AIS trajectory classification based on IMM data

Daniel Amigo Herrero, David Sánchez Pedroche, Jesús García Herrero, José Manuel Molina López
Group GIAA, University Carlos III of Madrid, Spain
Email: { damigo, davsanch, jgherrer } @inf.uc3m.es, molina@ia.uc3m.es

***Abstract*— The importance of the maritime vehicles makes necessary the implementation of systems capable of ensure the safety and security. This paper presents an analysis on Automatic Identification System (AIS) data processed with Interacting Multiple Model (IMM) filter in order to help trajectory data analysis for predictive tasks. The main objective is building a system capable of classifying ships trajectories into different categories as the ship type or the type of activity (fishing, under way with engines, etc.) based on the kinematic and other filter outputs. An automated processing system is implemented to use raw AIS data, preparing and organizing it in order to classify them in ship types and maneuvering state. The appropriate modelling with dynamic models and transition probabilities allow the identification of patterns helpful for trajectory reconstruction and classification. Important aspects as data cleaning, processes parallelization and parameter analysis are dealt on the paper, with results obtained from an available data set.**

***Index Terms*— AIS, anomaly detection, classification, data mining, IMM filter, maritime surveillance, trajectory reconstruction**

I. INTRODUCTION

The maritime traffic is an indispensable element for our actual society since covers a wide range of activities such as the fishing industry, passenger transport, trade of goods and entertainment. But between maritime vehicles there are others that make illegal activities like piracy, drug smuggling, contraband or illegal fishing.

The importance of the maritime traffic makes indispensable the control of these kind of vehicles to ensure the safety and security of those that are good for the society and to prevent the illegal activities. Therefore, it is needed a continuous development and improvement of maritime monitoring systems, especially decision support systems for ports and coast authorities in order to avoid congestions and at the same time maximize safety and security [1].

Currently a large amount of technologies allows the tracking and localization of maritime vehicles all around the globe, like AIS plots, but it is necessary to recollect and process all the data collected by such technologies to obtain information that could be useful for the maritime surveillance systems.

However, the large amount of information makes impossible for a person to consider all the possible information to decide. Therefore, systems of these characteristics could use artificial intelligence (AI) techniques to obtain useful information that could help to improve the results in tasks as the surveillance. Between that AI techniques, the classification process allows us to make prediction algorithms that could use some input variables to obtain some output information. In the case of the maritime surveillance can use all the movements of a ship to predict things like their destination, their actual maneuver or their type of ship. Because of this, use that new information to inform the system operator about atypical scenarios and allow the fast reaction to the possible dangers that the surveillance systems seek to avoid. Besides, building a unified and reliable picture for maritime surveillance requires contextual reasoning about the observations to relate them with situation elements and previous domain knowledge [2]. In general, context is used for several key tasks such as explaining observations according to the situation, constrain the processes or refine the estimations [3].

In this work, the analysis of trajectory data for automatic classification can be labelled as context learning process, providing knowledge useful to describe the behavior of ships accordingly to areas, routes, types of operations, which could be used to reason about particular trajectories. Context can be extracted from recorded data with appropriate training/learning procedures to build models useful to describe patterns, predict trajectories or identify anomalous behaviors [4].

The main objective of this study aims to the capability to create a classifier capable of read ships trajectories as an input, with the objective of classify them into different categories as the ship type or the type of behavior of the ship (the maneuver that the ship does in a specific trajectories).

This classifier could work as a subsystem in other systems helping to control anomalies detection, for example a cargo ship doing fishing movements could be illegal fishing.

This classifier uses real world data which implies a lot of noise and atypical information that could affect the training of the classifier and in consequence the future class predictions.

That is why is needed an implementation of a system capable of preprocessing raw data, to clean it and to extract useful information that the classifier can use in its training and its future predictions. The system is going to use an IMM filter stage to reduce the atypical kinematics and to add some new information extracted from its dynamic models and transition probabilities. Also, is necessary to implement a data clean step to reduce the misinformation and a step of trajectories segmentation to make them comparable in the classifier.

It is remarkable that a construction of a system in modules makes it so much configurable, being possible to create, for example, another filtering module and replace it for the actual IMM filter without modify the other modules. That also open the possibility to not use some of the models without needing to modify the other (except for the file reading).

This paper is organized as follows: In section II the state-of-art methods in classification of maritime vehicles tracks are analyzed. In section III is explained the source of the data used for the investigation and in section IV is related the implemented system and in V results of the work are shown.

Finally, the conclusions and perspectives for future works are presented in section VI.

II. Related work

Due to the ubiquity of AIS-equipped ships worldwide, there are numerous studies on trajectory data, specifically in the area of maritime surveillance there are previous works addressing the search of traffic patterns to enhance Situational Awareness in maritime domain, especially to organize (cluster), reconstruct and classify trajectories, including prediction of activities and anomaly detection.

The work [5] processes AIS messages with deep learning framework (recurrent neural networks with latent variables) to address real aspects such as noisy data and irregular time-sampling for tasks of trajectory reconstruction, anomaly detection and vessel type identification.

The work in [6] has a proposal for a representation of routes as spatial grids built with AIS data to model the navigational patterns. It is extended in [7] to perform trajectory classification and anomaly detection in a system named as Traffic Route Extraction and Anomaly Detection (TREAD) based on extraction of frequent routs to classify real-time trajectories and trigger anomaly detection.

A survey of techniques proposed for mining trajectory data in multiple domains is provided in [8], focusing on data preparation, preprocessing, management and mining tasks (pattern mining, outlier detection, and trajectory classification), while a specific survey of maritime anomaly detection is provided in [9], distinguishing available data, methods, systems and user aspects. In [10], an analysis of AIS trajectory clustering is presented, with appropriate distance measures and dimensionality reduction. Aspects related with efficiency and scalability are dealt in [11], with data organization based on quad trees and modelling with Gaussian Mixtures.

This work proposes the use of computed dynamic parameters in order to perform trajectory mining tasks based on the available information. The proposed system is a multiple-mode filter which several dynamic models in parallel and transition probabilities. In [12] a similar work was applied to segment classify missiles head and debris based on their dynamics.

III. Data source

This study started with a reliable dataset from which obtain enough information to create classifiers that could learn correctly and provide new useful information in maritime traffic surveillance. For this task, a fundamental element would be obtaining pre-labeled data to use supervised classification algorithms, giving the classifier a class to learn how to classify it. Additionally, the operation of the IMM filter to be implemented requires the existence of a measurement of the position in each instant with appropriate refresh rate and precision, so that the algorithm can filter it and generate new reliable information for the classifier.

With these characteristics in mind, and after evaluating some AIS free data sources, the AIS data repository provided by Danish Maritime Authority [13] was selected, in which millions of raw AIS contacts are available every day from 2006 to the present. The contacts of the ships are detected on the coast of Denmark and in the surrounding areas, generating files of approximately 1.8 GB per day.

This source provides a practically unlimited amount of information to generate a useful dataset for the objective of the study. Also, in the raw AIS plots, is possible to identify two possible classes for the subsequent classification process. The first of them is a "ship type" category that could be useful to obtain information about the ship based on data about its movement. The second one is the "navigational status" category which is useful to classify the type of maneuver that a ship is doing in a specific moment.

IV. Implemented system

The characteristics of the selected database implies the necessity of a very strong preprocessing of the raw data to clean and extract the information that could be useful for the classification, removing the irrelevant data for this task.

Figure 1 shows an overview of the whole implemented system and the different processing steps.

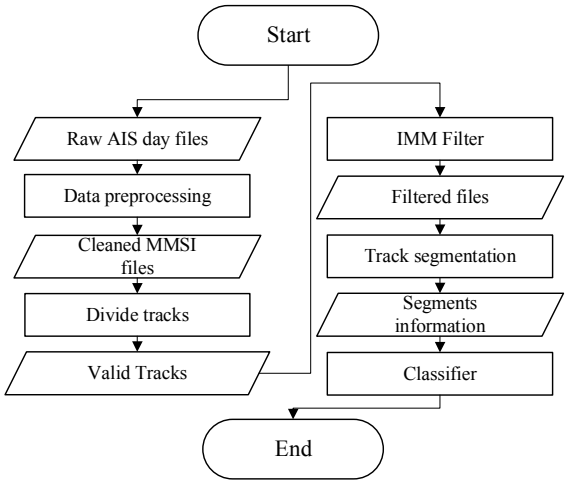

Figure 1. System overview

This approach uses an Interacting Multiple Model filter for its well proven performance in problems of trajectory processing and smoothing. In other words, it provides the capability of reduce the influence of atypical plots and adapt to maneuvers into this classification problems.

One of the most interesting aspects of the utilization of the IMM algorithm is the information that provides the capability of detection of the movement variations. This is achieved by one of the principal characteristics of this filter that it's the variation between different prediction modes that works internally in the filter.

The use of this filtering algorithm introduces the counterpart of needing a stable trajectory to process, being difficult to include in the classification stage trajectories that have big time gaps that could confuse the filter (due to fast increase of prediction covariances). To solve this counterpart, is needed another stage in the processing to divide all the movement of a ship in different trajectories that is possible to process with the filter without those problems.

Other condition needed for the classification is a set of data

without useless information (as data without classes or only a few plots) or even worse misinformed data (as could be a ship that does not move accordingly to the declared maneuver class).

So, is necessary a stage to clean the data and remove intervals that do not have all the variables that are needed or have some errors in its information.

The final step necessary for a correct classification is the creation of comparable inputs. For this step, is needed to make trajectories that could have thousands of plots comparable to those that maybe only have 100 plots. To achieve this, the proposed solution divides each trajectory into equal size segments making each segment an instance for the classification.

To characterize the information of all the segments, some statistical variables were selected to summarize the knowledge contained in every segment, in conjunction of the statistical information is viable to use the information of the IMM to improve the extracted knowledge.

It should be underlined that the size of the segments are needed big enough to be considered as a trajectory on its own (is going to be the minimum trajectory size in the previous stage of the IMM) but also smaller to make the statistics more importance (if the segment is bigger it could have more types of different movements that could misinform the statistics calculations).

The details of the implemented system are presented in the following sections.

A. *Data preprocessing*

Starting with a file with all the measured plots of one day, around ten million of AIS plots, the first step consists in dividing the file in multiple files, each of one containing the information of one specific ship discretized by the MMSI (Maritime Mobile Service Identity).

On these files divided by the MMSI some static information is replicated, since different AIS emitters could not send all the information. Specifically, the replicated information is the type of ship and its dimensions (width and length of the ship), which are variables that may be interesting for the subsequent classification (the type of ship to be a class and the dimensions for being able to provide useful information).

Lastly, once every MMSI file have all their data correctly replicated, some MMSI files are discarded depending on them usefulness for these concreted objectives.

All the files are separated first corresponding to static base stations, which are not consider useful for this study since they would introduce noise in the classification. Then, the remaining files are divided according to the difference between the two defined classes, if each file does not contain any of the classes, it contains only one of them or contains both classes in all the plots. If the file contains both classes but some of the maneuver instances are undefined the process is going to split the file by keeping to the next stage those plots with both classes and extracting the plots without all the classes to the “only ship” group.

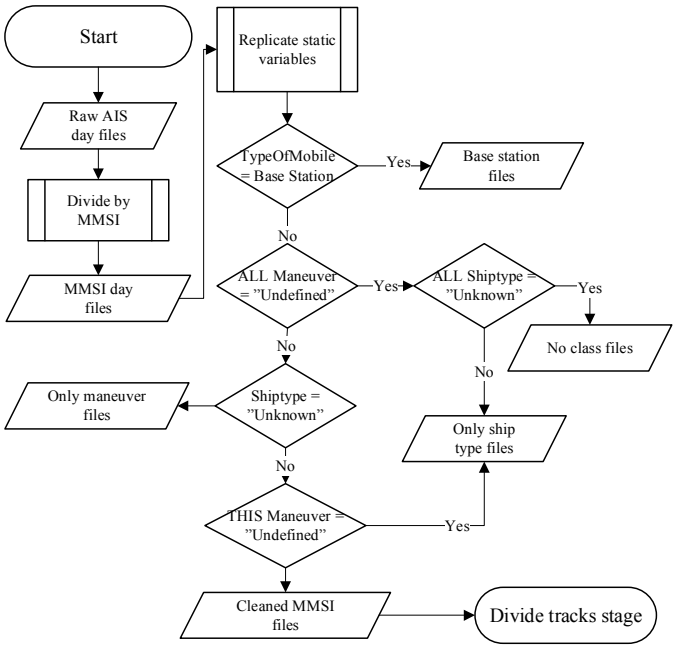

Figure 2. Data preprocessing

Figure 2 shows a diagram to explain all the process of this stage. Once the cleaning each file it could be used in the next step where each file can be divided into different tracks.

B. *Track division*

All the tracks of a ship in a day are not a treatable dataset because, being real information, they are not prepared to enter a filter. This real dataset provides movements that could be noisy in some parts and contain time gaps because of the range of each receiver and the different maneuvers a ship could perform.

Therefore, the solution for this task is to divide each trajectory into different tracks, each of one containing a minimum number of sequential plots with a maximum time gap between them.

The constraint to have a minimum number of sequential plots is to ensure a minimum size for the trajectories, while the maximum time gap is due to prevent possible outlier values due to abnormal extrapolation intervals.

Also, this sequence must be of the same maneuver type to use it correctly at the classification stage. Although trajectories with different maneuvers could imply a maneuver change trajectory, them are considered as two trajectories of one maneuver to simplify the classification problem.

Figure 4 is an example of the division of trajectories using the maximum time gap, and Figure 4 has another example using the minimum 50 consecutives with position variation.

With this definition of trajectory in mind the algorithm implemented is shown in Figure 3.

First, the cleaned files are prepared assessing the plots separation time dividing into different sequential groups if the time gap is over a certain threshold.

The next step tests the maneuver of the sequential plots and makes a division if there are more than one maneuvers.

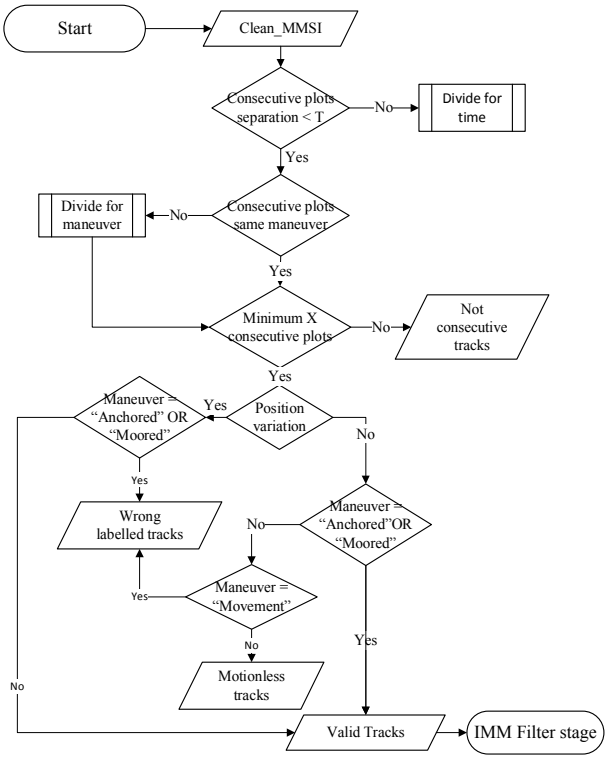

Figure 3. Track division

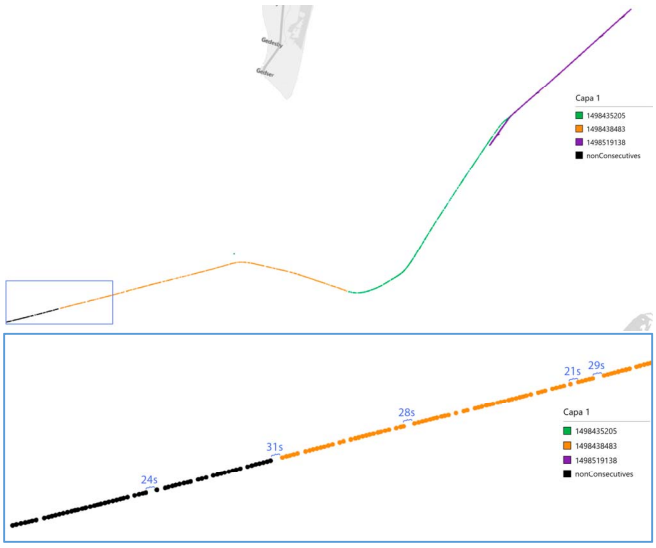

Figure 4. Example of the division of trajectories using maximum time gap

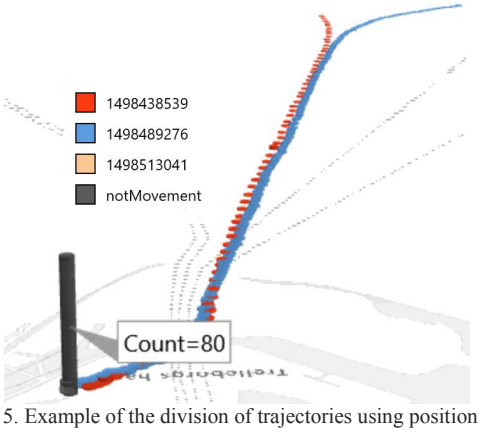

Figure 5. Example of the division of trajectories using position variation

Finally, the algorithm makes a validation of the resulting trajectories, discarding the ones that does not have enough plots by sending them into their own directory. Those that have the same maneuver with enough sequential plots are tested for the movement of the ship, if there is variation between the position or if the maneuver is "anchored" the algorithm introduce them into a directory of valid tracks for the next step.

The trajectories that does not have movement are separated because a movement maneuver that does not vary its position is due to a repetition of plots or noisy data. In other words, an outlier that could make the classification algorithm to work wrongly. In this case, with a study of the plots, is decided to use limits of minimum 50 plots with a maximum time gap of 30 seconds as a suboptimal solution.

With the trajectories obtained as a result of these algorithm is possible to proceed to the next stage, where the IMM filter is going to process these trajectories to obtain some new information and to smooth the possible outliers in the position measurement.

C. *Interacting Multiple Model (IMM)*

The IMM filter is an adaptative tracking technique which provides a satisfactory tradeoff between the complexity of the algorithm and its performance. It needs an accurate model of the sensor's performance and the target dynamics (movement pattern). This filter is used to combine multiple filter models (*N=2* models for this paper) into one state estimation based on the estimated probabilities (μ_j), the estimated state vectors ($\hat{x}_j$) and their corresponding covariance matrices (P_j).

$$\hat{x}[k] = \sum_{j=1}^{N=2} \mu_j[k] \cdot \hat{x}_j[k] \tag{1}$$

$$\begin{cases} P[k] = \sum_{j=1}^{N=2} \mu_j[k] \cdot P_j[k] + X \\ X = \sum_{j=1}^{N=2} \mu_j[k] \cdot \left(\hat{x}_j[k] - \hat{x}[k]\right) \cdot \left(\hat{x}_j[k] - \hat{x}[k]\right)^t \end{cases} \tag{2}$$

Each mode of the combined filter is defined to represent different movement patterns. In this case are selected two models, one to represent linear movement model and a second one to represent the target maneuvers. Both models use an Extended Kalman Filter and its difference comes from the Q matrix at the prediction equations.

$$\begin{cases} \hat{x}_{jp}[k] = F_j(\Delta t) \cdot \hat{x}_{0j}[k-1] \\ P_{jp}[k] = F_j(\Delta t) \cdot P_{0j}[k-1] \cdot \left(F_j(\Delta t)\right)^t + Q_j(\Delta t) \\ \mathrm{j} = 1, 2, \ldots, \mathrm{N} \end{cases} \tag{3}$$

Where F_j is the transition matrix, Q_j is the covariance matrix both for the mode j and Δt is the time passed since the last contact. As a result of the prediction, the probabilities of transition between modes could be updated to apply them in the explained combination of models.

$$S_j[k] = R[k] + H P_{jp}[k] H^t \tag{4}$$

$$\alpha_j[k] = \left(z[k] - H\hat{x}_{jp}[k]\right)^t S_j[k]^{-1} \left(z[k] - H\hat{x}_{jp}[k]\right) \tag{5}$$

$$\Lambda_j[k] \equiv \left|2\pi S_j[k]\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\alpha_j[k]\right) \tag{6}$$

$$\Lambda[k] = \sum_{j=1}^{N} \Lambda_j[k]\, \mu_j{}^-[k] \tag{7}$$

$$\mu_j[k] = \frac{\Lambda_j[k]\mu_j^-[k]}{\Lambda[k]} \quad (8)$$

Where the H matrix is the output matrix, the matrix R is the covariance matrix of the measurement noise, $z[k]$ is the observation and $S_j[k]$ is the covariance matrix of the innovation in mode j at time k.

D. Trajectories segmentation

Once the filtering of each trajectory is finished, is necessary to make a segmentation of trajectories in blocks of the same size, with the objective of making comparable all the inputs of the classification. Besides, that segmentation allows the exploration of sub-trajectory patterns that can't be extracted from an entire trajectory.

This stage also works as a preclassification step to clean the instances. An analysis of the classes values shows that there are instances with remaining empty values (" ", "-" or "Other").

For the maneuver class, there are some instances with the value "Reserved for future amendment [HSC]" that also is considered as a non-instanced value and is consequently cleaned. Finally, there are some minority classes with less than 0.0005% instances. Instances with that few quantities are considered useless for the classifier training and consequently the processing step also makes a clean of those values.

As a final step, the system is going to subtract the useful information of each segment to make it usable by the classification. To avoid making a lot of information subtractions with each classifier test this step will take all the possible information in one execution and then each version of the classifier uses the information to test as an input.

It is important to introduce inputs that could represent all the information of a segment and not only specific plots, that is why the selected variables use statistical values:

Table 1. Statistical values

Class	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE
Average	Mode	Standard deviation
Maximum	Minimum	3 quartiles

Specifically, to represent the movement information the selected variables are:

- Speed of the target.
- Speed variation within the segment.
- The length of the movement.
- The heading variation.
- The time duration of the movement.

The IMM filter has the capacity of locating the activation of maneuvers through the changes between models. In other words, is possible to use the mode probabilities (μ_j) value to track what type of movement is doing the target.

Based on the mode probabilities values the algorithm subdivides the segment into five categories accordingly to the following descriptors:

- Descriptor 1: Linear movement probability over 0.9 (the other one less than 0.1).
- Descriptor 2: Linear movement probability between 0.9 and 0.6 (the other mode between 0.1 and 0.4)
- Descriptor 3: Both probabilities between 0.6 and 0.4.
- Descriptor 4: Maneuver movement probability between 0.9 and 0.6 (the other one between 0.1 and 0.4).
- Descriptor 5: Maneuver mode probability over 0.9 (the other one less than 0.1).

Also, there is a global descriptor that englobes all the information of each segment.

With this information one of the input variables would be the average speed of the target in descriptor 1, meaning the average speed of the plots with a linear movement probability of 0.9.

Finally, with the definition of this descriptors is possible to explore the temporal changes between them along the segment:

- Count of the times when the track stays in the same descriptor.
- Count of the changes between two specific descriptors.
- Count of the times in which the track enters a descriptor.
- Count of the times in which the track enters a descriptor plus the moments in which the track stays in that descriptors.

These would be another type of input variable to the classification stage, giving information as the number of times the target changes from the descriptor 1 to the descriptor 2, meaning the start of a maneuver movement.

E. Classification problem

The approach proposed is the analysis of the best representational input variables useful for the classification problem. This analysis will allow improvements in future works based on a preprocessing environment useful for many data mining strategies.

The implemented algorithm for this study is a multiclass binary decision tree, trained with the 70% of data and tested with the other 30% [14].

As a result of the classification a decision tree is obtained, and it is useful to predict future data inputs, and its output can be used to prove its good or bad classification results.

V. EXPERIMENTS AND RESULTS

The implemented system has been developed using MATLAB R2018b version. In addition, to process such large data sets, the performance has been improved by applying a parallelization that takes advantage of the 16 cores of the experiment server.

The first approach of the study was a Principal Component Analysis (PCA) with the objective of a dimensionality reduction of the data. However, the results provided by this analysis were not satisfactory, so the analysis focused in alternative approaches for variable selection

In the next sections the results of the experiments are shown.

A. Input analysis

The final analysis presented in this paper consists in a variation of all the input variables of the classifier algorithm to find the best variables to represent this problem.

The default input variable data set has 2 different classes (ship type and maneuver) which possible values are presented in Table 2 and the following 283 input attributes:

- 2 comes from the ship dimensions.

- 240 comes from applying the 8 statistical values to each of the 5 variables over each one of the 6 descriptors (including the global).
- 1 comes from the count of times the stays in the same descriptor.
- 5 of enter each descriptor.
- 5 of exit each descriptor, and another 5 considering also the permanency in the same descriptor.
- 25 of the counts of changes between descriptor.

Table 2. Classes values

Ship type class		
Anti-pollution	Cargo	Dredging
Fishing	HSC	Law enforcement
Medical	Military	Passenger
Pilot	Pleasure	Port tender
Reserved	Sailing	SAR
Tanker	Towing	Towing long/wide
Tug		
Maneuver class		
At anchor	Constrained by her draught	Engaged in fishing
Moored	Restricted maneuverability	Under way sailing
Under way using engine		

To analyze these variables, the next cases of study are proposed according to the used variables.

- Case 0: Using all the variables except the ship dimensions. The IMM filter configuration uses an error measurement (R) of 10 and the first switch probabilities (μ_j^-) of the Figure 8.
- Case 1: Case 0 with the inclusion of the ship dimension in the classification.
- Case 2: Modifying IMM filter configuration. 9 configurations are performed including case 0, generated by the combinations of the error measurement (R) between 1, 5 and 10; and the three switch probabilities (μ_j^-) presented in Figure 8.
- Case 3: Modification of case 0 excluding the statistical values of all the subdivisions corresponding to IMM descriptors, leaving only the information of the global descriptor for the whole segment.
- Case 4: Use the case 0 configuration to process only the trajectories that stay in the first descriptor less than 90% of plots and those that stay less than 80% of the plots.
- Case 5: Exclusion from case 0 of each statistical value.
- Case 6: An exclusion of each variable of movement information (speed, speed variation…).

$$\begin{pmatrix} 0.999 & 0.001 \\ 0.01 & 0.99 \end{pmatrix} \quad \begin{pmatrix} 0.99 & 0.01 \\ 0.1 & 0.9 \end{pmatrix} \quad \begin{pmatrix} 0.95 & 0.05 \\ 0.1 & 0.9 \end{pmatrix}$$

Figure 8. Variations of switch probabilities

To explain the classifier results is needed to introduce some results examples of the IMM filter. Figure 6 presents a trajectory that a ship type "law enforcement", composed of 3 segments of 50 plots. Also, it could be observed how the IMM mode probabilities commute with the curse change and some velocity variations. Groundspeed and IMM probabilities are preprocessed as discussed in previous sections to prepare the input for the classifier. Figure 7 provides an example of other class, "Engaged in fishing", where it can be seen how the velocity variates constantly commuting the IMM filter modes although there are not direction changes.

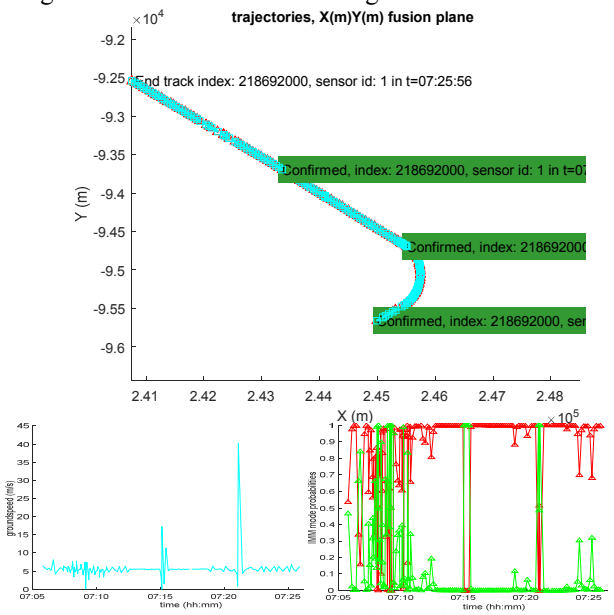

Figure 6. Ship type example of law enforcement

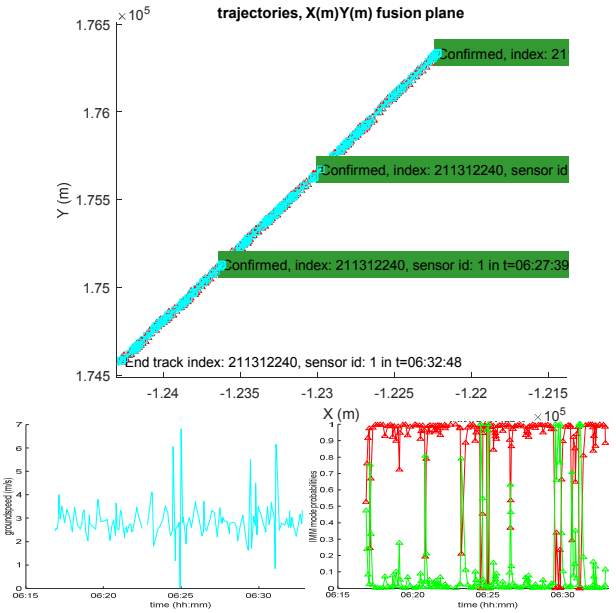

Figure 7. Maneuver example of engaged in fishing

B. *Files statistics*

The experiment starts with three days of AIS plots in June 2017. In total, they consist of more than 32 million AIS plots spread over 13833 different MMSI. After applying the data preprocessing, 1.7% of the total MMSI codes were lost because they are base stations. Furthermore, 22% of plots are removed from the usable dataset because they do not correctly define any of the classification classes. At the end, in the useful dataset are approximately 74% of the original plots. The next step is the track division, discarding 13.3% of the 24 million plots in the dataset, due to time-sequence constraints, corresponding to 39089 different trajectories.

Finally, once the IMM filter is processed and divided all the filtered trajectories into segments of the same size, a total of 19907806 segments are obtained, bringing with it a loss of 4.5 percent of the tracks in this last step. This useful dataset was the final for the classifier.

C. Classification results

The results of case 0 (Table 3) are used in each of the following cases as comparison. Each result table contains the success rate of the test dataset, measured by the results of each trajectory segments by its own and the result of applying a mode operation to all segments classes in order to obtain the majority predicted class along the whole trajectory.

Table 3. Results of Case 0

Class	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE
Ship type	35.4125%	42.5948%
Maneuver	69.1483%	65.6332%

1) Case 1

Is remarkable that including the ship dimensions into the classification has a dramatic impact over the results, especially in ship category. However, these variables are very dependent of the AIS emitter so only have values in some of the plots (in the pre-processing the values are replicated to enlarge the data with those values) making difficult their usability in a real system that process real time data to the maritime surveillance.

Also, the problem approach deals with the use of the type of movement to try the classification and the dimension variables will not imply classifying with that movement (see Table 4).

Table 4. Case 1: Use of ship dimensions to classify

Class	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE
Ship type	92.9957%	88.7195%
Maneuver	93.5869%	88.6179%

2) Case 2

Table 5 shows the results of this case for the ship type class whereas the

Table 6 presents the results with the navigational status class. These results are quite similar although the most consistent results are for the sigma 10 in the first table and for the 5th and 9th rows of the second table.

Table 5. Case 2: Ship type class results by IMM filter variations

AIS Error (meters)	SP 1to1	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE
1	0.99	34.7964%	42.3086%
1	0.999	34.8216%	41.8714%
5	0.95	34.9306%	42.412%
5	0.99	35.2317%	42.1417%
5	0.999	35.1047%	42.2848%
10	0.95	35.673%	42.9287%
10	0.99	35.2129%	42.6187%
10	0.999	35.4125%	42.5948%

Table 6. Case 2: Maneuver class results by IMM filter variations

AIS Error (meters)	SP 1TO1	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE
1	0.95	68.2159%	65.665%
1	0.99	68.1162%	65.2953%
1	0.999	68.2862%	65.5736%
5	0.95	68.4315%	65.5934%
5	0.99	68.4137%	66.1579%
5	0.999	68.2951%	65.1443%
10	0.95	68.4733%	65.0727%
10	0.99	68.7097%	65.4861%
10	0.999	69.1483%	65.6332%

That results could be also combined with an observation of all the made classification test, the best classification results we shown that the AIS error of 10 is clearly predominant while the switch probabilities of 0.999 also appears in most of the cases followed by the 0.99 probability.

3) Case 3

Table 7 present the results of this case which shows little improvement in some aspects to the case 0 but with worst results in the trajectory success rate.

Table 7. Case 3: Only global descriptor comparation

Class	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE
Ship type	35.9279%	43.191%
Maneuver	68.9883%	65.5815%

4) Case 4

The reduction of trajectories to only those with descriptor commutation implies a considerable degradation in the classifier results as the static maneuvers are an important part of instances that are well classified.

5) Case 5

The difference of removing statistical values is not so significative because the information could be extracted from another variable. The most relevant statistical value is the inclusion of the quartiles because it affects both classes, but the average value for the maneuver class and the maximum value for the ship type class have also considerable losses in their classifier results (see Table 8).

Table 8. Case 5: Statistical values modification

Input variables	Class	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE
Without Quartiles	Ship type	35.8167%	42.4199%
Without Maximum	Ship type	35.4311%	43.0042%
Without Quartiles	Maneuver	68.3381%	64.8382%
Without Average	Maneuver	69.0816%	65.4583%

6) Case 6

The exclusion of speed variation reduces significantly most of the success rates of Case 0 in both classes. The time period and the direction variation are also relevant variables although their removal from the inputs make a reduction of the success rate for both classes (see Table 9).

Table 9. Case 6: Variable modification

Input variables	Class	TEST SEGMENT SUCCESS RATE	TEST TRAJECTORY SUCCESS RATE

Without speed variation	Ship type	35.1306%	42.4477%
Without Direction variation	Ship type	35.0376%	42.714%
Without Time period	Ship type	35.51%	42.3722%
Without Speed variation	Maneuver	68.7808%	65.657%
Without Time period	Maneuver	69.1283%	65.5696%

VI. Conclusions and perspectives

As a conclusion of the obtained results, the implementation of a system with the capacity of raw data processing and transformation to make it useful for classifiers and other data mining problems is achieved. With other cases analysis the conclusions are that the maneuver class is more influenceable by the cases variation and that the ship type has improvement from joining all the segment classes into a trajectory class whereas the maneuver has better results if we use only the segment classification (without joining all into a trajectory).

The fact that the ship class can improve from the joint of the segment classes whereas the maneuver has poorer results is because the segments are useful for sub-trajectory patterns that improve the maneuver prediction whereas a big trajectory could identify better the ship class that is doing it.

The maneuver classifier obtains clearly more accurate results than the ship type classifier, this is because the inputs are movement parameters which could be quite similar in different models of ships but not in their maneuvers.

It is relevant the consideration for future works of search an improvement of the input variables by finding new information that could distinguish between the models of ships. One option would be the ship dimensions but those may not appear in all the cases. Other future work would be a study to adjust the implemented limits like the 50 plots segmentation or the trajectory division with a separation of 30 seconds. Another interesting study of the trajectory division is to reconsider the splits between different maneuvers, making the system consider of trajectories with maneuver changes (with more than one maneuver within segments).

Also, it will be interesting the evaluation of information loss in the trajectories segmentation (from 0 to 49 plots), by an algorithm capable of discard those with the less interesting part of the segments or use them to a more specific analysis that could be added to the classifier results.

Finally, it is remarkable that some frequent errors are present in input data, such as instances with wrong labels (for example, anchored ships that have nonzero speeds). It could be interesting to include in the system some additional preprocess steps able of locate these errors (our system locates some of them, but not all) with the objective of improving results of classifiers (to train them with less wrong data) and also the possibility to use this process into its own system that could raise warnings to the maritime surveillance system operator.

VII. Acknowledgement

This work was funded by public research projects of Spanish Ministry of Economy and Competitivity (MINECO), reference TEC2017-88048-C2-2-R.

VIII. References

[1] M. A. McNicholas, B. Wilson, and S. D. Genovese, *Maritime Security: An Introduction*. 2016.

[2] J. Gómez-Romero, M. A. Serrano, J. García, J. M. Molina, and G. Rogova, "Context-based multi-level information fusion for harbor surveillance," *Inf. Fusion*, vol. 21, no. 1, pp. 173–186, 2015.

[3] J. Garcia, L. Snidaro, and J. Llinas, "Architectural Aspects for Context Exploitation in Information Fusion," in *Context Enhanced Information Fusion - Boosting Real World Performance with Domain Knowledge*, Springer, 2015.

[4] B. J. Rhodes, "Taxonomic knowledge structure discovery from imagery-based data using the neural associative incremental learning (NAIL) algorithm," *Inf. Fusion*, vol. 8, no. 3, pp. 295–315, 2007.

[5] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, "A Multi-task Deep Learning Architecture for Maritime Surveillance using AIS Data Streams," 2018.

[6] V. F. Arguedas, G. Pallotta, and M. Vespe, "Maritime Traffic Networks : From Historical Positioning Data to Unsupervised Maritime Traffic Monitoring," vol. 19, no. 3, pp. 722–732, 2018.

[7] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction," *Entropy*, vol. 15, no. 6, pp. 2218–2245, 2013.

[8] M. Riveiro, G. Pallotta, and M. Vespe, "Maritime anomaly detection: A review," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 5, 2018.

[9] Y. U. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, p. 29, 2015.

[10] H. Li, J. Liu, R. W. Liu, N. Xiong, K. Wu, and T. H. Kim, "A dimensionality reduction-based multi-step clustering method for robust vessel trajectory analysis," *Sensors (Switzerland)*, vol. 17, no. 8, 2017.

[11] A. Graser and P. Widhalm, "Modelling massive AIS streams with quad trees and Gaussian Mixtures," *21st Int. Conf. Geogr. Inf. Sci. (AGILE 2018)*, pp. 1–5, 2018.

[12] K. Yoo, J. Chun, and J. Shin, "Target Tracking and Classification for Missile Using Interacting Multiple Model (IMM)," *2018 Int. Conf. Radar, RADAR 2018*, no. Imm, pp. 1–6, 2018.

[13] Danish Maritime Authority, "AIS data." [Online]. Available: https://www.dma.dk/SikkerhedTilSoes/Sejladsinformation/AIS/Sider/default.aspx#.

[14] L. Breiman, *Classification and regression trees Regression trees*. Boca Raton, FL: Chapman & Hall, 1984.