Real-time incremental machine learning for anomaly detection and surveillance in water quality

X. Rigueira[1], D. N. Olivieri[2], M. Araújo[1], M. Pazo[1], E. Alonso[1]

[1]CINTECX, Department of Natural Resources and Environmental Engineering, University of Vigo, Vigo, Galicia, 36310, Spain
[2]Department of Computer Science, University of Vigo, Vigo, Galicia, 32004, Spain

Keywords: incremental machine learning, water quality, anomaly detection, environmental modelling.

Presenting author email: xurxo.rigueira@uvigo.es

Water covers over 70% of the earth's surface and is crucial for the survival of most living organisms. Rivers, in particular, are essential for the environment, social well-being, and economic growth. Despite being abundant, the availability of drinkable water is limited. A variety of factors, including urbanization, agriculture, and industrial activities, can impact the quality of river water, making real-time monitoring crucial for ensuring its safety. In this context, anomalies are generally defined as rare events, or observations that differ significantly from the mean behavior, therefore their correct detection can help protect river systems and those who depend on them.

Machine learning (ML) is an essential component of artificial intelligence that enables a computer system to learn from data and use that knowledge to categorize or forecast future events. The methods and techniques employed in machine learning rely on complex mathematical models that can detect patterns from vast sets of data features or time-series. In this work we assess the combination of several data imputation methods and real-time incremental machine learning models for the reliable detection of anomalous water quality events.

The data studied, which spans from January 2005 to December 2022, belongs to 8 environmentally sensitive points in the Ebro River. This information was collected systematically by the two main control and monitoring networks in the Spanish river basins —SAIH (Automatic Hydrological Information Systems) and SAICA (Automatic Water Quality Information System) networks—. Both networks record data points of different variables every 15 minutes. While the SAIH network works with hydrological variables —precipitation, flow, surface water level—, the SAICA counterpart collects information on physicochemical parameters, which are indicators of water quality—water temperature, pH, dissolved oxygen, electrical conductivity, ammonium levels, and turbidity—. These data can provide valuable insights into the behaviour of the environment and early warning systems. However, the relationships between the parameters that are correlated to anomalies are not fully understood.

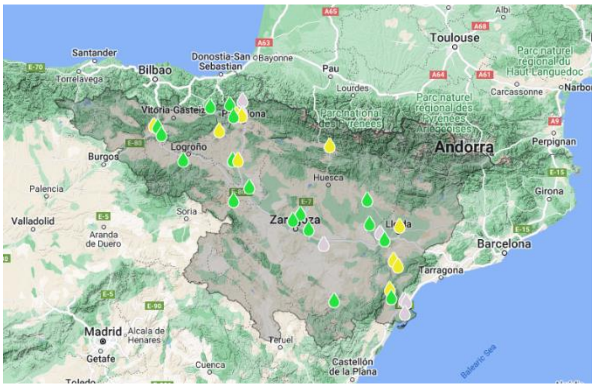

Figure 1. Ebro river basin and the SAICA and SAIH stations.

Anomalies are generally defined as rare events, or observations that differ significantly from the mean behaviour. In the case of water quality and hydrological variables, examples of anomalies include pollution events, or drastic natural changes resulting into a decrease in water quality and potential risks to aquatic life and public health. Given the availability of correctly labelled anomalies in a dataset would allow for a supervised learning approach to be used. Formally, a function of $f: X \rightarrow Y$ is learned, where X is a space of inputs and Y is the space of outputs that predicts on instances drawn from the joint probability distribution $p(x, y)$ on $X \times Y$. In the training phase, the model has access to a set of samples $(x_1, y_1), \dots (x_n, y_n)$ and attempts to minimize a loss function given by $V: Y \times Y \rightarrow \mathbb{R}$, such that $V(f(x), y)$ measures the difference between the predicted value $f(x)$ and the true value y. This is defined in Equation 1:

$$I[f] = \mathbb{E}[V(f(x), y)] = \int V(f(x), y) dp(x, y) \tag{1}$$

With an incremental, or online ML approach the corresponding model continually learns the best predictor f_t based upon new input (x_{t+1}, y_{t+1}) in the timeseries data; thereby updating model parameters. Recently, specific incremental ML learning libraries have become available, such as *River* (Montiel et al., 2021) that implement this paradigm. A key feature of such approaches is that it enables the models to adapt to drift of the underlying probability distributions without the need for ad-hoc retraining algorithms.

Data imputation has been performed with the KNNImpute algorithm (Troyanskaya et al., 2001), linear regression (Rubin, 1987; Van Buuren, 2007), missForest (Stekhoven & Bühlmann, 2012), and by pure deletion of those instances with one or more missing variables. As for the incremental ML models implemented we have compared the performance of logistic regression, Hoeffding tree, half space trees (Tan et al., 2011), an online variant of Isolation Forests (Liu et al., 2008) a stochastic implementation of the one-class SVM algorithm, Aggregated Mondrian Forest (AGF) (Mourtada et al., 2019), Adaptative Random Forest (ARF) (Gomes et al., 2017), Extremely Fast Decision Tree (Manapragada et al., 2018), and Stochastic Gradient Tree (SGT) (Gouk et al., 2019). Concept drift was detected with ADaptive WINdowing (ADWIN) (Bifet & Gavaldà, 2007).

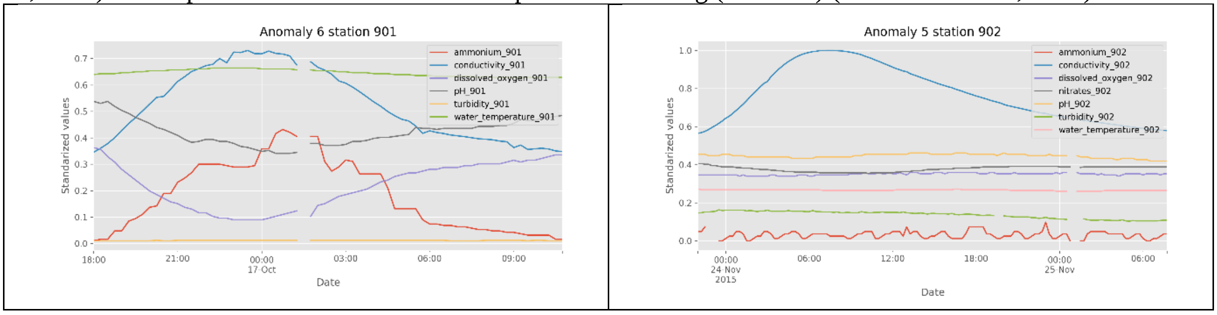

Figure 2. Multivariate representation of two different anomalies showing a complex time structure.

The results show that the best-performing combination is the imputation of the missing values with missForest, which recovers 21% of the data, and logistic regression or ARF with an accuracy over 90%. Consequently, this research demonstrates the feasibility of real-time incremental machine learning for anomaly detection in water quality and hydrological labeled data. The continuous learning approach allow the model to adapt quickly to any changing conditions in data over time, making deployment and maintenance straightforward. As such system offers a valuable tool for monitoring and managing freshwater ecosystems in real-time. Lastly, future research will focus on studying the time dependency of the anomalies and the implementation of more complex models.

Funding

This research was funded by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033), grant number PID2020-116013RB-I00

References

Bifet, A., & Gavaldà, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, 443–448. https://doi.org/doi:10.1137/1.9781611972771.42

Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfharinger, B., Holmes, G., & Abdessalem, T. (2017). Adaptive random forests for evolving data stream classification. *Machine Learning*, *106*(9), 1469–1495. https://doi.org/10.1007/s10994-017-5642-8

Gouk, H., Pfahringer, B., & Frank, E. (2019). Stochastic Gradient Trees. In W. S. Lee & T. Suzuki (Eds.), *Proceedings of The Eleventh Asian Conference on Machine Learning* (Vol. 101, pp. 1094–1109). PMLR. https://proceedings.mlr.press/v101/gouk19a.html

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Manapragada, C., Webb, G. I., & Salehi, M. (2018). Extremely Fast Decision Tree. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1953–1962. https://doi.org/10.1145/3219819.3220005

Montiel, J., Halford, M., Mastelini, S. M., Bolmier, G., Sourty, R., Vaysse, R., Zouitine, A., Gomes, H. M., Read, J., Abdessalem, T., & Bifet, A. (2021). River: Machine learning for streaming data in python. *Journal of Machine Learning Research*, *22*.

Mourtada, J., Gaïffas, S., & Scornet, E. (2019). AMF: Aggregated Mondrian forests for online learning. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *83*.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest — non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Tan, S. C., Ting, K. M., & Liu, T. F. (2011). Fast anomaly detection for streaming data. *IJCAI International Joint Conference on Artificial Intelligence*, 1511–1516. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-254

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*, 219–242. https://doi.org/10.1177/0962280206074463