Science of the Total Environment 749 (2020) 142368

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

A novel dynamic multi-criteria ensemble selection mechanism applied to drinking water quality anomaly detection

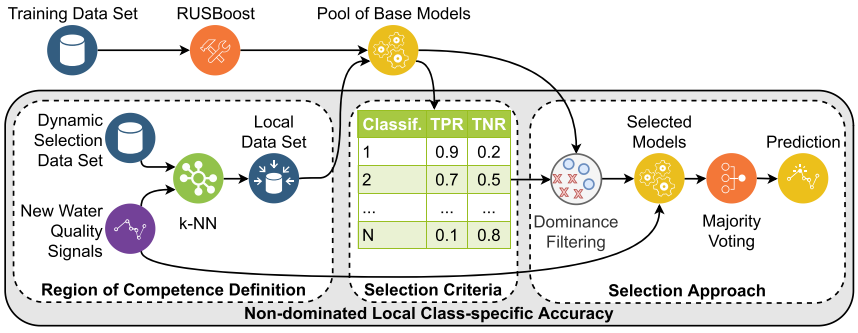
Victor Henrique Alves Ribeiro [a,*], Steffen Moritz [b], Frederik Rehbach [b], Gilberto Reynoso-Meza [a]

[a] Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS), Pontifícia Universidade Católica do Paraná (PUCPR), Rua Imaculada Conceição, 1155, 80215-901 Curitiba, PR, Brazil
[b] Institute of Data Science, Engineering, and Analytics, TH Köln, Campus Gummersbach, Steinmüllerallee 1, 51643 Gummersbach, Germany

HIGHLIGHTS

- The solution for a real-world drinking water anomaly detection problem is presented.
- Feature engineering and dynamic ensemble selection are explored to solve the task.
- A novel multi-criteria dynamic ensemble selection algorithm is proposed.
- The new algorithm outperforms all other tested dynamic ensemble selection methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 June 2020
Received in revised form 8 September 2020
Accepted 11 September 2020
Available online 15 September 2020

Editor: Ashantha Goonetilleke

Keywords:
Anomaly detection
Drinking water quality
Time series classification
Machine learning
Ensemble learning
Dynamic ensemble selection

ABSTRACT

The provision of clean and safe drinking water is a crucial task for water supply companies from all over the world. To this end, automatic anomaly detection plays a critical role in drinking water quality monitoring. Recent anomaly detection studies use techniques that focus on a single global objective. Yet, companies need solutions that better balance the trade-off between false positives (FPs), which lead to financial losses to water companies, and false negatives (FNs), which severely impact public health and damage the environment. This work proposes a novel dynamic multi-criteria ensemble selection mechanism to cope with both problems simultaneously: the non-dominated local class-specific accuracy (NLCA). Moreover, experiments rely on recent time series related classification metrics to assess the predictive performance. Results on data from a real-world water distribution system show that NLCA outperforms other ensemble learning and dynamic ensemble selection techniques by more than 15% in terms of time series related F_1 scores. As a conclusion, NLCA enables the development of stronger anomaly detection systems for drinking water quality monitoring. The proposed technique also offers a new perspective on dynamic ensemble selection, which can be applied to different classification tasks to balance conflicting criteria.

1. Introduction

Water is essential to life sustenance. Access to and availability of good quality water have been established as critical factors for both national security and public health, having a significant impact on life expectancy (Dogo et al., 2019). In 2010, the United Nations General Assembly recognized the access to safe and clean drinking water as a human right through Resolution 64/292 (UN General Assembly, 2010).

* Corresponding author.
E-mail addresses: victor.henrique@pucpr.edu.br (V.H. Alves Ribeiro), steffen.moritz@th-koeln.de (S. Moritz), frederik.rehbach@th-koeln.de (F. Rehbach), g.reynosomeza@pucpr.br (G. Reynoso-Meza).

https://doi.org/10.1016/j.scitotenv.2020.142368
0048-9697/

Therefore, the provision of clean and safe drinking water is a critical task for water supply companies from all over the world (Rehbach et al., 2018). However, water distribution systems (WDS) are highly sensitive to contamination, which can disseminate many water-related diseases. Recently, there have been reports of drinking water quality issues in the city of Rio de Janeiro, Brazil, where muddy and foul-smelling water affected more than 6 million people (Phillips, 2020; Barbon, 2020).

Large scale water suppliers often provide hundreds of thousands, if not millions of people with clean drinking water. While this allows for efficient processes, it also means that unrecognized drinking water contamination could sicken equally large amounts of people. A variety of natural and anthropogenic sources can cause water contamination, which includes impurities at the water source, problems at the water treatment plant, and defects in the water network. Thus, continuously surveying the water quality throughout the supply grid is essential for ensuring the populations' safety.

There is an expectation of a further increase in the importance of early recognition in the next years. Increasing numbers of extreme weather events caused by climate change (Hansen et al., 2012) pose further challenges to water suppliers (Castell-Exner and Zenz, 2010). Additionally, demographic changes may lead to problems with polluted water sources and wrongly dimensioned water networks. As an example, in industrialized nations like Germany, a combination of shrinking populations and more economical use of drinking water is likely to result in current water networks being oversized in the future. With slower water flow rates than initially planned, depositions can build up in the pipes. When demand peaks increase the flow rate, these depositions are refloated and subsequently pose a risk to the customers (Korth et al., 2007).

Beyond the early recognition of accidental contaminations, anomaly detection also plays a role as a precautionary measure against deliberate manipulations of the drinking water. Since contaminated water is a comparatively easy way to affect a large portion of the population, water supply networks must be considered as a potential target of terrorist attacks. Thus, government agencies like the National Homeland Security Research Center of the US EPA are actively developing and looking at anomaly detection systems as counter-terrorism measures (Hart et al., 2007).

Currently, water quality is monitored by water suppliers with rather simple tools. One common approach is that engineers occasionally check the current situation during the day by looking at visualizations of the sensor data on a monitor. In addition to this, each variable has a fixed threshold. If one parameter surpasses its threshold value for a defined period, it triggers an alarm.

This approach has two main drawbacks: First, to avoid vast numbers of false alarms, the thresholds are usually set quite high. Second, to prevent false alarms caused by momentary outliers in the sensor data, these set-points must be surpassed for a defined period. Therefore, in case of a real contamination scenario, the implemented system will likely miss the build-up phase of the event and trigger an alarm rather late.

During the day, an observant engineer can recognize these slight deviations and abnormal patterns before an event and trigger an early alarm. However, usually, the engineer's main task is not to continuously monitor the system all day long. To ensure early detection of possible events independent of office hours and working times, a more advanced automated anomaly detection system is required. Ideally, such a system is capable of accurately identifying contamination events in the same way or even surpassing the accuracy of the local engineer, while performing better than the described threshold technique.

Recent literature has focused on developing reliable automatic anomaly detection systems. Most studies are based on machine learning techniques, while only Hernandez-Ramirez et al. (2019) perform statistical analysis for factors and outliers in Atoyac River, Mexico. Deng and Wang (2017) use a hybrid model composed of k-nearest neighbors (k−NN) for similarity detection and autoregressive integrated moving average (ARIMA) model for signal modeling in Yangtze River, China. In Potomac River, USA, Shi et al. (2018) employ artificial neural networks (ANNs) for regression and residuals analysis, while Liu et al. (2020) use Bayesian autoregressive (BAR) model for signal modeling and isolation forest (IF) to detect residuals' outliers. Leigh et al. (2019) compare different classification and regression methods for anomaly detection in the Rivers from the Great Barrier Reef, in Australia. Finally, a recent industrial competition for anomaly detection in drinking water quality by Rehbach et al. (2018) inspired contributions using bidirectional long short term memory (LSTM) with convolutional neural network(CNN) ensembles (Chen et al., 2018), LSTM (Fehst et al., 2018), ANN, logistic regression (LR), and support vector machine (SVM) (Muharemi et al., 2019), k-NN and ANN (Ali et al., 2019), deep learning (DL) (Dogo et al., 2019), and ensemble learning (Ribeiro and Reynoso-Meza, 2018).

Therefore, there is an active research field in such more advanced systems. Yet, until now, they are only slowly being adapted into practice by water suppliers. A common worry among water suppliers is that the benefits of such systems do not counterbalance the effort required. Not only a distributed online sensor network throughout the water grid costs a significant amount of money, but it also must be maintained regularly, and every false alarm entails additional effort for the engineers. The intended earlier alarm signaling also implies a higher chance of triggering a false alarm because the pattern might not yet be as unambiguous as some minutes later. Engineers might simply begin to ignore them altogether if false alarms become too frequent. Thus, the crux for being accepted and implemented by water suppliers is providing a reasonable false negative (FN) to false positive (FP) ratio.

To develop strong classification systems for the task, ensemble learning techniques are recommended given their stronger predictive performance (Sagi and Rokach, 2018). In addition to this, dynamic ensemble selection (DES) techniques can further improve classification results (Cruz et al., 2018). However, usual DES techniques do not consider the trade-off relation between FPs and FNs, which can lead to “wrongly neglecting certain aspects of realism” for the given task (Roy, 2016). Therefore, the following contributions are highlighted.

- This work proposes the novel nondominated local class-specific accuracy (NLCA) technique, the first DES mechanism that considers multiple conflicting criteria to achieve a better trade-off between FPs and FNs in binary classification problems.
- Recent time series related precision and recall metrics are employed to analyze the results of the proposed technique on data from a real-world problem regarding anomaly detection in drinking water quality.
- In comparison to state-of-the-art DES techniques (META-DES and DES-P), NLCA achieves the best results considering scenarios that reward both early warning and recall in anomaly detection.

The remainder of this work is organized as follows: Section 2 presents the materials and methods, such as the case study and data collection, the feature engineering steps, the classification procedures for anomaly detection, the proposal of the novel NLCA mechanism for DES, and the evaluation metrics. Next, Section 3 discusses the experiments and results. Finally, Section 4 concludes the paper with some final remarks and future research directions.

2. Materials and methods

This section brings the materials and methods. First, Subsection 2.1 describes the real-world problem addressed in this work along with the data collection procedures. Next, Subsection 2.2 details the necessary feature engineering steps for the problem at hand in terms of signal processing, feature extraction, and feature selection. Subsequently, Subsection 2.3 details the machine learning techniques compared in the proposed case study, such as ensemble learning and its dynamical

procedures, the proposal of the novel NLCA mechanism, and a final output filtering step. Finally, Subsection 2.4 formulates the evaluation metrics to analyze and compare the tested models.

2.1. Real-world problem description

To test the performance of an anomaly detection system, adequate and realistic data is required. This work uses a real-world data set generated in a research project on drinking water safety and energy efficiency,[1] which included several German water suppliers. More specifically, the data set consists of data from Thüringer Fernwasserversorgung, a major German water supplier located in central Germany. The company supplies over a million people with high-quality drinking water and operates more than 60 dams and reservoirs. The data set is a particularly suitable choice for testing new algorithms, since a series of competitions about anomaly detection for drinking water was held in major international conferences (Rehbach et al., 2018). Thus, there are already some results and papers available to get a good baseline of what is possible to achieve. The data and additional documentation are available for download for anyone interested (Moritz et al., 2018).

The data is recorded at the water suppliers (Thüringer Fernwasserversorgung) measurement stations with sensor panels. It includes sensor values of turbidity (TURB), water temperature (WT), pH value, chlorine dioxide (ClO2), Redox potential, water flow rate (FR), and electrical conductivity (EC) recorded in one-minute intervals over six months, as well as the EVENT target that labels the anomalous data (Table 1). Some minor technical preprocessing is done on a Raspberry Pi next to the sensor panel, which is required for data transmission. New measurements are then directly transferred via a secure transmission channel to a server, where the data is processed and anomaly detection can be performed. Results from the anomaly detection are intended to be displayed in a control room, which is staffed 24 h a day. The measurement frequency of 1 min implies that online operation requires the current time step to be processed within 1 min. Otherwise, the queue would increase with unprocessed data.

Luckily, serious contaminations are extremely rare. Most of the time, everything runs as expected, and even minor problems seldom occur. On the downside, this means just the recorded data alone does not suffice to test the capabilities of an anomaly detection algorithm. Therefore, a small-scale water network testing-grid at a TH Köln research facility, equipped with the same sensor panels as the Thüringer Fernwasserversorgung grid, aided the modeling of artificial events. Such a testing-grid enabled the addition of contaminants and subsequent recording of the sensor patterns during the aforementioned research project. These patterns were then projected to the original data with different signal strengths. Thus, real events (expert labeling) and artificial events (simulation) populate the anomaly label of our data set to train more robust machine learning models.

Since the data was recorded in the water network of Thüringer Fernwasserversorgung, it also features operational issues that come along with online sensor measurements. There are occurrences of missing values and outliers due to sensor outages and maintenance present in the data. An anomaly detection algorithm must deal with these problems and should still produce reasonable results.

The goal of the water supplier is to use this monitoring data to quickly recognize problems in the water network to increase overall safety. Therefore, the anomaly detection system must not miss major contamination events. At the same time, it must also avoid too many false alarms.

Fig. 1 plots the training data set for the nine available sensor measurements in the drinking water quality problem. The images on the left show a broad overview of the time series, while the close-ups on the right focus on anomalous data, which can occur in many different fashions for any of the given variables. Also, anomalies in our data are not point-based, but rather range-based, lasting for several minutes, even hours. Typical noticeable patterns are steep increases/decreases and more subtle, slow changes of sensor values. These patterns can be seen in one or multiple sensors at a time and can have different signal strengths. Also, alternating high/low patterns occur in the data (like in the zoomed time series for $ClO2_1$). In real drinking water time series, alternating high/low patterns are highly uncommon (up to non-existent). However, they are an important test for anomaly detection algorithms. During these patterns, the moving average stays the same, which poses a challenge for some algorithms. Even if uncommon, algorithms should not miss such events should they ever occur. Overall, the data features a mix of easy and rather difficult detectable anomalies.

Table 1
Available parameters from the Thüringer Fernwasserversorgung dataset.

Parameter	Unit	Description
Time	*datetime*	Time Stamp
WT	*°C*	Water Temperature
$ClO2_1$, $ClO2_2$	*mg/l*	Chlorine Dioxide (2 values)
pH	*pH*	pH Value
Redox	*mV*	Redox Potential
EC	*µS/cm*	Conductivity
TURB	*NTU*	Turbidity
FR_1, FR_2	m^3/h	Water Flow Rate (2 values)
EVENT	*binary*	Anomaly Label

2.2. Feature engineering

The case study presents the classification task of a multivariate time series, which suffers from issues related to missing data, concept drift, and class imbalance (Souza et al., 2016; Krawczyk, 2016). Missing data is usually caused by data acquisition issues, such as sensor malfunction or network connection errors. Concept drift is related to a change in the relation between the input and the output variables, which is not modeled during training. Trends in the input data can cause such an issue. Class imbalance indicates that the number of observations between different classes differs significantly. Therefore, the raw data is not able to provide reliable and meaningful information to infer a classification model that generalizes well to future data. Nevertheless, feature engineering steps are important to "formulate the most appropriate features given the data, the model, and the task" (Zheng and Casari, 2018).

This work performs the following three steps: (1) signal preprocessing, where imputing and detrending operations mitigate issues related to missing data and concept drift; (2) feature extraction, which computes and combines statistical features; and (3) feature selection, which focus on dimensionality reduction.

2.2.1. Signal processing

As mentioned previously, signal processing is necessary to handle missing data and concept drift. Imputing is the first operation employed in this work, which handles missing data. There are many possible solutions to this, such as using the mean or median value of the whole data set (Zheng and Casari, 2018). Since this work deals with time-related data, and concept drift is existent, missing values ($\varnothing$) are imputed using the previously available data point for each feature f as follows.

$$x_f(t) = \begin{cases} x_f(t), & x_f(t) \neq \varnothing \\ x_f(t-1), & \text{otherwise} \end{cases} \quad (1)$$

Next, it is important to remove time-related distortions from the signals. Despite being common for physical measurements to present time-related variations, such as temperature across the year, fluctuations can deteriorate the predictive performance of machine learning

[1] Research Project 'IMProvT' (2015–2019) https://www.th-koeln.de/informatik-und-ingenieurwissenschaften/_56166.php.

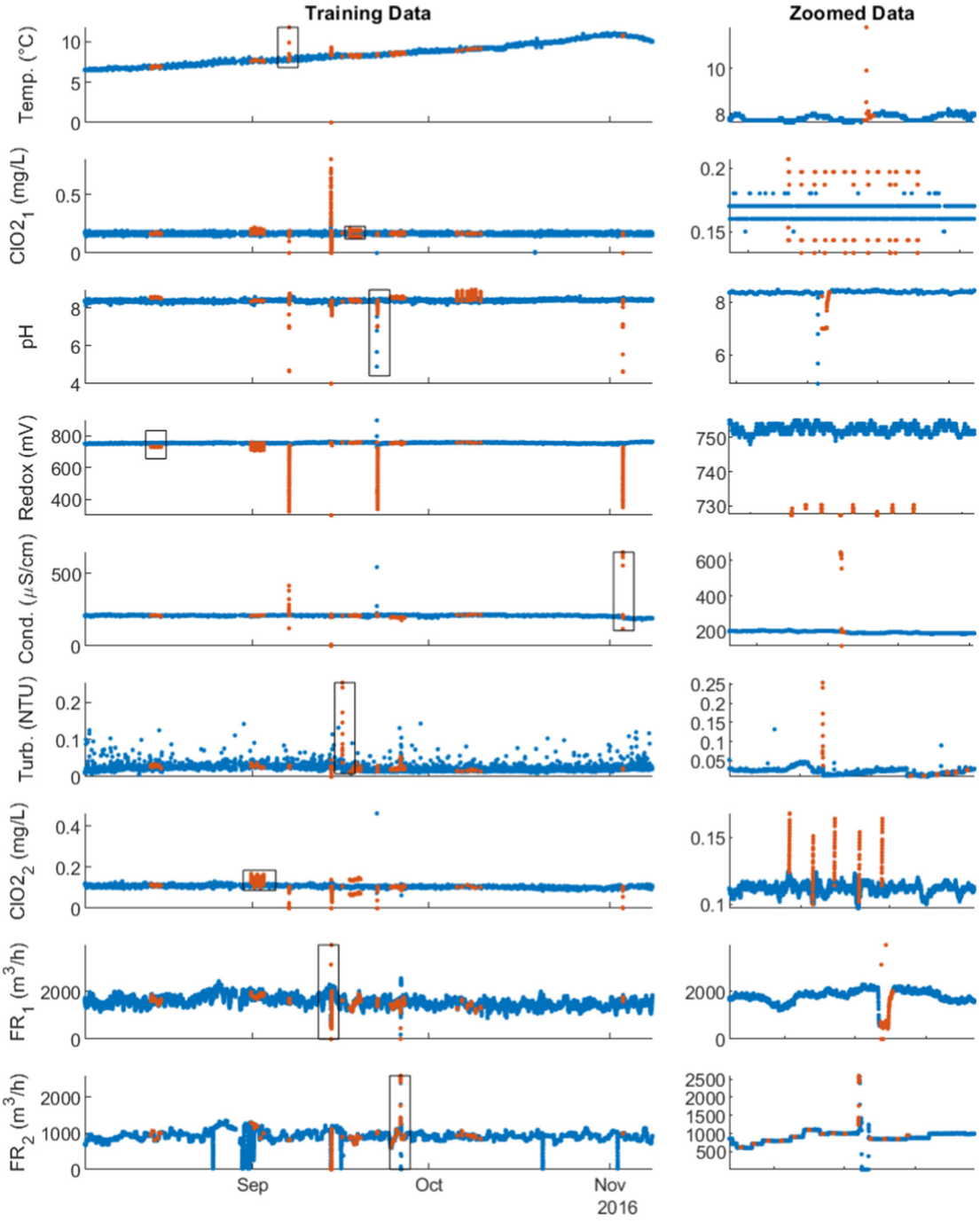


Fig. 1. Training data for the drinking water quality anomaly detection problem. Blue points indicate data under normal operation, while red points indicate anomalies. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

models. Detrending is performed by removing the simple moving average ($\bar{x}_f^{sm}(t,w)$) with a time window of w=1440 min, or 24 h, as follows.

$$x_f(t) = x_f(t) - \bar{x}_f^{sm}(t,w) \tag{2}$$

where

$$\bar{x}_f^{sm}(t,w) = \frac{1}{w}\sum_{i=0}^{w-1} x_f(t-i) \tag{3}$$

2.2.2. Feature extraction

One interesting feature to compute for a time series $x(t)$ is the first order difference to the previous observation $\Delta x(t)$. Such a feature captures the high frequency variations in the signal, being a discrete approach to the differentiation operation ($\frac{dx}{dt}$). Nevertheless, the difference signal is sensitive to noisy signals. The difference is computed according to Eq. (4) for each raw signal f, generating F=9 additional features.

$$x_{F+f}(t) = \Delta x_f(t) = x_f(t) - x_f(t-1) \tag{4}$$

Moreover, statistical features are computed to obtain abstraction from the raw signals. Given a time series $x(t)$, the mean ($\bar{x}(t)$) captures the series' tendency, aiding in filtering noisy signals. The standard deviation ($x^{\sigma}(t)$) captures the data variability. The maximum ($x^{max}(t)$) and minimum ($x^{min}(t)$) indicate the signal's peak values. Finally, the median

($x^{med}(t)$) detects the value that splits the signal's values in half. Such features are computed as follows, using a time window of l=30 min.

$$\overline{x_f}(t) = \sum_{i=0}^{l-1} x_f(t-i)/l \tag{5}$$

$$x_f^{\sigma}(t) = \sqrt{\frac{1}{l-1}\sum_{i=0}^{l-1}\left(x_f(t-i)-\overline{x_f}(t)\right)^2} \tag{6}$$

$$x_f^{max}(t) = \max\left\{x_f(t), x_f(t-1), \ldots, x_f(t-l+1)\right\} \tag{7}$$

$$x_f^{min}(t) = \min\left\{x_f(t), x_f(t-1), \ldots, x_f(t-l+1)\right\} \tag{8}$$

$$x_f^{med}(t) = \mathrm{med}\left\{x_f(t), x_f(t-1), \ldots, x_f(t-l+1)\right\} \tag{9}$$

Finally, to obtain a higher level of abstraction, the following features are computed for each of the 18 signals (the 9 original signals plus their differences): (1) the standard deviation; (2) the difference between the signal at the current time and 30 min before; (3) the difference between the signal at current time and the mean; (4) the total signal amplitude; (5) the difference between the maximum and current value; (6) the difference between the current value and the minimum; and (7) the difference between the median and mean values. As a result, the seven features for each of the 18 signals generate a total of 126 features.

$$X_f(t) = \begin{vmatrix} x_f^{\sigma}(t) \\ x_f(t)-x_f(t-l+1) \\ x_f(t)-\overline{x_f}(t) \\ x_f^{max}(t)-x_f^{min}(t) \\ x_f^{max}(t)-x_f(t) \\ x_f(t)-x_f^{min}(t) \\ x_f^{med}(t)-\overline{x_f}(t) \end{vmatrix} \tag{10}$$

2.2.3. Feature selection

The high number of features makes it difficult for learning algorithms to generalize a model to new observations, which is known as the *curse of dimensionality* (Domingos, 2012). Moreover, data could be affected by multicollinearity. However, tree-based models have shown to be robust to such a problem in terms of predictive performance (Piramuthu, 2008; Kotsiantis, 2013; Tomaschek et al., 2018). Therefore, to solve both issues, random forests (RF) (Breiman, 2001) is employed as an embedded feature selection model, which has shown good results for imbalanced classification problems (Hasan et al., 2016). To this end, the algorithm is configured with 100 deep decision trees, randomly selecting subsets of features from all the 126 available ones at each split. Additionally, the training data is balanced using random under sampling (Seiffert et al., 2009) before training the RF. Features are selected (f^s) according to the features' computed importances f^i given the following rule for each feature k.

$$\mathrm{f_k^s} = \begin{cases} 1, & \mathrm{f_k^i} \geq \sum_{k=1}^{K} \mathrm{f_k^i}/K \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

2.3. Classification

After the feature engineering step, the classification task can be performed for anomaly detection. Classification is a task that learns how to map a set of inputs x to an output y, given that $y \in [c_1, \ldots, c_n]$ for a problem with n classes (Murphy, 2012). To this end, such a task typically consists of a data set, an inducer, and a classifier, where the inducer is a learning algorithm that generates a classifier for the given data set. Additionally, it is possible to combine the output of different classifiers with an ensemble methodology to improve classification performance (Sagi and Rokach, 2018).

2.3.1. Ensemble learning

The ensemble learning methodology is composed of three steps: pool generation, where a mechanism trains diverse base models; selection, where another component prunes the ensemble; and combination (or aggregation), where some selected rule, such as majority voting, weighted voting, maximum probability, among others, combine the outputs (Sagi and Rokach, 2018).

For the pool generation step, this work employs random undersampling boosting (RUSBoost) (Seiffert et al., 2009). Such an algorithm is an adaptation of the famous adaptive boosting (AdaBoost) (Freund and Schapire, 1997), but is designed for imbalanced data sets. At each learning cycle, data from the majority class is randomly under sampled to balance the classes. According to high classification scores achieved in Ribeiro and Reynoso-Meza (2020), RUSBoost is configured with 200 shallow decision trees, using a maximum of 10 decision splits, and a ratio of 2 majority class observations for each minority class observation.

After the pool generation step, it is possible to simply combine the outputs of all the trained base models. However, studies have shown that selecting only the best of such models can improve classification performance (Sagi and Rokach, 2018). Nevertheless, the most successful selection approaches rely on dynamical mechanisms (Britto Jr et al., 2014; Cruz et al., 2018).

2.3.2. Dynamic classifier and ensemble selection

Dynamic selection approaches are employed in ensembles to select different classifiers on the fly, based on the characteristics of each predicted sample (Britto Jr et al., 2014; Cruz et al., 2018). The term dynamic highlights the adaptability of the new ensembles in contrast to the static behavior existent in usual ensemble models, where the same classifiers are always employed. To enable the dynamic behavior, such mechanisms consist of the following steps: region of competence definition, which identifies the similarity between known instances and the new observation; selection criteria, which computes evaluation metrics for each classifier on the region of competence; and selection approach, which selects one or more classifiers to predict the output for the new observation (Cruz et al., 2018).

Initial techniques relied on dynamic classifier selection (DCS), where only the single best classifier is employed for each new sample. For instance, Woods et al. (1997) define the region of competence for each new sample and compute the overall local accuracy (OLA) for each base classifier, selecting the one with the highest score. Most recently, Ribeiro et al. (2020) presented the overall local class-specific accuracy (OLCA) as an evolution of OLA, where the preference ranking organization method for enriched evaluation (PROMETHEE) (Brans et al., 1986; Brans and De Smet, 2016) technique ranks the base models given their local true positive ratio (TPR) and true negative ratio (TNR) scores.

Different from DCS, which only dynamically selects one classifier, DES selects and combines multiple classifiers for each new testing sample. Currently, META-DES (Cruz et al., 2015) can be considered as one of the state-of-the-art DES techniques. Such an algorithm uses a trainable classifier to select which base models shall be combined to form a final prediction. As input, the classifier uses selection criteria employed by other DCS and DES techniques, namely OLA (Woods et al., 1997), local class accuracy (LCA) (Woods et al., 1997), a posteriori (Giacinto and Roli, 1999), and k-nearest output profiles (KNOP) (Cavalin et al., 2013). Cruz et al. (2015) indicate that in case one of the selection criteria fails, the others shall still enable the selection of the proper ensemble. Another powerful dynamic selection technique is dynamic ensemble selection based on performance (DES-P), which selects all the base classifiers that achieve a better predictive performance than a random classifier on the region of competence (Woloszynski et al., 2012).

Despite this, to the present authors' knowledge, there are still no DES mechanisms that take into account the trade-off relation between multiple conflicting selection criteria. Complex engineering problems tend to have multiple, often conflicting criteria. The drinking water quality anomaly detection presents a binary classification task, where there is a trade-off between the number of FP and FN predictions. In such a complex problem, it is difficult to generate a single model that achieves a perfect predictive performance. Instead, there exist many different solutions that present distinct trade-off relations between the number of FPs and FNs. Therefore, when building a DES mechanism, using a single criterion could lead to "wrongly neglecting certain aspects of realism" (Roy, 2016). By considering both the problems of FPs and FNs, a DES mechanism could benefit from achieving a more desirable trade-off. To this end, this work proposes the novel NLCA technique.

2.3.3. Nondominated local class-specific accuracy

This work proposes the novel NLCA mechanism. In contrast to existing DES techniques, such an algorithm considers multiple conflicting selection criteria, namely TPR and TNR. Also, the selection approach employs dominance filtering to select only Pareto-optimal base models. This enables NLCA to achieve a better trade-off between the number of FPs and FNs for the anomaly detection problem. The following steps clearly detail the proposed solution using NLCA in terms of pool generation, region of competence definition, selection criteria, selection approach, and combination, which are also depicted in Fig. 2.

1. **Pool generation**: Given a training data set (X_{train}), this work employs RUSBoost (Seiffert et al., 2009) according to Subsection 2.3.1, which results in a pool of $M = 200$ base models.
2. **Region of competence definition**: Given a dynamic selection data set (X_{dsel}) and a new sample to be classified (x_{test}), NLCA uses k-NN to build a local data set (X_{local}) with 20 instances from X_{dsel} that present the shortest distances to xtest.
3. **Selection criteria**: The 200 available base models classify the instances from X_{local}. With the predictions, NLCA computes the TPR and TNR for each base model. Such selection criteria indicate the accuracies for the positive and negative classes, being computed as follows given the numbers of true positives (TPs), true negatives (TNs), FPs, and FNs.

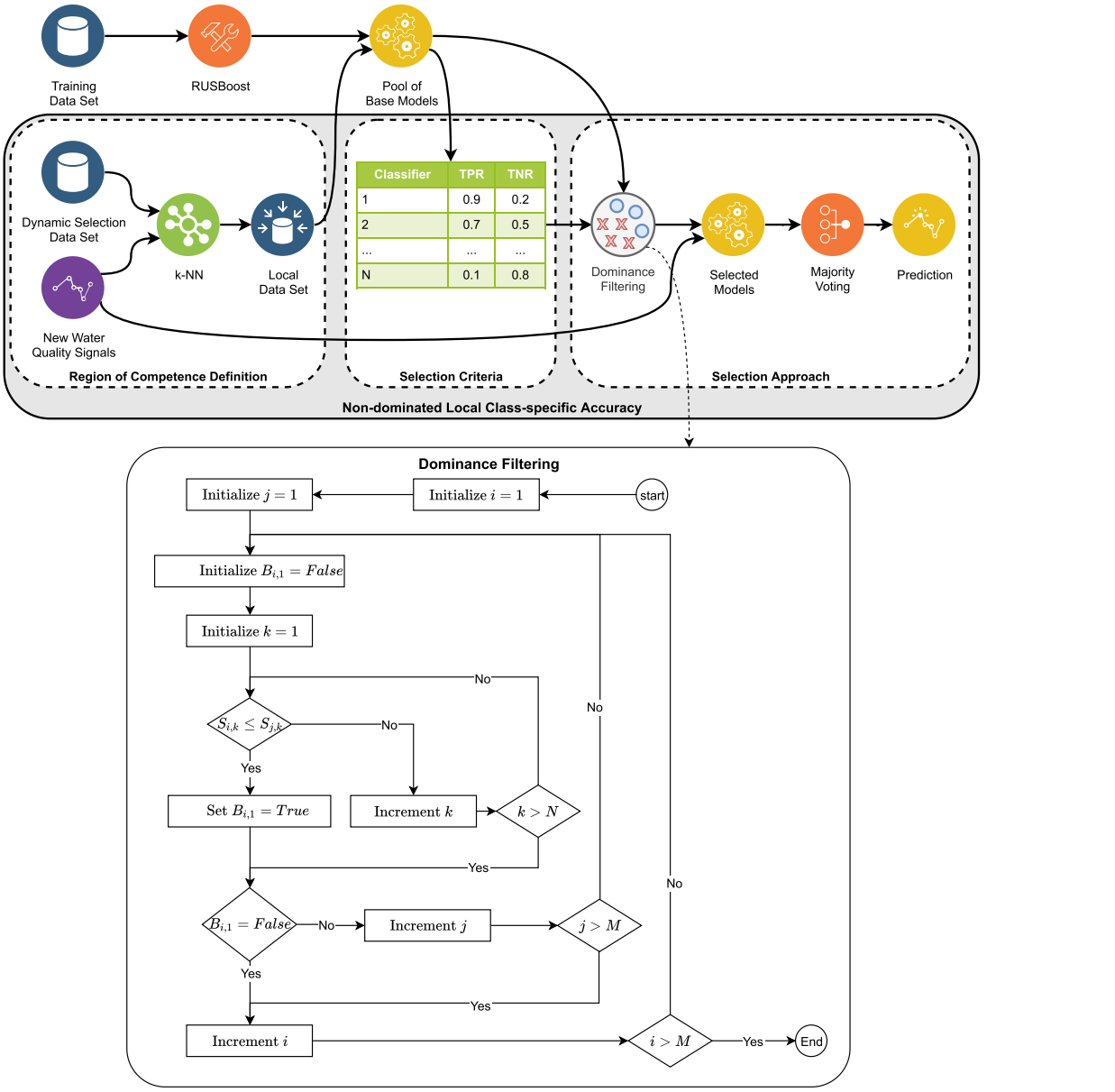


Fig. 2. The proposed non-dominated local class-specific (NLCA) dynamic multi-criteria ensemble selection method.

$$TPR = TP/(TP + FN) \tag{12}$$

$$TPR = TP/(TP + FN) \tag{13}$$

4. **Selection approach**: NLCA employs a dominance filter (Fig. 2) to select the non- dominated base-models, which uses the previously computed selection criteria in a matrix $S_{M,\ N}$ of M models and N criteria to define a Boolean dominance vector $B_{M,\ 1}$ of M models. With this procedure, NLCA selects only the best base models in terms of both TPR and TNR, providing a better trade-off between FNs and FPs for anomaly detection.
5. **Combination**: The selected base models predict the output for x_{test}. Next, NLCA uses the majority voting rule (Eq. (14)) or the weighted voting rule (Eq. (16)) to compute the final output. In the majority voting method, $y_k(x)$ is the prediction of the k^{th} classifier for input x, while ci is the i^{th} class. In the weighted voting method, $p_{k,\ i}(x)$ is the prediction probability of the kth classifier for input x and ith class c_i, while wk. is the weight assigned by RUSBoost to the k^{th} classifier.

$$class(x) = \underset{c_i \in\ dom(p)}{\operatorname{argmax}} \left(\sum_k g(y_k(x), c_i)\right) \tag{14}$$

given

$$g(y, c) = \begin{cases} 1, & y = c \\ 0, & y \neq c \end{cases} \tag{15}$$

or

$$class(x) = \underset{c_i \in\ dom(p)}{\operatorname{argmax}} \left(\sum_k p_{k,i}(x) \cdot w_k\right) \tag{16}$$

Similar to its predecessor, OLCA (Ribeiro and Reynoso-Meza, 2020), NLCA uses k-NN to define the region of competence, while TNR and TPR are the employed selection criteria. However, the selection approach differs significantly from one another. OLCA is a DCS technique that uses an additional multi-criteria decision-making (MCDM) technique to dynamically select a single classifier for each tested sample. NLCA, on the other hand, is a DES technique that uses a simple dominance filter to select multiple classifiers for each tested sample, which are combined through majority voting. Therefore, NLCA has two main advantages over OLCA. First, it presents better results due to the combination of multiple classifiers (Sagi and Rokach, 2018; Cruz et al., 2018). Second, NLCA only needs to configure the number of desired neighbors in the local pool while OLCA needs to configure additional parameters for the embedded MCDM technique.

In contrast to the current state-of-the-art DES techniques, DES-P and META-DES, NLCA differs mostly in terms of the selection criteria and selection approach. Regarding the selection criteria, the proposed algorithm is the first DES method that considers two conflicting objectives, the TPR and TNR. Regarding the selection approach, the dominance filter enables the selection of base models that are considered Pareto-optimal in terms of both selection criteria. With such a procedure, NLCA presents the advantage over META-DES and DES-P of balancing the trade-off between TPR and TNR in the drinking water quality anomaly detection problem, achieving superior results.

2.3.4. Output filtering

After the classification step using ensemble learning and NLCA, output filtering is analyzed for further improvements due to FPs reduction in noisy situations. On the downside, the filter causes detection lag. The filtered output $y_f(t)$ is computed as follows given the model's output $y(t)$ and a window of the lf previous minutes. Such a filter performs a majority voting scheme by calculating the mean of the previous l_f outputs.

$$y_f(t) = \begin{cases} 1, & \left(\sum_{i=0}^{l_f-1} y(t-i)/l_f\right) \geq 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

2.4. Evaluation metrics

Finally, it is important to define the evaluation criteria employed to compare the models. One of the main evaluation metrics for imbalanced data sets is the F_1 score (Ribeiro and Reynoso-Meza, 2020). Such a score is the harmonic mean between precision and recall, computed according to Eq. (18). On the one hand, recall (or TPR) indicates the proportion of positive samples that are correctly predicted as such. On the other hand, precision indicates the number of positive predictions that are correct. This work considers the positive class as the anomalous data.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{18}$$

where

$$precision = TP/(TP + FP) \tag{19}$$

$$recall = TPR = TP/(TP + FN) \tag{20}$$

Such a metric is principally concerned with point-based data. That is, the F_1 score does not take into account the relation between subsequent data, or ranges. However, the problem at hand does not present point-based anomalies, but rather ranges where the anomalous data can be found. Therefore, this work employs time series adaptations of F_1 score, precision, and recall (Tatbul et al., 2018).

Different from the point-based precision and recall, the range-based metrics consider full sequences of anomalies and predictions for computing the scores. With such, the scores can be modified to task-specific objectives by configuring three different parameters. For instance, the metrics can reward existence (α), position (δ), and cardinality (γ). The existence term (α) can be used to reward the detection of an event, even if only a single point in the whole range is predicted correctly. The positional bias (δ) can be used to reward early or late predictions. Finally, the cardinality term (γ) is employed to penalize the overlapping of multiple predictions. The full formulation for the range-based metrics, precision(α, δ, γ) and recall(α, δ, γ), can be found in Tatbul et al. (2018).

This work makes use of three different configurations of F_1 score, precision and recall, being: (1) classical, using the point-based metrics; (2) optimistic, where any detection is rewarded with the $\alpha = 1$ configuration; and (3) early warning, where the early detection is focused by removing the existence term ($\alpha = 0$). The front-end positional bias function for recall (δ_{recall}) and precision ($\delta_{precision}$) are detailed in Eqs. (21) and (22), respectively, where l is the total length of a sequence of continuous anomalous or normal data and p is the position of the current point being analyzed in such a sequence. Finally, the cardinality term is defined as $\gamma = 1$ for both the optimistic and early warning configurations.

$$\delta_{recall}(p, l) = l - p + 1 \tag{21}$$

$$\delta_{precision}(p, l) = 1 \tag{22}$$

Fig. 3 depicts a fictional anomaly detection scenario, where a detection delay of three samples is found. Additionally, the prediction incorrectly detects two normal samples after the anomalous points. For such a scenario, the different configurations will compute different scores. In the classic configuration, $F_1 = 0.6667$, Recall $= 0.6250$, and Precision $= 0.7143$. In the optimistic configuration, $F_1 = 0.8333$, Recall $= 1.0000$, and Precision $= 0.7143$. Finally, in the early warning configuration, $F_1 = 0.5263$, Recall $= 0.4167$, and Precision $= 0.7143$. Such

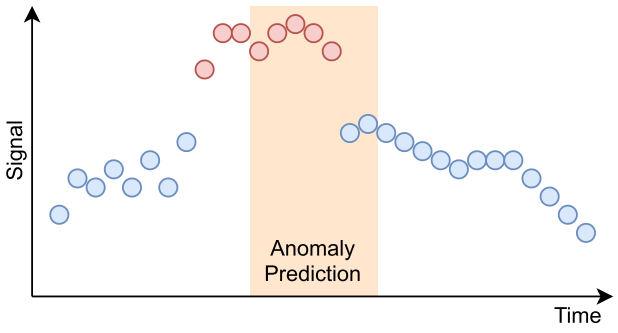


Fig. 3. Fictional time series anomaly detection scenario. The blue dots indicate data under normal operation, while red dots indicate anomalous data. The yellow area indicates the classifier output. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

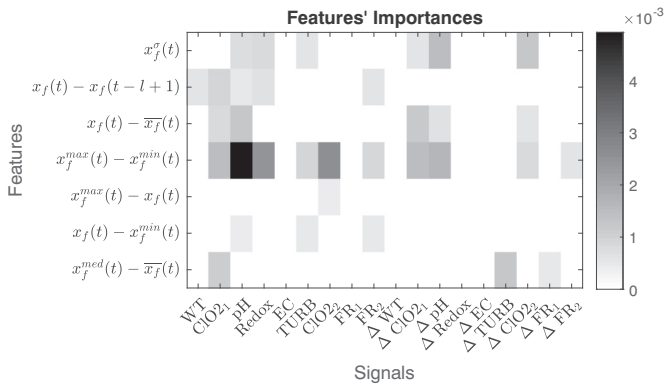


Fig. 4. Relevance of each specific signal and feature for the drinking water quality anomaly detection problem. Colors indicate the features for each signal.

values clarify the existence reward in the optimistic configuration and the delay penalization for the early warning configuration in the recall and, consequently, F_1 scores.

3. Results and discussion

This section discusses the results for anomaly detection in drinking water quality using the proposed methodology and NLCA. First, the feature engineering steps are analyzed. Next, results from the comparison of NLCA with other state-of-the-art dynamic selection mechanisms are shown. Subsequently, the effects of filtering the classifiers' outputs are compared. All results are analyzed in terms of the evaluation criteria (Section 2.4). Finally, the results are summarized along with remarks regarding implementation in different water networks.

3.1. Feature analysis

The first analysis demonstrates how feature engineering is able to improve the results for the anomaly detection task. In total, four scenarios are considered, one with only data imputing (Section 2.2.1), the next including signal detrending (Section 2.2.1), the following including feature extraction (Section 2.2.2), and a final one including feature selection (Section 2.2.3). It is important to mention that validation of such steps is performed using two-fold cross-validation with the whole training data. Such a configuration is selected to simulate the test scenario, where half of the available observations are used for training and the other half for testing.

Table 2 depicts the results, where the best values are marked in bold. Classification using only the raw data (with imputing) achieves the worst results for all F1 and precision configurations. Next, the detrended data results demonstrate the effect of removing concept drifts for time series classification, where the best precision results and second best F1 scores are attained. Once the new features are created, recall results are improved, but both precision and F1 scores drop in comparison to the detrended data. Finally, after the feature selection step, results indicate improvements in both F1 and Recall values, where the best results are achieved.

The presented results confirm the importance of the feature engineering process. First, it is interesting to notice how the F_1 scores for the classification of the raw data are low. This is due to the low precision of the predictions, where many FPs are generated due to poor generalization. The precision is greatly improved by detrending the data, also greatly increasing the F_1 scores. However, since no time-related information is used, time series related feature extraction and selection procedures further improve the results. The highest results for recall and F_1 are attained due to an abstraction in the feature space to the time series scope, which greatly reduces the number of FNs at the cost of a slight increase in the number of FPs. Such improvements can be further examined and confirmed by the high optimistic and early warning scores.

In conclusion, the feature engineering stage results in 33 features, which are used in the remainder of the experiments. Additionally, the importance of each signal and feature combination is depicted in Fig. 4, as computed by the RF. For this specific problem, the selection procedure indicates that pH, ClO2, and Redox potential values are the most influential signals, while WT and EC are the less important ones. There is indeed some physical significance to these findings. The problem of anomaly detection in this study aims to detect contaminations in the drinking water, which are usually caused by unwanted particles polluting the water. Since the water volume is enormous compared to these pollutions, other sensors will show changes while there are no visible variations in WT (at the available sensor resolution). Moreover, the temperature is a very slow-moving parameter. Similarly, EC presents smaller changes when compared to pH value and Redox potential due to the high volume of water in comparison to the pollutions, which can be efficiently measured by the two latter signals. Such findings do not mean that there is no possible physical link between contaminations and WT/EC, but rather that these signals are hard to recognize with the current sensors and the involved water volume. It is also possible to notice how the total amplitude ($x_f^{max}(t) - x_f^{min}(t)$), standard deviation ($x_f^{\sigma}(t)$), difference between current and mean values ($x_f(t) - x_f(t)$), and difference between last and first values ($x_f(t) - x_f(t-l+1)$) are the most important features.

3.2. Model comparison

To compare the classification approaches, an ensemble built with RUSBoost is considered. Therefore, NLCA is compared to state-of-the-

Table 2
Cross-validation scores for different data conditions.

Data	F1			Recall			Precision	
	Classic	Optimistic	Early warning	Classic	Optimistic	Early warning	Classic	Range
Raw	0.1353	0.2849	0.2584	0.6072	0.7647	0.4934	0.0761	0.1750
Detrended	0.5983	0.4392	0.3755	0.5492	0.4706	0.3451	0.6570	0.4117
New Features	0.5124	0.3158	0.2779	0.5585	0.6471	0.4153	0.4732	0.2088
Selected Features	0.6085	0.5032	0.4553	0.7132	0.8431	0.6235	0.5306	0.3586

Table 3
Test scores for different ensemble models.

Data	F1			Recall			Precision	
	Classic	Optimistic	Early warning	Classic	Optimistic	Early warning	Classic	Range
Oracle	0.9469	0.5194	0.5193	0.9987	1.0000	0.9996	0.9002	0.3508
RUSBoost	0.5814	0.4755	0.4229	0.6715	0.8043	0.5663	0.5126	0.3375
OLA_{acc}	0.1824	0.0954	0.0917	0.4560	0.8261	0.4837	0.1140	0.0506
OLA_{F1}	0.0811	0.1098	0.1072	0.5620	0.8913	0.6336	0.0437	0.0585
OLCA	0.2652	0.1586	0.1473	0.4182	0.8478	0.4656	0.1941	0.0875
DES-P_{acc}	0.4618	0.5146	0.4150	0.4234	0.8478	0.4733	0.5080	0.3694
DES-P_{F1}	0.0931	0.0694	0.0673	0.4152	0.8913	0.4906	0.0524	0.0361
META-DES	0.4407	0.6572	0.4275	0.3405	0.7391	0.3346	0.6244	0.5916
$NLCA_w$	0.4905	0.5105	0.4091	0.4873	0.8261	0.4585	0.4937	0.3693
$NLCA_f$	0.4752	0.6344	0.4894	0.4392	0.8478	0.4731	0.5175	0.5068

art DES and DCS techniques, as well as an ensemble without dynamic selection. Two variations of NLCA are employed, one with the majority voting scheme ($NLCA_f$) and another with the weighted voting scheme ($NLCA_w$).

The compared DCS techniques are OLCA (Ribeiro et al., 2020), OLA_{acc} (Woods et al., 1997), and OLA_{F1}. Different from the original OLAacc, which considers the global accuracy as selection criterion, OLA_{F1} considers the F_1 score. Finally, the compared DES techniques are META-DES (Cruz et al., 2015), DES-P_{acc} (Woloszynski et al., 2012), and DES-P_{F1}. Similar to OLA_{acc} and OLA_{F1}, DES-P_{acc} considers the global accuracy as selection criterion while DES-P_{F1} considers the F_1 score.

The first step to train the base models and perform the dynamic selection is to split the training data. To this end, 70% of the initial training data is selected to train the base models while the latest 30% performs the dynamic selection. Such values fall within common split values found in literature for the validation of machine learning models (Hastie et al., 2009; Kohavi et al., 1995). During the testing phase, the models are evaluated using a separate test set with 139,566 new observations (Rehbach et al., 2018).

Table 3 brings the results for all the models. In it, the Oracle appears in the first row, which is an abstract model that always selects the correct classifier for a given sample, if such a classifier exists (Cruz et al., 2018). Next, the RUSBoost method without dynamic selection achieves the highest classic F_1 and recall values. Subsequently, the DCS models are shown. Of such methods, OLA_{F1} achieves the best results for the optimistic and early warning recalls. Despite this, all DCS methods achieve low F_1 and precision scores due to a high number of FPs. Finally, the DES models are detailed. Of such models, META-DES attains the best precision scores and best optimistic F_1 value, while the fixed weight version of NLCA obtains the highest early warning F_1score. Therefore, there is a dominance of $NLCA_f$ and META-DES over other models in terms of the early warning and optimistic F_1 scores.

The presented results indicate how dynamic selection mechanisms improve the results of the ensemble learning approach considering a time series perspective. Despite the best results for the classic configuration of F_1 and recall, RUSBoost can be further improved when considering the range-based scores. Unfortunately, DCS methods achieve the lowest scores for the given problem, since they cause a high number of FPs. Therefore, DES mechanisms are preferred, especially META-DES and NLCA. On the one hand, META-DES presents the best precision scores, but lowers the recall. On the other hand, $NLCA_f$ is able to slightly improve the precision without such a high drop in recall scores. With such, the former DES mechanism improves the optimistic F_1 score while the latter improves the early warning F_1 score. Thus, the preferable model will depend based on the preferred type of detection, optimistic or early warning. Nevertheless, further improvements can be analyzed with output filtering.

3.3. Output filtering

Table 4 details the results when output filtering is applied to the best methods, namely RUSBoost, META-DES, and $NLCA_f$. For each method, four configurations are tested: no filtering, two-minutes filtering, three-minutes filtering, and five-minutes filtering. In such a scheme, the best classic F_1 and recall scores, as well as the early warning recall, are attained by the unfiltered RUSBoost. Moreover, the best precision scores are attained by META-DES with a five-minutes filter. Finally, the best optimistic F_1, optimistic recall, and early warning F_1, are attained by $NLCA_f$ with five-minutes, three-minutes, and no filtering, respectively.

For all the compared models, the two-minute filter highly decreases the F_1 and recall scores, while only slightly modifying the precision. This is caused by the fact, that two sequential predictions must exist to generate a filtered prediction. Therefore, the number of positive predictions is decreased, lowering the number of both TPs and FPs, and consequently the recall. Thus, such a filtering time window is not recommended for any model.

By using the three-minutes and five-minutes filtering configurations, the optimistic F_1 and range-based precision scores are improved

Table 4
Test scores for different filtering times for the three best models.

Data	F1			Recall			Precision	
	Classic	Optimistic	Early warning	Classic	Optimistic	Early warning	Classic	Range
Oracle	0.9469	0.5194	0.5193	0.9987	1.0000	0.9996	0.9002	0.3508
RUSBoost	0.5814	0.4755	0.4229	0.6715	0.8043	0.5663	0.5126	0.3375
2 min	0.5709	0.4012	0.3477	0.6136	0.6957	0.4534	0.5338	0.2819
3 min	0.5779	0.5797	0.4848	0.6595	0.8043	0.5213	0.5142	0.4531
5 min	0.5725	0.6024	0.4805	0.6518	0.8043	0.4796	0.5104	0.4815
META-DES	0.4407	0.6572	0.4275	0.3405	0.7391	0.3346	0.6244	0.5916
2 min	0.3455	0.5949	0.3001	0.2370	0.5652	0.1972	0.6374	0.6279
3 min	0.4361	0.6791	0.4179	0.3302	0.6522	0.2964	0.6419	0.7084
5 min	0.4337	0.6705	0.3947	0.3267	0.6304	0.2724	0.6449	0.7160
$NLCA_f$	0.4752	0.6344	0.4894	0.4392	0.8478	0.4731	0.5175	0.5068
2 min	0.4276	0.4976	0.3769	0.3658	0.6304	0.3480	0.5145	0.4110
3 min	0.4635	0.6640	0.4939	0.4255	0.7609	0.4252	0.5090	0.5891
5 min	0.4516	0.6969	0.4806	0.4126	0.7609	0.3837	0.4987	0.6428

for all models. However, all recall scores drop when filtering in such configurations. For the early warning F_1 score, META-DES presents a slight performance loss, RUSBoost achieves better results, and $NLCA_f$ has little changes. Nevertheless, the final range-based F_1 performance for $NLCA_f$ confirms the better trade- off between precision and recall for the multi-criteria based dynamic selection mechanisms, which also enables a more suitable trade-off between existence and early warning F_1 score configurations for time series classification.

The two best models, $NLCA_f$ with three and five-minutes filters, dominate all other solutions in the range-based F_1 score context. That is, such models present the best trade-off between the number of FN and FP, as parts of recall and precision, respectively. Therefore, one such model can be selected to implement the anomaly detection system for drinking water quality monitoring. The final selection must take into account that, (1) both models can detect the same percentage of anomalies (76%); (2) a three-minute filter enables a faster anomaly detection (higher early warning recall); and (3) a five-minute filter reduces the number of FPs (higher range-based precision). Therefore, a decision maker must define what is more important for the problem at hand, a faster detection and lower risks to public health, or a lower financial burden and higher risk to public health.

3.4. Final remarks

In the context of anomaly detection for the drinking water quality problem, the three sequential experiments indicate the following: (1) feature engineering is of great importance to correctly detect anomalies and reject data under normal operation; (2) DES mechanisms are able to improve the predictive performance of strong ensemble learning methods; and (3) output filtering plays an important role in fixing noisy detections, reducing both the number of FPs and FNs. Nevertheless, results also show that NLCA outperforms all other techniques in terms of range-based predictive performance. This indicates that DES can greatly benefit from considering multiple conflicting criteria. Therefore, the proposed technique contributes to such a field by introducing the multi-criteria perspective, which can be taken into account by researchers, scientists, and engineers when developing new selection mechanisms.

It is important to notice that the proposed solution can also be implemented on other water networks, given that enough training data is available. In case different signals are employed, modifications would be necessary in the feature engineering step (Section 2.2). Most specifically, more attention should be given to signal processing techniques, since different variables could need different forms of treatment. Nevertheless, it is expected that the proposed feature extraction and selection mechanisms are agnostic to the nature of the signal and robust enough to be applied in different time series related problems.

Despite this, the use of different training data is likely to change the efficiency of machine learning models. When implementing the solution with RUSBoost and NLCA on a new water network, the number of base models, training class ratio, and number of samples in the region of competence can be tuned to achieve better predictive performance. To this end, grid search is recommended (Andonie, 2019). Moreover, the performance evaluation during hyper-parameter selection can follow the cross-validation scheme presented in Section 3.

4. Conclusions

To solve the problem of anomaly detection in drinking water quality, this work proposed the non-dominated local class-specific accuracy (NLCA), a novel multi-criteria dynamic ensemble selection (DES) mechanism for multiple classifier systems. Different from existing DES techniques, NLCA considers the relation between TPR and TNR to achieve a better trade-off between FPs and FNs in classification tasks. The method is compared to current state-of-the-art ensemble learning algorithms, such as META-DES and RUSBoost. Moreover, an output filtering process is incorporated to analyze its effects on time series classification, improving anomaly detection for the given problem. Results indicate the success of the proposed technique, which attains the best scores considering early warning and optimistic scenarios for anomaly detection. NLCA achieves optimistic and early warning F_1 scores of 0.6969 and 0.4939, respectively, outperforming both META-DES (0.6791 and 0.4275) and RUSBoost (0.6024 and 0.4848). That is, NLCA attains more than 15% improvement over META-DES and RUSBoost in terms of early warning and optimistic F_1 scores, respectively.

As a conclusion, NLCA improves the relation between FPs and FNs for the task of anomaly detection in drinking water quality monitoring, enabling the development of more reliable anomaly detection systems in water network systems. Future works shall study novel multi-criteria-based techniques for dynamic selection, such as different rules for ensemble and classifier selection using different MCDM ranking techniques. Moreover, the comparison of dynamic multi-criteria selection mechanisms with current state-of-the-art DES techniques must be performed in different types of problems, such as regression and multi-class classification. To this end, statistical comparison on multiple benchmark problems is recommended.

CRediT authorship contribution statement

Victor Henrique Alves Ribeiro The author performed the experiments and wrote the manuscript.

Steffen Moritz The author performed data collection/curation and wrote the manuscript.

Frederik Rehbach The author performed data collection/curation and wrote the manuscript.

Gilberto Reynoso-Meza The author supervised the experiments and revised the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and the Fundação Araucária (FAPPR) - Brazil - Finance Codes: 159063/2017-0-PROSUC, 310079/2019-5-PQ2, 437105/2018-0-Univ, 51432/2018-PPP, and PRONEX-042/2018.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.142368.

References

Ali, A.A., Rasheeq, A., Logofatu, D., Badica, C., 2019. Anomaly detection procedures in a real world dataset by using deep-learning approaches. Asian Conference on Intelligent Information and Database Systems. Springer, pp. 303–314.

Andonie, R., 2019. Hyperparameter optimization in learning systems. Journal of Membrane Computing 1–13.

Barbon, J., 2020. Police Investigate CEDAE Staff in Rio Water Crisis. Folha de São Paulo https://www1.folha.uol.com.br/internacional/en/brazil/2020/01/police-investigate-cedae-staff-in-rio-water-crisis.shtml.

Brans, J.P., De Smet, Y., 2016. PROMETHEE methods. Multiple Criteria Decision Analysis. Springer, pp. 187–219.

Brans, J.P., Vincke, P., Mareschal, B., 1986. How to select and how to rank projects: the PROMETHEE method. Eur. J. Oper. Res. 24, 228–238.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Britto Jr., A.S., Sabourin, R., Oliveira, L.E., 2014. Dynamic selection of classifiers - a comprehensive review. Pattern Recogn. 47, 3665–3680.

Castell-Exner, C., Zenz, T., 2010. Klimawandel und wasserversorgung. Energie, Wasser-Praxis 61, 20–23.
Cavalin, P.R., Sabourin, R., Suen, C.Y., 2013. Dynamic selection approaches for multiple classifier systems. Neural Comput. & Applic. 22, 673–688.
Chen, X., Feng, F., Wu, J., Liu, W., 2018. Anomaly detection for drinking water quality via deep biLSTM ensemble. Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 3–4.
Cruz, R.M., Sabourin, R., Cavalcanti, G.D., Ren, T.I., 2015. META-DES: a dynamic ensemble selection framework using meta-learning. Pattern Recogn. 48, 1925–1935.
Cruz, R.M., Sabourin, R., Cavalcanti, G.D., 2018. Dynamic classifier selection: recent advances and perspectives. Information Fusion 41, 195–216.
Deng, W., Wang, G., 2017. A novel water quality data analysis framework based on time-series data mining. J. Environ. Manag. 196, 365–375.
Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C., 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J. 16, 235–248.
Domingos, P., 2012. A few useful things to know about machine learning. Commun. ACM 55, 78–87.
Fehst, V., La, H.C., Nghiem, T.D., Mayer, B.E., Englert, P., Fiebig, K.H., 2018. Automatic vs. manual feature engineering for anomaly detection of drinking-water quality. Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 5–6.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139.
Giacinto, G., Roli, F., 1999. Methods for dynamic classifier selection. Proceedings 10th International Conference on Image Analysis and Processing. IEEE, pp. 659–664.
Hansen, J., Sato, M., Ruedy, R., 2012. Perception of climate change. Proc. Natl. Acad. Sci. 109, E2415–E2423.
Hart, D., McKenna, S.A., Klise, K., Cruz, V., Wilson, M., 2007. CANARY: a water quality event detection algorithm development tool. World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat, pp. 1–9.
Hasan, M.A.M., Nasser, M., Ahmad, S., Molla, K.I., 2016. Feature selection for intrusion detection using random forest. J. Inf. Secur. 7, 129–140.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.
Hernandez-Ramirez, A., Martinez-Tavera, E., Rodriguez-Espinosa, P., Mendoza-Pérez, J., Tabla-Hernandez, J., Escobedo-Urı´as, D., Jonathan, M., Sujitha, S., 2019. Detection, provenance and associated environmental risks of water quality pollutants during anomaly events in river Atoyac, Central Mexico: a real-time monitoring approach. Sci. Total Environ. 669, 1019–1032.
Kohavi, R., et al., 1995. A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI, Montreal, Canada, pp. 1137–1145.
Korth, A., Petzold, H., Böckle, K., Hambsch, B., 2007. Coliforme umweltkeime in trinkwasserverteilungssystemen - vorkommen, anreicherung und vermehrung. Abschlussbericht DVGW-Forschungsvorhaben W 6/03/04.
Kotsiantis, S.B., 2013. Decision trees: a recent overview. Artif. Intell. Rev. 39, 261–283.
Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence 5, 221–232.
Leigh, C., Alsibai, O., Hyndman, R.J., Kandanaarachchi, S., King, O.C., McGree, J.M., Neelamraju, C., Strauss, J., Talagala, P.D., Turner, R.D., et al., 2019. A framework for automated anomaly detection in high frequency water-quality data from in situ sensors. Sci. Total Environ. 664, 885–898.
Liu, J., Wang, P., Jiang, D., Nan, J., Zhu, W., 2020. An integrated data-driven framework for surface water quality anomaly detection and early warning. J. Clean. Prod. 251, 119145.
Moritz, S., Rehbach, F., Chandrasekaran, S., Rebolledo, M., Bartz-Beielstein, T., 2018. GECCO Industrial Challenge 2018 Dataset: A Water Quality Dataset for the 'Internet of Things: Online Anomaly Detection for Drinking Water Quality' competition at the Genetic and Evolutionary Computation Conference 2018, Kyoto, Japan. https://doi.org/10.5281/zenodo.3884398.
Muharemi, F., Logofatu, D., Leon, F., 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of Information and Telecommunication 3, 294–307.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT Press.
Phillips, D., 2020. "It Tastes Like Clay": Residents of Rio Alarmed by Murky, Smelly Tap Water. The Guardian https://www.theguardian.com/world/2020/jan/16/brazil-rio-de-janeiro-tap-water-pollution.
Piramuthu, S., 2008. Input data for decision trees. Expert Syst. Appl. 34, 1220–1226.
Rehbach, F., Chandrasekaran, S., Rebolledo, M., Moritz, S., Bartz-Beielstein, T., 2018. GECCO Challenge 2018: Online Anomaly Detection for Drinking Water Quality. URL. http://www.spotseven.de/gecco/gecco-challenge/gecco-challenge-2018/.
Ribeiro, V.H.A., Reynoso-Meza, G., 2018. Online anomaly detection for drinking water quality using a multi-objective machine learning approach, in: proceedings of the genetic and evolutionary computation conference companion. ACM. pp. 1–2.
Ribeiro, V.H.A., Reynoso-Meza, G., 2020. Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. Expert Syst. Appl. 147, 113232.
Ribeiro, V.H.A., Domingues, P.H., Cavalin, P.R., Reynoso-Meza, G., Ayala, H.V.H., Azevedo, L.F.A., 2020. Dynamic multi-criteria classifier selection for illegal tapping detection in oil pipelines. 2020 International Joint Conference on Neural Networks (IJCNN). IEEE (in press).
Roy, B., 2016. Paradigms and challenges. Multiple Criteria Decision Analysis. Springer, pp. 19–39.
Sagi, O., Rokach, L., 2018. Ensemble learning: a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8, e1249.
Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A., 2009. RUSBoost: a hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40, 185–197.
Shi, B., Wang, P., Jiang, J., Liu, R., 2018. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. Sci. Total Environ. 610, 1390–1399.
Souza, F.A., Araújo, R., Mendes, J., 2016. Review of soft sensor methods for regression applications. Chemom. Intell. Lab. Syst. 152, 69–79.
Tatbul, N., Lee, T.J., Zdonik, S., Alam, M., Gottschlich, J., 2018. Precision and recall for time series. Advances in Neural Information Processing Systems, pp. 1920–1930.
Tomaschek, F., Hendrix, P., Baayen, R.H., 2018. Strategies for addressing collinearity in multivariate linguistic data. J. Phon. 71, 249–267.
UN General Assembly, 2010. The human right to water and sanitation. UN Resolution 64, 292.
Woloszynski, T., Kurzynski, M., Podsiadlo, P., Stachowiak, G.W., 2012. A measure of competence based on random classification for dynamic ensemble selection. Information Fusion 13, 207–213.
Woods, K., Kegelmeyer, W.P., Bowyer, K., 1997. Combination of multiple classifiers using local accuracy estimates. IEEE Trans. Pattern Anal. Mach. Intell. 19, 405–410.
Zheng, A., Casari, A., 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc.