Towards a hybrid human-machine discovery of complex movement patterns*

Natalia Andrienko[1,2][0000−0003−3313−1560], Gennady Andrienko[1,2][0000−0002−8574−6295], Alexander Artikis[3,4][0000−0001−6899−4599], Periklis Mantenoglou[5,3][0009−0002−3275−1522], and Salvatore Rinzivillo[6][0000−0003−4404−4147]

[1] Fraunhofer Institute IAIS, 53757 Sankt Augustin, Germany
[2] City, University of London, UK, http://www.geoanalytics.net
[3] NCSR "Demokritos", Greece
[4] University of Piraeus, Greece
[5] National and Kapodistrian University of Athens, Greece
[6] KDDLab, CNR-ISTI, Italy

Abstract. Results of automated detection of complex patterns in temporal data, such as trajectories of moving objects, may be not good enough due to the use of strict pattern specifications derived from imprecise domain concepts. To address this challenge, we propose a novel visual analytics approach that combines expert knowledge and automated pattern detection results to construct features that effectively distinguish patterns of interest from other types of behaviour. These features are then used to create interactive visualisations enabling a human analyst to generate labelled examples for building a feature-based pattern classifier. We evaluate our approach through a case study focused on detecting trawling activities in fishing vessel trajectories, demonstrating significant improvements in pattern recognition by leveraging domain knowledge and incorporating human reasoning and feedback. Our contribution is a novel framework that integrates human expertise and analytical reasoning with ML or AI techniques, advancing the field of data analytics.

Keywords: Movement data analysis · Trajectory data · Pattern detection · Feature-based pattern classification · Interactive visual analytics · Human-computer analysis workflow.

1 Introduction

One of common tasks in analysing time-referenced data, such as multivariate time series and trajectories of moving objects, is to find time intervals where the manner, or pattern, of data variation is indicative of particular kinds of

* This work was supported by EU in project *CrexData* (grant agreement no. 101092749) and by Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the *Lamarr Institute for Machine Learning and Artificial Intelligence* (Lamarr22B).

dynamic behaviour. Automatic detection of such patterns by means of computer algorithms requires precise specification of what values may occur and how the data are expected to vary. In many application domains, however, patterns of interest have no exact definitions. What can be elicited from domain experts is often far from being distinct and precise, for example, "A flock is a large enough group of objects moving close to each other for a certain time". Translation of such description to a form suitable for automated search involves introducing parameters and thresholds; see, for example, the formal definition of the flock pattern [8]: "Let $m, k \in N$, and let $r > 0$ be a constant. Consider a set of trajectories, where each trajectory consists of T line segments. A flock in a time interval $I = [t_i, t_j]$, where $j - i + 1 \geq k$, consists of at least m entities such that for every point in time within I there is a disk of radius r that contains all the m entities".

Formalisation of vague definitions elicited from domain experts entails two problems. First, the choice of appropriate parameter settings may not be obvious, while different choices may lead to very diverse results. Second, after the parameters are set, the definitions become rigid and intolerant to even minor data noise and small deviations from the thresholds. Imagine, for example, that just for a single time moment one of the m entities moving in a flock steps out from the disk of radius r. This breaks the time interval I in which the conditions of the formal definition of a flock hold. If the lengths of the sub-intervals are less than k, the flock will not be detected.

We encountered the problem of definition rigidity in exploring the work of a knowledge-based system designed to detect complex activity patterns in vessel movement [17]. The system applies Event Calculus [5] to a set of formal definitions, many of which involve constant thresholds such as speed bounds, minimal change in movement direction, frequency of changes, and minimal duration of an activity. Upon observing that the system fails to recognise a significant number of visually identifiable pattern instances, we employed interactive visualisation to investigate the data used for the inference. We found that the minimal activity duration was often not formally reached due to occasional breaks in the fulfillment of the rule conditions, which, in turn, happened because of data noise and small variations of attribute values around the thresholds.

Hence, formalisation of human-defined concepts may not be a good approach in tasks requiring the tolerance and flexibility of human reasoning. Probabilistic methods of pattern recognition (e.g., [16]) can be less sensitive to data noise, but they still assume that pattern specifications obtained from experts are complete and precise, which is not always the case.

Opposite to specification-driven approaches, machine learning methods strive to acquire the ability of pattern recognition by generalizing from labelled data examples. Due to the generalization, the resulting classification models can be sufficiently flexible regarding data variability. However, machine learning methods require large numbers of representative training examples, which may be very problematic. While domain experts can usually easily identify a pattern (or pattern absence) given an appropriately represented piece of data, their time is

too costly to be spent for considering and labelling a large number of individually shown examples.

The inherent problems of the knowledge- and data-driven approaches call for hybrid solutions that would be able to effectively leverage expert knowledge while accommodating the flexibility of human reasoning [2, 3], abstractive perception and capability to give meaning to visual patterns [4].

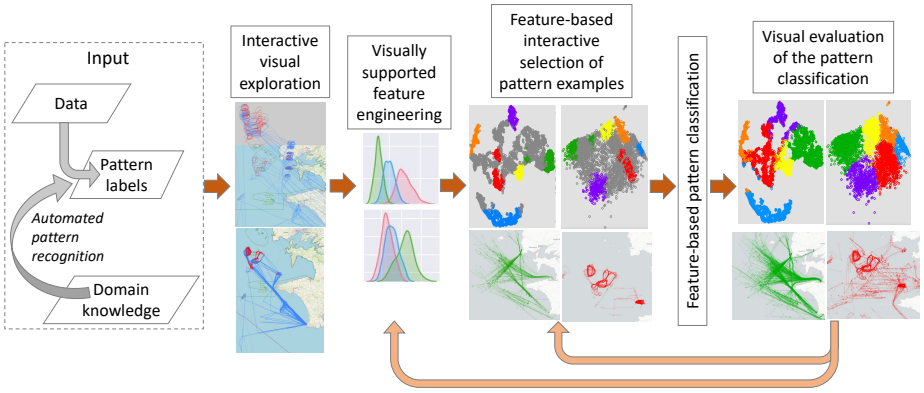

Fig. 1. A schematic representation of the workflow of the hybrid human-machine pattern classification.

To address this challenge, we propose a novel visual analytics approach (see Fig. 1), in which domain knowledge is used for constructing features capable of effectively distinguishing patterns of interest from other types of behaviour. It is essential to note that these features need to characterise the *behaviour* of relevant variables *on time intervals*, whereas raw data consist of elementary values referring to individual time steps. Hence, feature construction requires knowledge of (a) what aspects of the behaviour are important, e.g., the range of the values or the development trend, and (b) what kinds of computationally derivable aggregate characteristics can represent these aspects. The derived variables are utilised to generate interactive visualisations enabling experts to select, view, and label groups of similar data items. These labeled examples of different pattern types are then supplied to a machine learning method for developing an automated classifier. By actively involving a human analyst in the process, our approach achieves flexibility in utilising domain knowledge and accommodating data variations. The interactive visual interface enables simultaneous consideration and labelling of multiple data items, which saves the precious time of the human while allowing creation of a sufficiently large set of data examples for model training. We evaluated the effectiveness of our approach through a case study focused on detecting trawling activities in fishing vessel trajectories. However, our approach is sufficiently general to be applicable to other domains facing similar challenges.

2 Approach introduced through a case study

We shall briefly demonstrate the process of human-guided development of a pattern recognition model using the case study on detection of trawling activity patterns in trajectories of 71 fishing vessels that operated in the waters northwest of France, from October 1, 2015, to March 31, 2016. This is a subset of an openly accessible dataset [18, 1, 6]). The process is illustrated in Fig. 2.

In our case study, the successfully recognised pattern instances from the output of the knowledge-based computer system can be used as supporting material to create an extended set of representative labelled examples that encapsulate the patterns of interest. However, the proposed approach does not depend on the availability of such supporting material. In another application (namely, recognition of football teams' playing styles), we had no examples of successful recognition. The key requirement is the availability of domain knowledge that includes descriptions of patterns of interest as well as criteria for distinguishing these patterns from the remaining data. In the maritime activities case study, the knowledge has been encapsulated in formal rules created earlier in communication with domain experts, while for the football application we gained the necessary information from special literature. The criteria differentiating the patterns are translated into computationally derivable features of data segments, so that the presence of a pattern can be indicated by a particular combination of feature values. If an initial set of successfully recognised pattern examples is available, it can be used to test how consistently and distinctly they are characterised by the feature values and adjust the feature set when necessary. The step of feature testing and adjustment may be repeated after obtaining some version of a classifier if it performs insufficiently well (see Fig. 1).

Time-variant data, such as trajectories or time series of attribute values, need to be partitioned into segments by dividing the time into intervals of suitable duration according to the expected duration of the behaviours or activities to be identified. This is done using a sliding time window, so that the data segments partially overlap ensuring that patterns of interest are not overlooked due to being fragmented into disjoint parts. The characteristic features [14] are computed for the resulting data segments.

In our case study, we divide the trajectories of the fishing vessels into segments, called episodes, of length 3 hours using a sliding window shifted by 1 hour. From time series of movement attributes, we construct features to distinguish trawling from other movements based on low speed and repeated changes of movement direction. For the characterisation of speed magnitudes, we compute the minimum, maximum, and quartiles of the speed values within each episode. To capture changes in movement direction, we calculate the following features: the amplitude of direction deviation from the start-end vector, the angle of the trend line in the progression of the distances from the start, the amplitude of value deviations from the trend line, and the Pearson correlation coefficient between the distance values and corresponding time moments. In total, we derive 9 numeric features. The available results of automated pattern detection are used to assess feature effectiveness. For this purpose, we utilise frequency his-

tograms to compare the distributions of feature values for episodes containing at least 50% trawling-classified points with the overall distributions. We observe that the feature distributions indeed exhibit distinguishing patterns for trawling activities.

As the approach strongly relies on cognitive capabilities of a human analyst, interactive visualisations play a crucial role. They enable the analyst to assess the distinctiveness of features, select representative examples of pattern classes, and evaluate results of example-based pattern recognition. Depending on the evaluation, the analyst may need to revisit previous steps, such as providing additional examples or modifying the set of features.

We utilise dimensionality reduction to create a 2D spatial embedding. In Fig. 2D, it can be seen that the purple-coloured points representing episodes with automatically detected trawling are concentrated within a relatively compact area in the centre of the projection plot. This observation additionally suggests that the extracted features effectively capture the essence of trawling activities and can provide a distinct separation from other types of movements.

In the next step, an expert can use an interactive projection plot to select groups of points that have close positions in the projection space and see a visual representation of the corresponding episodes (Fig. 2E and B). This allows the expert to specify examples of different types of movement patterns. In our case, there are two patterns (looping and sweeping movements) that may indicate trawling activities (Fig. 2H). As all embedding methods introduce distortions, we use simultaneously two embeddings with different properties: MDS [13] that attempts to faithfully represent all distances, and t-SNE [15] that aims at preserving local neighborhoods at the cost of long-range distances.

To classify remaining episodes, we apply, as a proof of concept, the kNN algorithm [10, 9]. The labelled data can also be used for creating different kinds of ML models or for refining the set of rules [12]. As the results of the first application of kNN are not satisfactory, we extend the set of examples (Fig. 2F). After evaluating the new results, we decide to improve the performance further by extending the set of features. The maps in Fig. 2H-J demonstrate iterative refinement of the classification results. In this process, we managed to decrease the number of false negatives from 202 to 166 and the number of false positives from 422 to 121. The number of correctly recognised trawling patterns increased from 1640 to 1676, which is a large improvement compared to the 530 automatically detected trawling episodes.

3 Discussion and conclusion

The case study has demonstrated that involvement of human intelligence can significantly enhance the detection of complex behavioural patterns reflected in data. Our approach is not specifically designed for movement data. In fact, we applied it to time series of values of multiple numeric attributes [7] rather than to spatial positions or geometric shapes. The key step is data abstraction [11] by defining relevant features characterising the behaviour of the attributes on

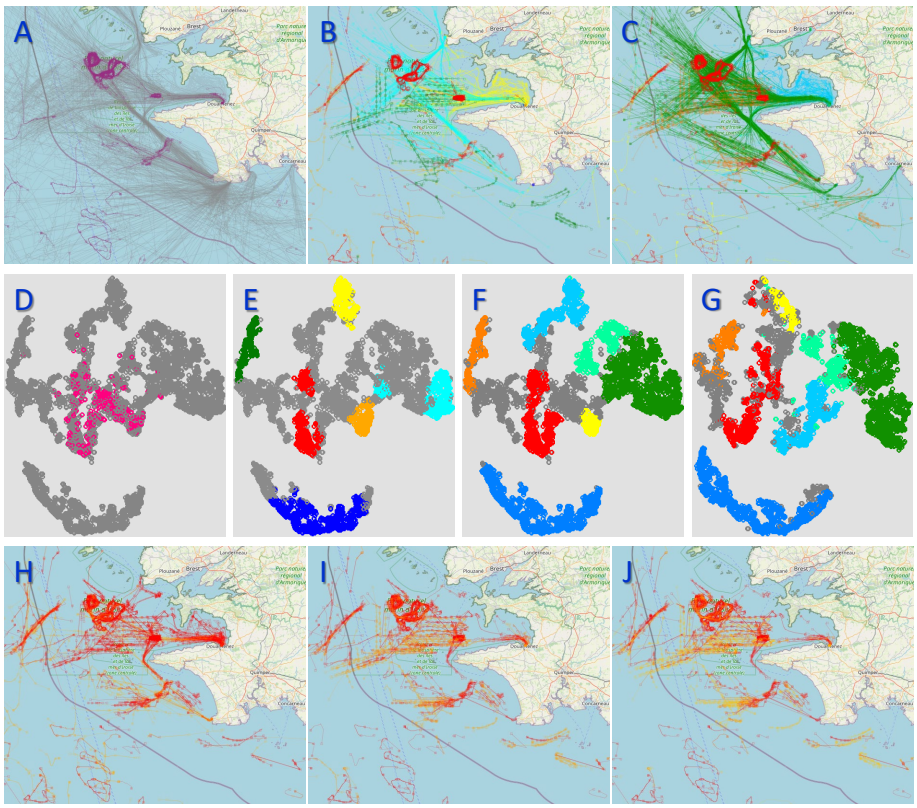

Fig. 2. Visualisations used throughout the workflow. A-C: Maps show episodes with automatically recognised trawling patterns (A, in purple), initial subset of pattern class examples (B), and extended subset of examples (C). D-F: 2D space embedding of the set of episodes based on initially chosen features is used for assessing the feature suitability (D), creation of an initial set of examples (E), and extending the set of examples (F). G: 2D space embedding based on an extended combination of features. H-J: Episodes recognised as containing trawling patterns based on similarities to the initial (H) and extended (I, J) sets of examples based on the initially chosen features (H, I) and the extended combination of features (J).

time intervals [14]. Effective feature engineering requires not only domain knowledge of pattern specifics (which may be incorporated in formal rules, as in our case study) but also good understanding of possible data transformations and capability to utilise different statistical metrics to characterise data properties and behaviours. In addition, it requires cognitive capabilities of a human to see and understand what data can tell and to derive meaningful concepts, such as pattern types, which can then be communicated to the machine and used in computational data analysis.

Our approach contributes to the field of data analytics by demonstrating a possible way to integrate human expertise and analytical reasoning with machine learning and artificial intelligence techniques. The key role in the approach belongs to interactive visualisations enabling human analysts to apply their cognitive capabilities.

References

1. Andrienko, N., Andrienko, G.: Visual Analytics of Vessel Movement, pp. 149–170. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-61852-0_5
2. Andrienko, N., Andrienko, G., Adilova, L., Wrobel, S.: Visual analytics for human-centered machine learning. IEEE Computer Graphics and Applications **42**(1), 123–133 (2022). https://doi.org/10.1109/MCG.2021.3130314
3. Andrienko, N., Andrienko, G., Fuchs, G., Slingsby, A., Turkay, C., Wrobel, S.: Visual analytics for data scientists. Springer (2020). https://doi.org/10.1007/978-3-030-56146-8
4. Andrienko, N., Andrienko, G., Miksch, S., Schumann, H., Wrobel, S.: A theoretical model for pattern discovery in visual analytics. Visual Informatics **5**(1), 23–42 (2021). https://doi.org/10.1016/j.visinf.2020.12.002
5. Artikis, A., Sergot, M., Paliouras, G.: An event calculus for event recognition. IEEE Transactions on Knowledge and Data Engineering **27**(4), 895–908 (2015). https://doi.org/10.1109/TKDE.2014.2356476
6. Artikis, A., Zissis, D. (eds.): Guide to Maritime Informatics. Springer (2021). https://doi.org/10.1007/978-3-030-61852-0
7. Beeram, S., Kuchibhotla, S.: Time Series Analysis on Univariate and Multivariate Variables: A Comprehensive Survey, pp. 119–126 (10 2020). https://doi.org/10.1007/978-981-15-5397-4_13
8. Benkert, M., Gudmundsson, J., Hübner, F., Wolle, T.: Reporting flock patterns. Computational Geometry **41**(3), 111–125 (2008). https://doi.org/https://doi.org/10.1016/j.comgeo.2007.10.003
9. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE transactions on information theory **13**(1), 21–27 (1967)
10. Fix, E.: Discriminatory analysis: nonparametric discrimination, consistency properties. USAF school of Aviation Medicine (1951)
11. Höppner, F.: Time series abstraction methods - a survey. In: Informatik Bewegt: Informatik 2002 - 32. Jahrestagung Der Gesellschaft Für Informatik e.v. (GI). p. 777–786. GI (2002)
12. Katzouris, N., Paliouras, G., Artikis, A.: Online learning probabilistic event calculus theories in answer set programming. Theory and Practice of Logic Programming **23**(2), 362–386 (2023). https://doi.org/10.1017/S1471068421000107

13. Kruskal, J.B.: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika **29**(1), 1–27 (Mar 1964). https://doi.org/10.1007/BF02289565
14. Lubba, C.H., Sethi, S.S., Knaute, P., Schultz, S.R., Fulcher, B.D., Jones, N.S.: Catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis. Data Min. Knowl. Discov. **33**(6), 1821–1852 (nov 2019). https://doi.org/10.1007/s10618-019-00647-x
15. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
16. Mantenoglou, P., Artikis, A., Paliouras, G.: Online probabilistic interval-based event calculus. In: Giacomo, G.D., Catalá, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., Lang, J. (eds.) ECAI 2020 - 24th European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, vol. 325, pp. 2624–2631. IOS Press (2020). https://doi.org/10.3233/FAIA200399
17. Pitsikalis, M., Artikis, A.: Composite Maritime Event Recognition, pp. 233–260. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-61852-0_8
18. Ray, C., Dréo, R., Camossi, E., Jousselme, A.L.: Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance (Feb 2018), https://doi.org/10.5281/zenodo.1167595