Multivariate temporal data analysis for vessels behavior anomaly detection

Rui Maia
Instituto Superior Tcnico
Lisboa
rui.maia@tecnico.ulisboa.pt

ABSTRACT

Temporal or sequential data analysis is a broad research field with different relevant applications such as cyber-security, health care, biomedicine, automotive industry, financial analysis or supporting civil and military operations. Anomaly detection in data series is a specific field of data analysis considered as crucial in multiple applications areas, since abnormal data behavior typically represent critical situations that should be addressed: whether it is a network traffic pattern change that might indicate a cyber attack, a different heart beating frequency that can help anticipate and prevent an heart-attack, or different vessels behavior that might help detecting smuggling.

Multivariate time-series anomaly detection is a challenging research field that has been studied mainly supported on the adaptation of univariate time-series anomaly detection techniques. In this work, in the scope of MARISA - EU H2020 Project - we experiment and propose new methods that can natively include the multivariate dimensions of time-series without loss of information. We aim at modeling the multiple dimensions of vessels movement and the relations between series, including contextual information data series.

KEYWORDS

Anomaly detection; Outliers; Multivariate; Time-Series; Vessel

1 INTRODUCTION

The growing number of systems collecting large quantities of data is leveraging the need for efficient data processing approaches. Complex data streams from network logs, financial information systems, social networks, radars, automotive industry, health sensors, wearables or industry facilities, are just a few examples of massive data series being produced at every second. Time-series and temporal data are special types of data series, since they include one contextual or base dimension: time. Furthermore, while univariate time-series only vary in one dimension, multivariate time-series will vary in multiple dimensions at each moment of time.

Modeling the behavior of time-series and temporal data is a broad research field. One the prominent perspectives of data series modeling is the anomaly detection. It tries to point out unusual behaviors of an observation or group of observations, which typically may represent critical situations that should be addressed. Eg.:network traffic pattern change might indicate a cyber-attack, abnormal heart beating frequency can help anticipate and prevent heart-attacks, vessels behavior might help detecting smuggling.

This research work is developed in the scope of Maritime Integrated Surveillance and Awareness (Marisa) H2020 European Project [1]. It aims to integrate and process heterogeneous datasets of temporal data for the development of new behavior analysis and anomaly detection methods. The provided data sources include, for example, weather and sea conditions, vessels characteristics, geographic positions, undersea informations or radar and satellite information. These are analyzed to build vessels behavior models. We specifically aim at anticipating possible irregular activities, to support Search And Rescue operations and other civil and military operations.

2 MOTIVATION

Most anomaly detection research is dominated by univariate time-series anomaly detection methods, which only include one dimension varying over time. Univariate approaches are adapted by many authors to the multivariate case, commonly involving feature engineering, data cleaning and other necessary processes required by the use of univariate approaches. Typically, univariate approaches cannot model multivariate anomalies and hidden relations between series, rather they help to identify anomalies in one or more features of a data series.

Multivariate anomaly detection problem raise distinct and complex challenges [3] due to the hidden data structure and semantics between time-series as illustrated in Figure 1. Observation, Sequence, Context and Collective anomalies, from the multivariate time-series perspective, are still open and complex research challenges.

In the context of MARISA project we aim at building vessels behavior models that unveil unexpected collective behaviors. We aim at anticipating possible irregular activities taken by two or more vessels in organized movements and to support search and rescue operations or other civil and military tasks. These objectives perfectly maps the less frequently explored research and application domain of geographically referenced tracks anomaly detection.

We underline four scientific challenges in this work:

- Multivariate time-series anomaly detection using complex networks of different sensors

[1] http://www.marisaproject.eu

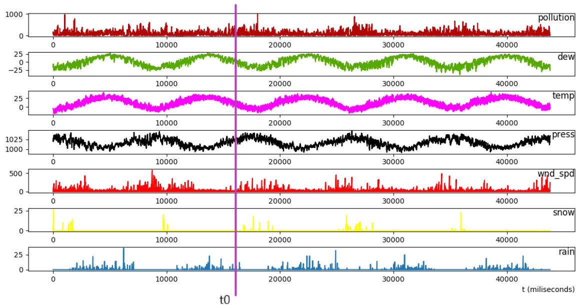

Figure 1: Multivariate Time-Series. Each observation in a specific time moment is associated with the multiple dimensions of a vessel movement. All dimensions might vary dependently or independently of each others and are influenced by other multivariate data series related to the domain (e.g.: sea and underwater conditions, atmospheric conditions, other vessels movements...)

- Using categorical and real valued parameters that might be manipulated by emitting entities
- Complex relations between different information areas (e.g.: weather, sea and undersea)
- Hidden semantic relations between different time-series

3 RELATED WORK

Relevant research works reviewing anomaly detection methods were published by Hodge et al. [5], Chandola et al. [3], Aggarwal [1] and Gupta et al. [4]. Some authors focused on specific application domains, but few have dedicated their attention to the multivariate time-series case, exploring and proposing new approaches.

In the scope of our work we investigate anomaly detection in multivariate time-series experimenting on a less explored context: vessel tracks in maritime context. Other authors analyzed more specific perspectives of anomaly detection which are related to our scope, for example: Ahmed et al. [2] focused on network anomalies and Zheng [8] explored trajectory data mining, including anomaly detection.

Very different approaches have been used to deal with anomaly detection in data series. We underline just three of recent published works from supported by prominent research perspectives: dimensionality reduction based on spectral analysis, probabilist approach, and neural networks.

Kane et al. [6] based their approach on dimensionality reduction and correlation analysis. They first represent a multivariate time-series as univariate, after what they apply Singular Values Decomposition (SVD) to uncover the most relevant latent features. The similarity between series is measured using simple methods such as Pearson's product moment coefficient. On other perspective, based on a probabilistic model, Zor et al. [9] proposed a framework for anomalous ferry (vessels) tracks detection. The method is

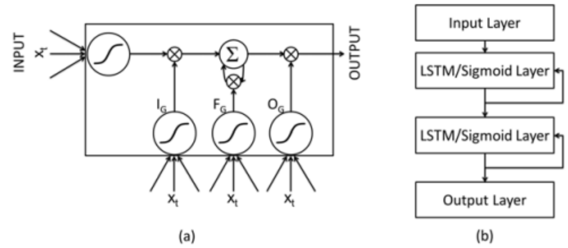

Figure 2: Long-Short Term Memory (LSTM) cells can arguably model the seasonal, or long term, behavior of a time-series.

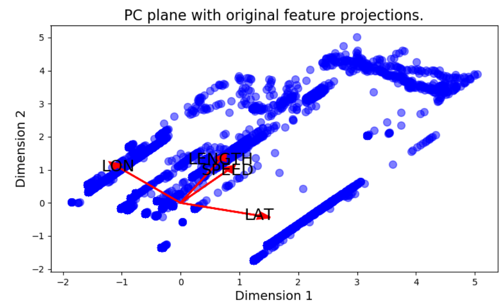

Figure 3: Multiple Vessel transmitted positions representation using two latent dimensions. The contributing dimension vectors are represented in the biplot.

based on speed, time and direction information. Experimenting on an Automatic Identification Systems dataset (AIS) from Thales, including one month of unlabeled data, the authors approach uses a Median Absolute Deviation algorithm to preprocess and clean the training data. Then, the base model is built on Markov Chains process regarding direction dimension, while and a Gaussian Process is applied for speed and time modeling. Finally, from neural networks perspective, Malhotra et al. [7] approach is based on a Long Short Term Memory (LSTM) network (as illustrated on Figure 2). They state that the use o LSTM can incorporate seasonal or long term tendencies of data series that are difficult to capture using other techniques.

4 EXPERIMENTS

The ongoing experimentation includes testing different approaches and algorithms adaptation. We are currently investigating the relations between multiple dimensions of each vessel track (see Figure 3). We aim to integrate this information in an LSTM adapted model, that can also integrate domain knowledge. This knowledge should help to model the multiple anomaly types: observation, context, sequence and collective anomalies. We are also testing a Support Vector Machine (SVM) based approach in order to define the characteristics of data series relations.

For the experimentation we are using MARISA datasets not publicly available, but also two open datasets, namely from: the Australian Maritime Safety Authority (AMSA[2]), the Office for Coastal Management of the National Oceanic and Atmospheric Administration (NOAA[3]), the National Centers for Environmental Information (NOAA[4]) and the Australian Ocean Data Network (AODN[5]). The former two data sources provide open access to Automatic Identification System (AIS) data for vessel positioning analysis, and the later are open access climate datasets, including atmosphere and ocean related information.

National military experts participating in MARISA project are involved in the validation of anomaly detection results, as well as contributing to the definition of abnormal events semantics. These validations and definitions are supporting unsupervised (SVM) and semi-supervised (LSTM) experiments.

ACKNOWLEDGMENTS

MARISA project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 740698.

REFERENCES

[1] Charu C Aggarwal. 2013. *Outlier Analysis*. Springer Science & Business Media.

[2] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. 2016. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications* 60 (2016), 19–31.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 15.

[4] Manish Gupta, Jing Gao, Charu C Aggarwal, and Jiawei Han. 2014. Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering* 26, 9 (2014), 2250–2267.

[5] Victoria Hodge and Jim Austin. 2004. A survey of outlier detection methodologies. *Artificial intelligence review* 22, 2 (2004), 85–126.

[6] Aminata Kane and Nematollaah Shiri. 2017. Multivariate Time Series Representation and Similarity Search Using PCA. In *Industrial Conference on Data Mining*. Springer, 122–136.

[7] Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. 2015. Long short term memory networks for anomaly detection in time series. In *Proceedings*. Presses universitaires de Louvain, 89.

[8] Yu Zheng. 2015. Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)* 6, 3 (2015), 29.

[9] Cemre Zor and Josef Kittler. 2017. Maritime anomaly detection in ferry tracks. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2647–2651.

[2] https://www.operations.amsa.gov.au/Spatial/DataServices/DigitalData

[3] https://coast.noaa.gov/htdata/CMSP/AISDataHandler/2014/index.html

[4] https://www.ncdc.noaa.gov/data-access/marineocean-data

[5] https://portal.aodn.org.au/search