Integrating human knowledge within a hybrid clustering-classification scheme for detecting patterns within large movement data sets

René Enguehard
Memorial University of Newfoundland
St. John's, NL, Canada
rene@computer.org

Benjamin Fowler
Memorial University of Newfoundland
St. John's, NL, Canada
b.fowler@mun.ca

Orland Hoeber
Memorial University of Newfoundland
St. John's, NL, Canada
hoeber@mun.ca

Rodolphe Devillers
Memorial University of Newfoundland
St. John's, NL, Canada
rdeville@mun.ca

Wolfgang Banzhaf
Memorial University of Newfoundland
St. John's, NL, Canada
banzhaf@mun.ca

Abstract

The visual analysis of large movement data sets can be a challenging task. This study proposes an approach for identifying interesting movement patterns that combines human knowledge and decision making with a hybrid clustering-classification method. Rather than performing an unsupervised clustering of the entire data set, a stratified random sample of the full data set is used to identify initial clusters that are verified and labelled by the analyst, and then used as input patterns for classifying the remainder of the data set using an iterative genetic program. Classifications suggested after each iteration are presented to the analyst for refinement based on their knowledge and experience. A geovisual analytics environment is provided to both show the outcomes of the clustering and classification, and to obtain the analyst's input during the hybrid clustering-classification process. Our approach allows data to be classified without *a priori* specification of classification patterns. Instead, the process takes advantage of human decision making within the automatic analysis of the data. The approach was tested with fishing vessel movement data in Eastern Canada.

Keywords: movement data, human decision-making, clustering, classification, genetic programming.

1 Introduction

Large movement data sets often contain millions of records, which make visual analysis a challenging task. Many methods have been used to help deal with this problem, including automated statistical analysis, filtering mechanisms, pattern recognition, unsupervised clustering, and supervised classification. Several of these methods are commonly used to provide users with groups of data that share some similarities, or on the contrary, identify data that do not fit some groups (i.e. anomalies).

Clustering is a data mining technique that attempts to group similar data in an unsupervised manner. Numerous clustering algorithms have been developed to deal with large data sets. Some, such as k-means, rely on the distance between data points and cluster centroids [6]. Approaches such as DBSCAN or OPTICS rely on the density distribution of the data [10,11], whereas expectation maximization attempts to group similar data based on their statistical distribution [8]. While an appealing method in terms of simplicity, clustering large data sets can often yield thousands of clusters, or require *a priori* knowledge to produce a smaller number of cohesive clusters.

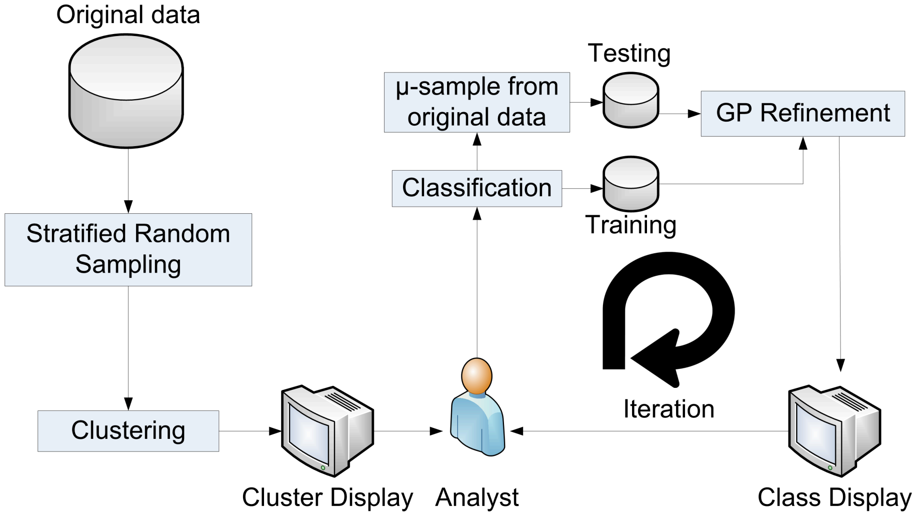
In contrast, classification operates in a supervised manner and has a lower computational cost, but requires *a priori* knowledge of the classes that are to be discovered in the data [9]. The classifiers are constructed based on an analysis of a robust set of training data. However, such data are often not available, particularly when the goal is to explore large data sets to find interesting, unusual, or anomalous patterns, the details of which may not be known in advance.

Combining clustering and classification methods in a hybrid approach has been explored in a number of different domains [1,7,12]. Our goal in this research is to integrate the analyst's knowledge within this process, not only for the interactive identification of patterns to be classified, but also through the iterative refinement of the classifiers.

2 Proposed Method

We propose a hybrid clustering-classification method that addresses some of the shortcomings of either method used in isolation, through the direct involvement of the analyst and an iterative class refinement within a geovisual analytics environment (Figure 1). While this approach may be useful for the analysis of different types of movement data, it was developed to support the visual analysis of fishing vessel movement data, with the goal of semi-automatically identifying interesting features within the data. The intent is to provide the analyst with an effective way of dealing with the large amounts of fishing vessel movement data currently collected throughout the world. Potential uses include the analysis of shipping lanes or support for fisheries enforcement activities.

Figure 1: A hybrid clustering-classification process that integrates human knowledge within the classification process.

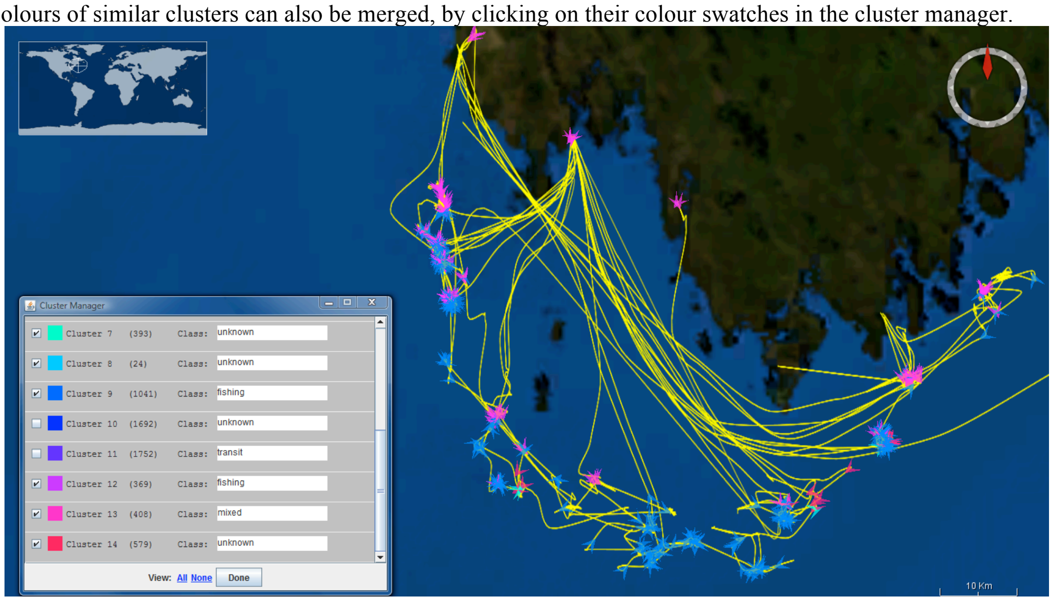


The data used in our study are based on the Vessel Monitoring System (VMS) and provide fishing vessel positions collected hourly by GPS for the scallop fishery in the Bay of Fundy and on the Scotian Shelf, Eastern Canada. The data set includes all of the fishing vessels that had a license to fish for scallops in this region over the 2008-2009 fishing season (approximately 2,000,000 data points), and consisted of the vessel identifier, timestamp, latitude, and longitude. A number of calculated or derived parameters were added to the data set, including vessel heading, velocity, bathymetry (ocean depth), slope of the ocean floor at the recorded position, distance to the coast, and distance to the previous data point (persistence of motion). These parameters were selected due to their relation to specific fishing activities; case studies in other domains may require a different set of supplemental attributes.

Our approach is to provide an analyst with a "first pass" clustering of the data, based on the parameters of a stratified random subset of the vessels. The parameters used in the clustering are the derived parameters, which incorporate the spatial characteristics of movement (heading, persistence), as well as temporal characteristics (velocity), and environmental characteristics (bathymetry, slope). Focusing on these six parameters, instead of the full range of parameters, allows the complex tasks of determining the number and composition of the clusters to occur in a more efficient manner due to the reduction in the size of the data set. The stratified nature of the sample ensures that entire vessel paths are considered, preserving individual vessel patterns such as going to and from fishing grounds, and ensuring statistical similarity to the original data set.

Based on these results, the analyst can then identify meaningful clusters for their intended task and assign these to user-defined classes of movement patterns (e.g., dredging for scallops, drifting, going to and from port). This process occurs within a geovisual analytics environment developed in prior research [5], which is based on NASA's WorldWind system. The initial clusters are shown to the analyst within a virtual globe. The analyst can then select, label, and colour these based on the types of activities occurring and the patterns they wish to extract from the data (see Figure 2).

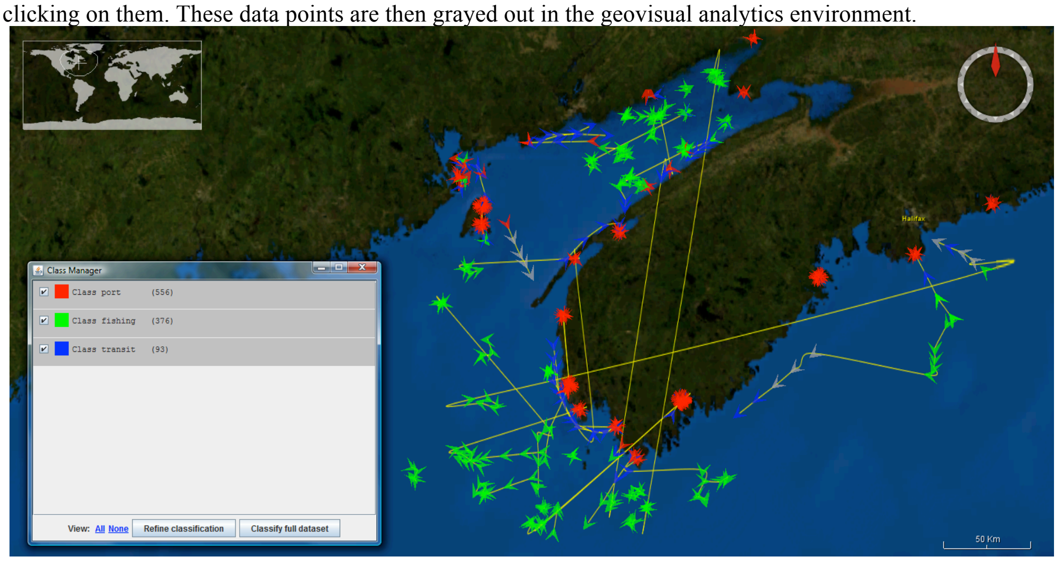
Figure 2: The geovisual analytics environment allows analysts to assign classes to the clustered subset of the data. The colours of similar clusters can also be merged, by clicking on their colour swatches in the cluster manager.



Cluster Manager

☑	🟩	Cluster 7	(393)	Class:	unknown
☑	🟦	Cluster 8	(24)	Class:	unknown
☑	🟦	Cluster 9	(1041)	Class:	fishing
□	🟦	Cluster 10	(1692)	Class:	unknown
□	🟪	Cluster 11	(1752)	Class:	transit
☑	🟪	Cluster 12	(369)	Class:	fishing
☑	🟪	Cluster 13	(408)	Class:	mixed
☑	🟥	Cluster 14	(579)	Class:	unknown

Figure 3: Analysts can inspect the GP classification outputs and flag data points as being incorrectly classified, by clicking on them. These data points are then grayed out in the geovisual analytics environment.



Class Manager

☑	🟥	Class port	(556)
☑	🟩	Class fishing	(376)
☑	🟦	Class transit	(93)

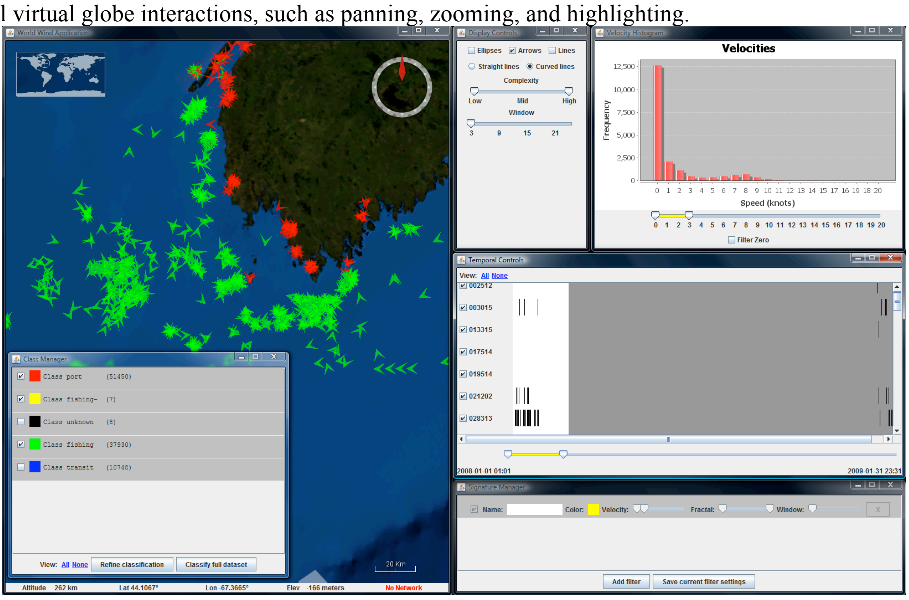
These classified data are then used as training data for classifying the remainder of the data using a heuristic-based Genetic Programming (GP) System [2,4] as a supervised classifier. However, rather than classifying the remainder of the data all at once, it is beneficial to provide the analyst with some control over the quality of the classification. To do this, another stratified random subset of the vessels is chosen, which is independent of the data used in the clustering phase. The GP system starts by classifying this subset, the results of which are shown to the analyst within the geovisual analytics environment. These data are coloured to match the previously assigned classes, allowing the analyst to easily relate them to the information previously provided. Any misclassified data can simply be selected and marked as incorrect (see Figure 3).

This knowledge is then integrated into the system by adding correctly classified data points to the training data. Misclassified data points are retained and added to the next subset selected from the data, and the process is repeated. Once the analyst finds that this interactive refinement of the classification is no longer leading to an improvement in the quality of the classifiers, this process can be terminated and the remainder of the data will then be classified all at once.

Upon completion, the final classification of the data is shown within the geovisual analytics environment, highlighting the classifications using colouring and labelling (see Figure 4). By visualizing the classified data spatially and across the entire data set, the analyst can readily identify the spatial distribution of the movement patterns and inspect other aspects of the data (e.g., velocity, path complexity [2]). The system allows the analyst to focus on specific classes or characteristics of movement, using a visual filtering interface integrated into a virtual globe.

While others have proposed similar approaches to combining clustering and classification techniques that allow the analyst to provide input into which clusters should be used in the classification of the data [1], the novelty of this approach is that input from the analyst is also integrated into the classification process. As a result, not only can the analyst provide information regarding interesting clusters initially extracted from a subset of the data, they can further guide the development of the classifiers in an iterative and interactive manner. Such an approach provides a great deal of control to the analyst for specifying how the data is to be grouped, taking advantage of the tacit knowledge they possess and integrating this within the semi-automated analysis process.

Figure 4: Once the entire data set is classified, it can be filtered and explored within the geovisual analytics environment. This environment allows data exploration using visual analytics tools, such as velocity, complexity, and temporal filtering, as well as traditional virtual globe interactions, such as panning, zooming, and highlighting.



Class Manager

☑	🟥	Class port	(51450)
☑	🟨	Class fishing-	(7)
☐	⬛	Class unknown	(8)
☑	🟩	Class fishing	(37930)
☐	🟦	Class transit	(10748)

Altitude 262 km Lat 44.1067° Lon -67.3665° Elev -166 meters No Network

Display Controls

☐ Ellipses ☑ Arrows ☐ Lines

○ Straight lines ● Curved lines

Temporal Controls

☑ 002512
☑ 003015
☑ 013315
☑ 017514
☑ 019514
☑ 021202
☑ 028313

2008-01-01 01:01 2009-01-31 23:31

Signature Manager

☑ Name: Color: 🟨 Velocity: Fractal: Window:

3 Conclusion & Future Work

Visual analysis of large movement data sets, while required in some contexts, can be challenging for the analyst. The approach proposed in this paper combines the simplicity of clustering a small data subset, the decision-making power of the human mind, and the classification power of an iterative GP system. As a result, it is possible to classify large data sets with no *a priori* knowledge of the data; instead, knowledge is added by the analyst through the identification and labeling of interesting clusters, and later through the interactive identification of misclassified data points.

The decrease in computational cost provided by clustering over a sample of the data allows for the analysis of very large data sets. The iterative classification, driven by GP, and combined with a geovisual analytics environment, allows the analyst to take an active role in the process. Visualizing the resulting classification within an interactive geovisual analytics environment further assists the analyst in the task of understanding and exploring the movement data sets.

In future work, we plan to perform a direct comparison of the proposed approach against other techniques, such as clustering the entire data set, non-iterative clustering-classification techniques, and pure GP approaches. Delegating the GP classification to a Graphics Processing Unit (GPU) would also allow the GP to evolve more generations per iteration, potentially decreasing the amount of iterations required to arrive at an acceptable classification [3].

Acknowledgements

The authors wish to thank Fisheries and Oceans Canada for providing the data used in this research. This work was supported by a Strategic Projects Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) held by the last three authors.

References

[1] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. Interactive visual clustering of large collections of trajectories. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, pages 3-10, 2009.

[2] G. Bakırlı, D. Birant, and A. Kut. An incremental genetic algorithm for classification and sensitivity analysis of its parameters. *Expert Systems with Applications,* 38(3):2609-2620, 2011.

[3] W. Banzhaf, S. Harding, W.B. Langdon, and G. Wilson. Accelerating Genetic Programming on Graphics Processing Units. In *Genetic Programming in Theory and Practice VI,* R. Riolo, T. Soule, and B. Worzel (eds.), Springer, New York, pages 229 - 248, 2009.

[4] W. Banzhaf, P. Nordin, R. Keller, and F. Francone. *Genetic Programming – An Introduction,* Morgan Kaufmann, San Francisco, 1998.

[5] R.A. Enguehard, R. Devillers, and O. Hoeber. Geovisualization of fishing vessel movement patterns using hybrid fractal/velocity signatures. In *Proceedings of GeoViz Hamburg,* pages 1-2, 2011.

[6] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 28(9):1450-1464, 2006.

[7] A. Kyriakopoulou and T. Kalamboukis. Using clustering to enhance text classification. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 805-806, 2007.

[8] R. Laxhammar, G. Falkman, and E. Sviestins. Anomaly detection in sea traffic - a comparison of the Gaussian Mixture Model and the Kernel Density Estimator. In *Proceedings of the International Conference on Information Fusion*, pages 756-763, 2009.

[9] O.D.A. Prima, A. Echigo, R. Yokoyama, and T. Yoshida. Supervised landform classification of Northeast Honshu from DEM-derived thematic maps. *Geomorphology,* 78(3-4):373-386, 2006.

[10] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization,* 7(3-4):225-239, 2008.

[11] J.A.M.R. Rocha, V.C. Times, G. Oliveira, L.O. Alvares, and V. Bogorny. DB-SMoT: A direction-based spatio-temporal clustering method. In *Proceedings of the IEEE International Conference Intelligent Systems*, pages 114-119, 2010.

[12] H-J. Zeng, X-H. Wang, Z. Chen, H. Lu, and W-Y. Ma. CBC: Clustering based text classification requiring minimal labelled data. In *Proceedings of the IEEE International Conference on Data Mining*, pages 443-450, 2003.