Discovering Ship Navigation Patterns towards Environmental Impact Modeling

Alberto Gutiérrez, David Buchaca, Josep Lluís Berral
firstname.lastname@bsc.es
Barcelona Supercomputing Center, Data Centric Computing department

Index Terms—**Pattern Mining, Machine Learning, Neural Networks, Ships, Environmental modeling.**

In this work a data pipe-line to manage and extract patterns from time-series is described. The patterns found with a combination of Conditional Restricted Boltzmann Machine (CRBM) and k-Means algorithms are then validated using a visualization tool. The motivation of finding these patterns is to leverage future emission model.

I. Introduction

According to the European Community Shipowners Associations (ECSA) in 2015, the maritime traffic has become a key component for European economy [1], being sea transportation more fuel-efficient than other modes of transport (e.g. trucks and trains). Also, according to a recent report by the International Maritime Organization (IMO), it is expected that this form of transport will continue increasing in the future due to globalization and the increase of global-scale trade [2], but at the same time, it is considered an important contributor to primary atmospheric emissions in coastal areas [3] and subsequently to European coastal air quality degradation [4], especially in the North Sea and the Mediterranean basin. Maritime traffic is also responsible for about 2.5% of global greenhouse gas (GHG) emissions, which are predicted to increase between 50% and 250% by 2050 [2].

Automatic Identification System (AIS) is the Global Positioning System (GPS) based tracking system used for collision avoidance in maritime transport, as a supplement to marine radars. AIS provides information such as a unique identification for each transport (MMSI identifier), the position as latitude and longitude (GPS positioning), the course and speed (from the on-board gyrocompass). Such information is used by maritime authorities to track and monitor vessel movements and transmitted through standardized VHF transceivers, mainly to prevent collisions amongst ships.

Discovering which patterns ships perform according to the data provided will help to give explanation to: 1) air pollutant concentrations in coastal zones and cities; and 2) degradation of sea life, by detecting unusual or even criminal activities from fishing fleets working in special sea-life protection zones.

AIS data can be considered a *time-series*, as each input updates the vessel status in time. There are several approaches for mining patterns for time series, from stream mining methods for learning on time-changing data [5], to series-aware neural network methods like *Recurrent NNs* and Hidden Markov Models [6]. Here we are focusing on a simplistic pipeline consisting in CRBMs to deal with time dimensions [7], and a classical clustering method like *k-Means* [8]. The reasons for choosing CRBMs is because our analytics goal passes to determine patterns through dimensionality reduction attempting to simplify clustering and pattern mining processes, and CRBMs have the ability to encode multidimensional input data and its history into a dimension and time aware k-length code, easier for feeding simplistic clustering techniques.

A. Dataset

The currently used dataset has been provided by the Spanish Ports Authority (*Puertos del Estado*), from their vessel monitoring database collecting the AIS signals from all registered ships navigating national waters. The dataset used for current experiments is a slice of data concerning the coastal area of Barcelona, including a week of maritime traffic. It is composed by more than 1.5 million entries and indicating 19 features, including the vessel identification, the position in longitude and latitude degrees, speed over ground, facing position, and other vessel properties like vessel category.

The data is cleaned and processed in order to ensure that the data is of enough quality for analyzing it. Several features are then generated from the original for using them in the analysis, e.g. rotation attribute generated from the GPS traces.

B. Methodology

The methodology here presented implements a data pipeline consisting in the preparation of data, then passing the data through a CRBM for data encoding and reduction of dimensions, then clustering / classifying it through a classical method like *k-Means*. Following subsections explains each step of the chain, also depicted on Figure 1.

The CRBM is trained with sample series of data, structured as explained before. The CRBM is not aware of time by itself, but is our $history$ input data what provides such notion. This allows training it through data batches and without forcing data order, as the notion of order is already present on each new instance. Best practices in modeling and prediction require to split training data with validation and testing data, to prevent the auto-verification of the model, so for this reason we performed this training process with a subset of the available time-series.

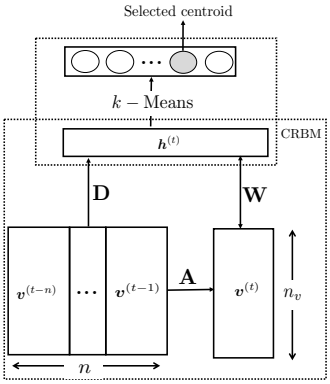

Fig. 1. Schema of the data pipeline, given a sequence at a particular time t (and its n predecessor values) the hidden activations of the CRBM are computed. Then the hidden activations are fed to a *k-Means* in order to cluster the sequence at time t.

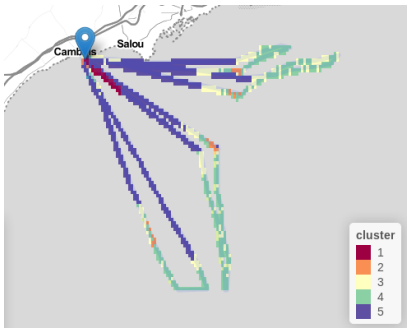

Fig. 2. Example of the visualizing tool for traces and categorization. In this case patterns are distinguished using colors over the traces of the ship.

Once CRBM and *k-Means* models are trained, new data can be fed to the pipe-line, encoding and classifying each input into a status. At this time, similitude between patterns are checked visually using a tool for ship trace visualization, created on behalf this and future analyses in our center, as can be seen in Figure 2.

The visualization tool also allows the superposition of traces for different ships, allowing us to detect geographical regions where clustering labels *cluster*, indicating where behaviors are caused by geographical causes.

II. Conclusion

Detecting and discovering patterns in maritime traffic is an important topic for modeling air quality in coastal urban zones and sea-life. Maritime emissions, combined with urban emissions (industry and road traffic), are responsible of pollution in such areas.

In this paper we presented a methodology for characterizing maritime traffic, understanding it as time series, and using an ensemble of CRBMs and clustering techniques like *k-Means* to reduce dimensionality of data while considering time, then clustering it into common patterns of traffic. Such methodology implies pre-processing data, knowing that AIS provides error-prone data. Such datasets can be cleaned using standard techniques, also aggregated features can be derived from the most reliable ones, i.e. GPS traces.

CRBMs have proven to be useful for reducing such dimensionality, as most time series contained more than 3000 observations, even after pre-processing and reducing the time scale from seconds to minutes. When tuning the CRBM hyper-parameters, we observed that it is not required to introduce a large history window ($<$ 20 minutes) or a high number of hidden units ($\sim$ 20) to achieve good results. Also *k-Means* appeared as a simple but effective approach to cluster the reduced outputs of CRBMs, comparable to real ship statuses.

By using the presented methodology, we observed identifiable patterns for real use cases, like vessel discrimination and operation modes. Such patterns can be used to complement or correct missing or erroneous data from AIS, trace ship behaviors and recognize their activity, and define geographical regions with common operation modes and behaviors. We also provided a tool for data and patterns visualization, available to the general public.

Acknowledgment

We would like to thanks to Marc Guevara and Albert Soret from Earth Sciences department for providing us with the use case, related information and their help on this project and Spanish Ports Authority (Puertos del Estado) for providing the data for this study. This project has received funding from the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (grant agreement No 639595). This work is partially supported by the Ministry of Economy of Spain under contracts TIN2012-34557, 2014SGR1051, and Severo Ochoa Center of Excellence SEV-2015-0493-16-5.

This paper has been submitted to a Data Mining conference and is under review process.

References

[1] E. C. S. A. (ECSA), "The economic value of the eu shipping industry. update," February 2015.

[2] T. S. et al., "Third imo greenhouse gas study," 2014.

[3] F. D. Natale and C. Carotenuto, "Particulate matter in marine diesel engines exhausts: Emissions and control strategies," *Transportation Research Part D: Transport and Environment*, vol. 40, pp. 166 – 191, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361920915001169

[4] M. Viana, P. Hammingh, A. Colette, X. Querol, B. Degraeuwe, I. de Vlieger, and J. van Aardenne, "Impact of maritime transport emissions on coastal air quality in europe," *Atmospheric Environment*, vol. 90, pp. 96 – 105, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1352231014002313

[5] A. Bifet and R. Gavaldá, "Learning from time-changing data with adaptive windowing," in *In SIAM International Conference on Data Mining*, 2007.

[6] Z. Ghahramani, "Hidden markov models." River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2002, ch. An Introduction to Hidden Markov Models and Bayesian Networks, pp. 9–42. [Online]. Available: http://dl.acm.org/citation.cfm?id=505741.505743

[7] V. Mnih, H. Larochelle, and G. E. Hinton, "Conditional restricted boltzmann machines for structured output prediction." in *UAI*, F. G. Cozman and A. Pfeffer, Eds. AUAI Press, 2011, pp. 514–522. [Online]. Available: http://dblp.uni-trier.de/db/conf/uai/uai2011.html#MnihLH11

[8] P. K. Agarwal and C. M. Procopiuc, "Exact and approximation algorithms for clustering," in *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '98. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1998, pp. 658–667. [Online]. Available: http://dl.acm.org/citation.cfm?id=314613.315040