1266, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.1266

[6] C. Iphar, A. Napoli, and C. Ray, "An expert-based method for the risk assessment of anomalous maritime transportation data," *Applied Ocean Research*, vol. 104, p. 102337, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0141118720304314

[7] J. Maillo, S. Ramírez, I. Triguero, and F. Herrera, "knn-is: An iterative spark-based design of the k-nearest neighbors classifier for big data," *Knowledge-Based Systems*, vol. 117, pp. 3–15, 2017, volume, Variety and Velocity in Data Science. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705116301757

[8] A. B. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, "Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach," *CoRR*, vol. abs/1409.0919, 2014. [Online]. Available: http://arxiv.org/abs/1409.0919

[9] D. R. Cutler, T. C. Edwards Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-0539.1

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1010933404324

[11] C. Ray, R. Dréo, E. Camossi, A.-L. Jousselme, and C. Iphar, "Heterogeneous integrated dataset for maritime

intelligence, surveillance, and reconnaissance," *Data in Brief*, vol. 25, p. 104141, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340919304950

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html