2024 IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR) | 979-8-3503-6193-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/AQTR61889.2024.10554265

Anomaly Detection in Water Consumption Patterns Using Prediction and Clustering Approaches

Codruta Maria SERBAN, Gheorghe SEBESTYEN, Anca HANGAN
Department of Computer Science, Technical University of Cluj-Napoca, Romania
email: Codruta.Serban@cs.utcluj.ro, Gheorghe.Sebestyen@cs.utcluj.ro, Anca.Hangan@cs.utcluj.ro

***Abstract*—Abnormal consumption patterns or potential leakages in the water distribution system can be identified at an early stage with the help of anomaly detection. This can prevent water wastage, encourages responsible usage and minimize environmental impact. Labeled data indicating the presence of an anomaly has a considerable impact in the accuracy of the anomaly detection model, but also for its validation.**

In this direction we propose two methods for anomaly detection in daily water consumption. The first approach revolves around predicting future water usage and comparing it with real-time values, enabling the identification of deviations from expected consumption. The other one is focused on classifying consumption behavior. With the help of a pre-trained classifier, the new measurements are labeled as anomalies or regular patterns.

***Index Terms*—water consumption, anomaly detection, prediction, classification**

I. Introduction

Along with the population growth and the industrial sector development, the demand for water in urban areas has increased considerably. Therefore, improving the existing administration systems has become a must.

In the context of water consumption, an anomaly means negative or excessively high value resulted from erroneous meter readings, or an unusual consumption pattern, i.e. constant consumption higher than zero day and night or average consumption considerably lower or higher than the previous average values. These changes can be caused by faulty sensors, broken pipes, faucets left open, an increase/decrease in the number of inhabitants of the house, if they went on vacation, if they are spending more time at home and so on.

This kind of information can be integrated in various systems, offering utility from both administrative and consumer perspectives. It is practical for the local water utilities administration as it helps planning the optimal distribution of water. Regarding the customers, insights into their water consumption can be valuable, for example to know their average usage, or to be notified about a possible problem as an unusual consumption occurred.

Anomaly detection techniques may be built upon unsupervised, supervised or semi-supervised learning methods, most of them being from the first category. Both, labeled and unlabeled data can be used for the training data set. Our approach focuses on techniques that don't require labels, as this use case is more common and many existing data sets are not labeled.

The current paper targets two types of anomalies: caused by measurement errors and by changes in consumption behavior. The centerpiece of our work consists in two solutions that can be integrated into a domestic water consumption monitoring system. As they focus on the second category of anomalies, the data set employed for analysis should be free from measurement inaccuracies. Both can identify anomalies in daily water consumption. The scenario in which these approaches can fit in is that after identifying the abnormal consumption pattern, the user is notified of the presence of possible anomalies, and it remains user's responsibility to identify their cause. Behind a fluctuation in consumption there may be, either a malfunction that leads to water loss, or a conscious action carried out by the occupants of the home.

The first approach is based on the predicted consumption for the next day. By comparing the actual and predicted water consumption the method can identify potential anomalies. Techniques used in this case are autoregressive integrated moving average (ARIMA) and adaptive threshold computation.

The second implemented method starts from grouping data into clusters, which were consider to be daily consumption behaviors. New measurements that deviate from the established patterns are classified as anomalies. As a first step, the initial data set was labeled based on a set of rules. Then, a classification method is used to determine whether an entry is part of normal consumption behavior or not.

The DAIAD data set [1] served as the initial data source. In the pre-processing step the erroneous data had to be eliminated. After having a clean data set, the developed methods were applied which successfully identified their targeted anomalies. There were cases in which both methods identified the same measurement as an anomaly.

The rest of the paper is organized as follows. Section II presents related work. Section III presents the problem statement. Section IV focuses on the anomalies caused by measurement errors, identified types in the DAIAD data set and how they have been fixed. Section V provides a comprehensive description of the two studied detection methods: prediction and evaluation, as well as clustering and classification. Section VI contains the results of each step taken to implement the mentioned methods and their limitations. Section VII concludes the paper.

II. Related work

In water consumption monitoring systems, there are several types of anomalies that can appear: negative, duplicate or missing values due to malfunctioning sensors, sudden changes in the consumption pattern or values that hardly fit in the dominant consumption behaviors [2] [5] [9].

The related work presented in this paper is focused around three aspects: anomaly detection, forecasting and detection of behavior change. A way to detect anomalies like the ones mentioned above is to have a forecasting model that based on former values would predict the future expected consumption. If it differs considerably from the real value read by the meter, an anomaly occurred. Another approach would be to identify behavior patterns in the water consumption of a home and if a new record sequence doesn't fit in any of them, it is considered an anomaly.

A. Anomaly detection

The authors in [2] introduce a new constrained-clustering-based semi-supervised approach for detecting periods of abnormal water consumption in supermarkets, called SSDO (Semi-Supervised Detection of Outliers). The method starts with an unlabeled data set and uses clustering to make the difference between normal and abnormal values. Based on this, it assigns an anomaly score to each recording. The score is calculated using the point deviation, cluster deviation, cluster size and a squashing function. A domain expert defines ground-truth labels that improve the model through active learning. Every time new labels are provided the label propagation step is performed in order to maximize the usefulness of the new acquired information. During this phase, the anomaly score of each example is adjusted based on the known labels of nearby examples. Consequently, even in the absence of initial labels, the method progressively refines its performance with feedback, becoming more adept at discerning more subtle differences in behavior. Vercruyssen et al. [3] also addresses semi-supervised anomaly detection in water consumption but it uses multi-domain active learning. The active learning bandits (Alba) method creates independent classifiers for each domain. Both, SSDO and Alba approaches have demonstrated that supplying the model with limited yet specific labels can significantly enhance its performance.

The technique described in [4] uses an anomaly score, calculated based on ten features which describe the mean daily flow, the lowest and highest flow in the last 24 hours and the mean water flow distributed in seven intervals throughout a day. The anomaly score for each day is evaluated based on its calendar context. This context, defined for a day d, contains a subset of days from the database D that are anticipated to exhibit similar water usage patterns as day d. In addition to detecting the abnormal consumption, the proposed solution finds possible causes for it in order to identify the physical event beside the problem.

In [5] the algorithm uses rules, historical context, and user location to detect leakage in the water distribution system of a household. The algorithm is executed at the end of a specific time interval, and based on the inputs, it checks if the water consumption matches any of the four leak scenarios: “negative trend”, “24-hour consumption”, “similar consumptions”, and ”anomalous high consumptions”. They are run in this order, and if any is verified then there may be a leak. Ten possible water consumption scenarios between normal and anomalous consumption are tested and the results show that the algorithm has 100% accuracy in leakage detection.

B. Forecasting

Boudhaouia and Wira [6] developed a real-time data analysis platform to forecast water consumption using Long Short-Term Memory (LSTM) and the Back-Propagation Neural Network (BPNN) algorithms. The measurements from distributed smart meters are stored as unevenly spaced time series. Using the previous mentioned machine learning techniques and only past water consumption (without contextual information) the platform is able to predict the quantity of water consumed in the next coming hours.

A hybrid model is presented in [7]. The stochastic portion of the original time series is extracted through a signal pre-treatment method and then an autoregressive (AR) model is applied in order to predict monthly water consumption for the next year. Ten years of historical water consumption data were used. The signal pre-treatment method consists in decomposing the raw time series into a number of sub time series. The results show that applying the signal pre-treatment method improved the capability of the forecasting model to predict monthly water consumption.

The authors of [8] compared multiple new hybrid models which uses ARIMA with seasonality (SARIMA), artificial neural network models (like Long Short Term Memory, LSTM, or the Multi Layer Perceptron, MLP) and a deterministic model based on a time function. The purpose is to predict hourly water consumption values of a building. In total, they test 7 approaches, the individual ones and other four combinations of them. The hybrid model that combines the deterministic model with ANN (LSTM for the week days and MLP for the weekend days) and SARIMA offers the lowest forecast error.

C. Behavior change detection

The coronavirus pandemic has become an exciting topic of research even from a technological point of view. We noticed in the literature some works that highlight the behavior change in water consumption before and after the pandemic.

Ludtke et al. [9] conducted a study which examines hourly and daily water consumption patterns of a utility in northern Germany during the initial phase of the pandemic. The consumption comparison is made using a linear mixed model. They used Bayesian statistics to remove climate-related influences, allowing them to isolate and estimate the specific impact of COVID-19. Their findings demonstrate a significant increase of approximately 14.3% in daily residential water consumption. This rise is marked by elevated demand peaks in both the morning and evening periods due to changed behavioral routines.

The methodology presented in [10] analyses the weekly water consumption data, at an hourly resolution. It identifies four clusters of households with distinctive temporal patterns by using the KMeans algorithm for pattern recognition. Labeling the data makes possible the pattern classification in order to predict the cluster for new input data. For this step two unsupervised algorithms are tested: K Nearest Neighbour(KNN) and Logistic Regression (LR). The last one shows best results with an accuracy between 93% and 94%.

Based on the research, it was decided to implement two methods that are able to identify anomalies caused by changes in the consumption behavior. The first approach uses forecasting to predict the future consumption. If the real value differs drastically, a sudden trend change is detected. The second approach classifies measurements as anomalies or not following a consumption patterns analysis. The purpose is to detect those values that hardly fit or do not fit at all in the most common consumption behaviors in the data set.

III. Problem statement

The problem this work wants to solve is the detection of water consumption anomalies. Anomalies may result from measurement errors or changes in consumption behavior. The latter being characterized by either a sudden shift in the user's consumption tendency or a deviation from the dominant consumption behaviors. It is considered that a smart water meter (SWM) is installed on the water supply pipe of each house or apartment. The device measures the amount of water consumed in the home at a well-established time interval (per hour). Each record is sent to an analyzer module. This module was developed on a NVIDIA Jetson Nano device; its job is to determine in real time whether there is an anomaly in the incoming data. The result obtained is then sent to the Cloud where there is a service that notifies the home owner about the existence of a possible problem. Afterwards, it is up to the user to find and solve the problem that produced that unusual consumption. Fig. 1 illustrates the main scenario, where anomaly detection plays a very important role. The current work focuses only on this step.

The anomalies caused by measurement errors may be due to a faulty SWM. They are handled first so that the two developed methods can be applied on a clean data set.

The current paper addresses two anomaly detection methodologies related to consumption behavior. The first one is based on predicting user consumption based on existing data. The approximate value is compared with the true consumption when it is recorded. Based on the differences between them it is determined if the amount of water consumed has an abnormal value.

The second methodology is based on consumption patterns. A user's data from the data set is clustered using the KMeans algorithm. In the labeled results, anomalies are determined using various criteria. With the new labeled data set, it is determined whether a new consumption reading is an anomaly or not using classification models.

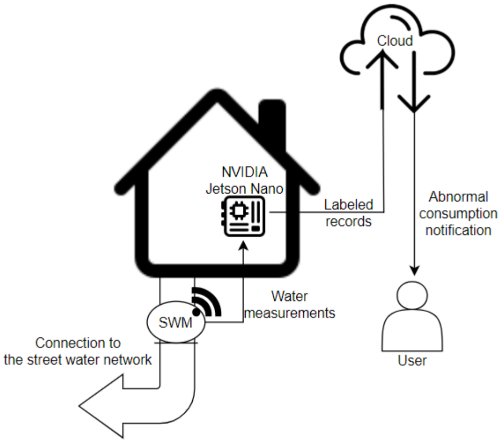

Fig. 1: Water consumption anomaly detection using an NVIDIA JetsonNano device

IV. Anomalies caused by measurement errors

The data used for anomaly detection is part of the DAIAD project [1]. Water consumption comes from 1,000 users, and the longest time period monitored is 01.01.2015-19.05.2017. Each instance indicates the amount of water used within an hour and is described by four values: a unique anonymous identifier of a home, the time when the SWM reads the consumption, the total amount of water used since the installation of the meter represented in liters, the amount of water consumed since the previous reading.

Three types of measurement errors were observed and fixed in the DAIAD data set:

- Negative values
- Duplicated recordings
- Missing values

These errors are considered a source of disturbance because they are not usual values that would characterise water consumption. The erroneous data were inserted by the measuring device and it was necessary to fix them in order to start with a clean data set. Their presence makes it impossible to determine the difference between an expected or an anomalous behavior. Fig. 2 illustrates the steps to eliminate the three types of errors encountered in order to obtain a clean data set, ready for anomaly detection.

The first two categories are correlated because duplicated values always appear in sequences with negative values for hourly consumption. Many cases were noticed in which two SWM values are recorded in one hour, and the second one is lower than the first one, resulting in a negative difference. This inconsistency can significantly harm the results of experiments because it is impossible for water consumption, over a time interval, to have negative value.

The solution for these two errors was to replace the values. Initially, the records that delimit the problematic sequence are identified, keeping in mind that water consumption increases with time. The last correct record before the first negative

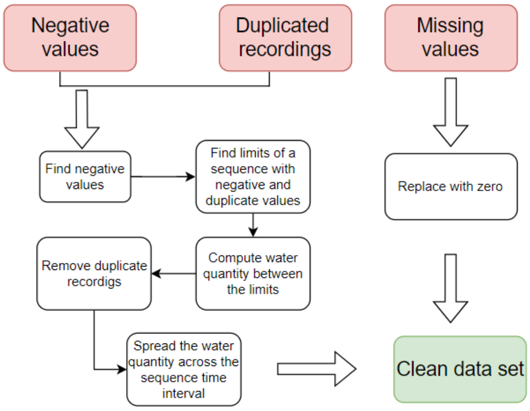
Negative values
Duplicated recordings
Missing values
Find negative values
Find limits of a sequence with negative and duplicate values
Compute water quantity between the limits
Remove duplicate recordigs
Spread the water quantity across the sequence time interval
Replace with zero
Clean data set

Fig. 2: The three measurement errors identified in DAIAD and their associated steps to be removed

value in the erroneous sequence and the first correct value after this sequence are spotted. Then the consumption between these moments was calculated. Where applicable, multiple entries from the same hour are removed so that there is only one reading for a given hour. The next step is to spread the previously obtained difference across the period specified by the initially set limits, including the upper limit. The measurements for the consumption columns are updated. In this way, the same amount of water is distributed over the same time interval, but without negative values.

A final issue encountered with this data set was the lack of measurements for several hours. The solution to this problem is to add values for these hours equal to zero for the hourly consumption. Missing records affect the time series by breaking its continuity. Adding zero consumption measurements values will not conflict with the known total amount of used water in each hour.

It should be noted that the anomaly detection methods presented in this paper use the records corresponding to a single home. Before performing any experiment, the measurements for several individual buildings were extracted from the DAIAD data set, and then the presented rules were applied to each one.

V. Proposed Techniques for Anomaly Detection in Water Consumption Behavior

A. Prediction and evaluation

With this method, the classification of measurements as anomalies or not is done by comparing a predicted value with the true consumption recorded by the SWM (Fig. 3). The algorithm used for anticipating future consumption is the ARIMA model.

After eliminating measurement errors, a new prepossessing step was applied. Its purpose is to reduce the sampling rate, from hourly sampled data towards daily data. This way, long sequences of zero consumption values were eliminated. These zero sequences would have a negative effect on the prediction algorithm.

The ARIMA model is used to predict future daily values based on historical measurements. Once the model is created (with 1, 1, 0 values for p, d and q) it can be used to approximate future consumption. These are compared with the actual values recorded by the water meter and if the difference exceeds a certain threshold it means that an anomaly is present. In order to determine the threshold, tests were performed using various methods, such as: the mean of the differences, standard deviation, cumulative sums or other variants that use these values.

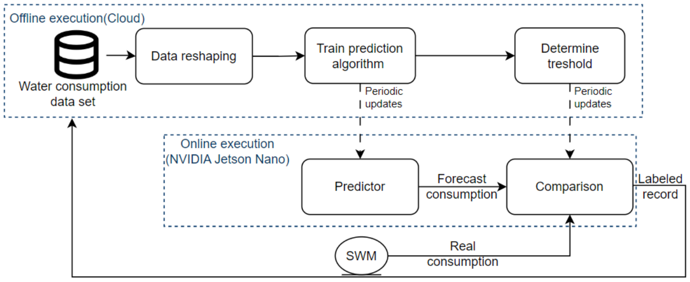

Fig. 3: Anomaly detection model based on prediction and evaluation of recorded consumption

Algorithm 1 Algorithm to determine the threshold for 'Prediction and evaluation' method

```
procedure FIND_THRESHOLD(predictions, real_values)
    $differences \leftarrow predictions - real\_values$
    $threshold \leftarrow 0$
    $sum\_aux \leftarrow 0$
    for $diff$ in $differences['water']$ do
        $sum\_aux \leftarrow sum\_aux + diff - threshold$
        if $sum\_aux < 0$ then
            $sum\_aux \leftarrow 0$
        if $sum\_aux > threshold$ then
            $threshold \leftarrow sum\_aux$
    return $threshold$
```

The developed method, as shown in Algorithm 1, starts with calculating the differences between the predicted and the actual values. Iterating through the obtained array, it calculates the cumulative sum of the subtraction between the current element and the threshold and in the end it returns the maximum value of the threshold encountered throughout the iteration. Basically, it accumulates positive differences, gradually increasing the threshold, and resets the accumulation to zero whenever negative differences occur. The final threshold represents a significant positive deviation from the actual values, capturing the fluctuations in the data.

The prediction model is trained offline and then uploaded on the NVIDIA board using serialization. The method for finding the threshold is periodically run (for example, monthly). This is necessary in case the residents changed their usual consumption behavior. At the end of the day, when all hourly consummations are received, the algorithm can be applied to label the daily usage of water from that day.

B. Clustering and classification

As Fig. 4 indicates, this method of categorizing measurements as anomalies or not is based on a classification

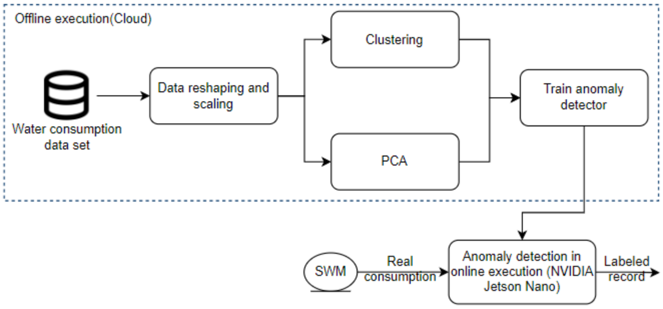
Offline execution(Cloud)
Water consumption data set
Data reshaping and scaling
Clustering
PCA
Train anomaly detector
SWM
Real consumption
Anomaly detection in online execution (NVIDIA Jetson Nano)
Labeled record

Fig. 4: Anomaly detection model based on clustering and classification

algorithm. Records from the DAIAD data set, after removing the measurements error, are preprocessed and labeled with the cluster value that characterizes them. For measurements that meet certain conditions, the cluster value is set to indicate an anomaly. Multiple classification models are trained with this data, and the most accurate one is uploaded to the NVIDIA Jetson Nano device. When it receives a new record it is classified, tagged with the result obtained and finally transmitted to the Cloud.

For this approach the data set was reshaped so that each record represents the consumption for one day distributed over six time intervals (an interval is a column in the data set): night, morning, late morning, afternoon, evening, late evening. The purpose of data reshaping is to reduce the number of zero values. Their presence is justified by the fact that during the night the residents do not use the water because they are sleeping, and the same happens when they are away from home. Standard scaling was then applied to this new data set.

The KMeans algorithm is used to identify behavior patterns based on water consumption. The records were divided into six groups. this number was obtained using the Elbow method.

In the data set, the daily consumption is described by six columns, one for each of the previously mentioned time intervals. In consequence, it is not possible to visualize the water usage in a more human friendly format, like in a Cartesian plane. The solution to this problem was to use principal component analysis (PCA) which reduces the dimensionality of the data. Now, the water consumption is indicated by only two features and can be visualized as points in a two dimensional plan.

Having the coordinates of each input and the label of the pattern they belong to, a new data set is obtained and it can be used to identify the anomalies. None of the found clusters represent a grouping of only the anomalous data. Anomalies can be met among the data of all clusters. A record is anomalous if:

- it's part of a cluster whose number of elements is less than 1% of the total number of values
- the integer part of the Euclidean distance from the point to the centroid of the cluster it belongs to, is greater than a threshold. This limit reflects the proximity to the centroid.

Fig. 5 contains the histogram of the distribution of points around the centroid for a selected cluster. In this cluster, the anomalies are the points whose distance from the centroid exceeds the value 4. The depth of the last two buckets is considerably smaller than the other ones. It is obvious that most of the points in the cluster are near the centroid, meaning that they are similar, describing the same consumption behavior.

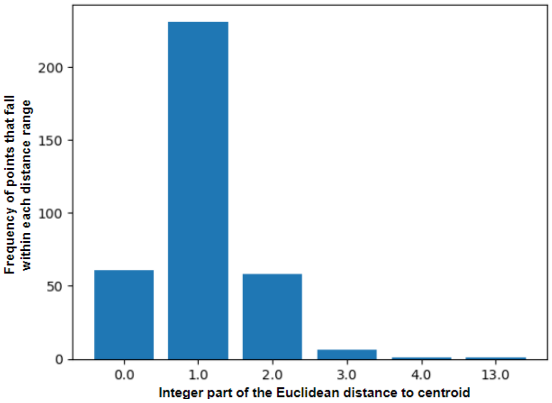

Fig. 5: Histogram of the distribution of points around the centroid for a seleted cluster

Based on these rules, the labels obtained by KMeans are updated and this data set is used to choose the classifier.

The data is split in training and testing, each containing anomalies. Anomaly detection was tested using four classifiers: K-Nearest Neighbor, Decision Tree, Random Forest and Isolation Forest. The evaluation metrics used are precision, recall and accuracy. Of these, the most important is the last one because it represents the number of correctly identified anomalies out of the total number of anomalies in test data.

Similar to the first methodology, this one can be used for the scenario where a pre-trained classifier is loaded on the NVIDIA Jetson Nano device and at the end of the day it decide whether that day's water consumption was abnormal or not.

VI. Results

A. Prediction and evaluation

Fig. 6 shows the anomalies detected by the prediction and evaluation based method. Those points are determined by a sudden and significant increase or decrease in water consumption and are successfully marked.

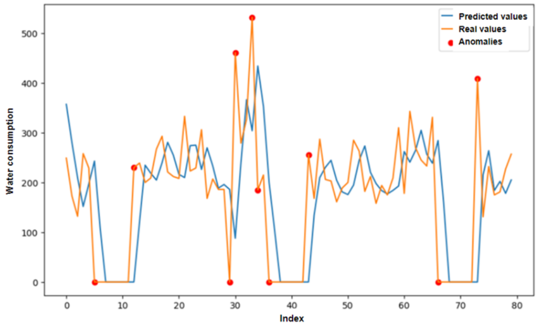

Fig. 6: Anomalies detected using prediction and evaluation based method

This solution is only validated by graph analysis because the data in DAIAD is not labeled to know for sure which records are anomalies. But it manages to satisfy what the project set out to do because it identifies significant trend changes.

B. Clustering and classification

In Fig. 7, the daily consumption for a user is represented by points. The labels are obtained using the KMeans algorithm, with the exception of the -1 label. This indicates abnormal values respecting the rules mentioned in the previous section. In some groups, there are points more spread out towards the extremities. Among those, the most distant ones may be anomalies because they differ from the majority of points in the cluster that are concentrated in certain areas.

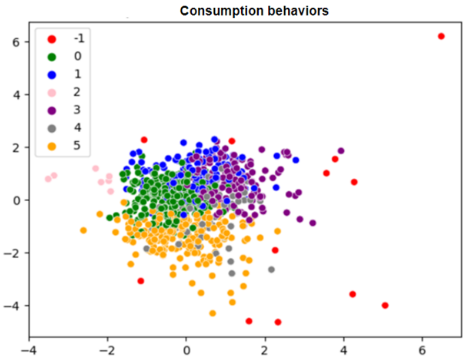

Fig. 7: Anomalies (red dots) detected using clustering and classification based method

Each of the models were trained and tested ten times. From our experiments, it was determined that Isolation Forest achieves the best results, detecting a larger number of anomalies compared to the other classifiers, being cases where its classification completely achieves the expected results. The other three models obtain the same results in most cases. This is caused by the low number of anomalies in the training data set.

C. Discussion

Each method has its own advantages and disadvantages. Regarding the number of historical values, ARIMA is able to make predictions having a limited number of past values. On the other hand the classifier needs a large data set of labeled historical values for training in order to make correct decisions.

For the first approach the model's input data must be univariate, it should describe the values of a single variable, in this case water consumption (the data set contains a single column). The more information the input provides, the more the algorithm learns and the more accurate the predictions are. A classifier can support as many features as are provided to it. To take advantage of this, new useful information was generated by grouping the consumption into six time intervals.

The results of the classifier are affected by the lack of labels in the initial data set because if our labeling is wrong then the classifier is automatically wrong. The data set used by the classifier contains less than 1% abnormal records. If the training and test data sets contains few anomalies, the examples the model can learn from are reduced and so are the cases for result evaluation.

Putting the results from both methods side by side, common points labeled as anomalies can be observed. It confirms they are able to identify very unusual consumption values. Even so, the double check from a human agent is still needed, as the models cannot be validated due to the lack of a labeled data set.

VII. Conclusions

In conclusion, the proposed methods manage to identify anomalies in the DAIAD data set, both sudden changes in trend, but also those records that hardly fit into the dominant consumption behaviors. Cases were observed where the same consumption was detected as abnormal by both approaches. An advantage of the developed solutions is that pre-trained models can be easily transferred to the edge device using serialization and in addition, do not require large amounts of memory. This paper's contribution also includes the methodology used for mitigating measurement errors, augmented by algorithms for computing the thresholds applied in labeling the data set. While they cannot be evaluated due to insufficient information, these methods demonstrate an ability to detect consumption changes that can later be verified by the user helping him to reduce water waste where appropriate.

References

[1] Athanasiou, Spiros, et al. "DAIAD: Open water monitoring." Procedia Engineering 89 (2014): 1044-1049.

[2] Vercruyssen, V.; Meert,W.; Verbruggen, G.; Maes, K.; Baumer, R.; Davis, J. Semi-supervised anomaly detection with an application to water analytics. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; Volume 2018, pp. 527–536.

[3] Vercruyssen, V., Perini, L., Meert, W., & Davis, J. (2022, September). Multi-domain Active Learning for Semi-supervised Anomaly Detection. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 485-501). Cham: Springer Nature Switzerland.

[4] Patabendige, S.; Cardell-Oliver, R.;Wang, R.; Liu,W. Detection and interpretation of anomalous water use for non-residential customers. Environ. Model. Softw. 2018, 100, 291–301.

[5] Fuentes, H.; Mauricio, D. Smart water consumption measurement system for houses using IoT and cloud computing. Environ. Monit. Assess. 2020, 192, 602.

[6] Boudhaouia, Aida, and Patrice Wira. "A real-time data analysis platform for short-term water consumption forecasting with machine learning." Forecasting 3.4 (2021): 682-694.

[7] Zubaidi, S. L., Al-Bugharbee, H., Muhsin, Y. R., Hashim, K., & Alkhaddar, R. (2020, July). Forecasting of monthly stochastic signal of urban water demand: Baghdad as a case study. In IOP Conference Series: Materials Science and Engineering (Vol. 888, No. 1, p. 012018). IOP Publishing.

[8] Ticherahine, Anissa, et al. "Time series forecasting of hourly water consumption with combinations of deterministic and learning models in the context of a tertiary building." 2020 International Conference on Decision Aid Sciences and Application (DASA). IEEE, 2020.

[9] Lüdtke, Deike U., et al. "Increase in daily household water demand during the first wave of the COVID-19 pandemic in Germany." Water 13.3 (2021): 260.

[10] Abu-Bakar, Halidu, Leon Williams, and Stephen H. Hallett. "Quantifying the impact of the COVID-19 lockdown on household water consumption patterns in England." NPJ Clean Water 4.1 (2021): 13.