Leakage Detection in a Real Water Distribution Network Through a Federated Prototype-Based Model

Diego Perdigão Sousa
Department of Teleinformatics Engineering
Federal University of Ceara
Fortaleza, Brazil
0000-0001-6408-2760

***Abstract*—This thesis explores important an intricate trade-off in intelligent systems: the efficiency of leakage detection tasks, the preservation of sensitive information, and the interpretability of the intelligent system. Through a comprehensive analysis of these interconnected domains, we aim to design a low-cost algorithmic solution that can efficiently learn from industrial data collected by distributed sensors while preserving sensitive information. We focus on the following topics: (i) efficient leakage detection on water distribution networks; and (ii) formulation of federated prototype-based models. In particular, this thesis mainly focuses on proposing an efficient and low-complexity distributed solution for identifying potential leaks in water distribution networks in municipal areas while ensuring the privacy of the hydraulic data. To this end, we explore and extend existing theories and methods from prototype-based learning and federated learning. We consider a hydraulic dataset, which includes water pressure and flow measurements obtained from pumps within district-metered areas in Stockholm, Sweden.**

***Index Terms*—federated learning, leakage detection, prototype-based models, water distribution network.**

I. Introduction

This thesis mainly focuses on proposing an efficient and low complexity distributed solution for identifying potential leaks in water distribution networks (WDNs) in municipal areas while ensuring the privacy of the hydraulic data. To this end, we explore and extend existing theories or methods on prototype-based models (PBMs) and federated learning (FL). Considering this principal objective, we summarize the following general questions that guided the development of our work.

- Research questions:
  1) How can we build a distributed solution for detecting anomalies that satisfies the subsequent conditions: (i) it demands low computational cost to process and predict the occurrence of anomaly behavior; (ii) it holds a high level of interpretation/understanding of the generated models; (iii) it can be formulated by using, adjusting and extending existing theories and methods from the literature; (iv) it can be extended preserving their level of interpretation/understanding.
  2) What performance can be achieved in a proof-of-concepts experiments applied to water leakage detection scenarios in which just a low level of description of the WDN is available?
  3) What are the minimum required knowledge about the network architecture of the WDN that is possible to build a reliable solution?
  4) What hydraulic feature(s) is(are) most relevant to generate the machine learning models?
- Research problems:
  1) How can one effectively utilize prototype-based models?
  2) What concepts and methods from federated learning can leverage models grounded on prototype-based learning?
  3) How to formulate a mathematical problem that properly describe the federated prototype-based models?
  4) How to demonstrate the interpretation potential of solutions built by using federated prototype-based models?
  5) How much can performance be enhanced by employing the proposed distributed approach as opposed to the corresponding centralized version?
  6) How can one design improved federated prototype-based models from our contribution to future work?

Thesis submitted to the Graduate Program in Teleinformatics Engineering of the Technology Center at the Federal University of Ceará, as a partial requirement for obtaining the title of Doctor in Teleinformatics Engineering. Concentration Area: Signals and Systems. Approved on January 30th, 2024. Supervisor: Prof. Dr. Charles Casimiro Cavalcante. Co-supervisor: Prof. Dr. Carlo Fischione.

II. Contributions and outline of the thesis

1) Prototype-based learning applied to water leakage detection: In the first part of our work, we have considered a traditional modeling paradigm, which contemplates the existence of a server responsible to collect and process all samples from every device and build machine learning models. Specifically, our goal is validating our assumption that PBMs can efficiently

detect leakages in WDNs. Therefore, we have investigated research questions 1(i-iii), 2, 3, and 4, and the first research problem along this stage of the work.

To achieve our goal, we designed representative sets with a reduced number of prototypes for generating a compact and realistic dataset for fault detection/classification of the monitored water distribution network. Specifically, we first clustered the observed water pressure data into understandable subgroups; in the following, we trained prototypes to represent the generated subgroups; finally, we used the trained prototypes to process operational condition predictions for newly observed water pressure data.

Within the context of the PBMs, we proposed low-complexity strategies based on both unsupervised and supervised learning. For the unsupervised method, we used the conventional K-means and cluster validation techniques. For the supervised method, we used crucial learning vector quantization (LVQ) classifiers. Specifically, we determined the number of prototypes through a clustering and cluster validation procedure per class label that can determine an adequate number of prototypes to obtain representative subsets of the input data. Then, we fine-tuned the prototypes of these generated subgroups using LVQ classifiers.

We have investigated this assumption and first reported our findings in [1]. Then, we published an extended version of our work in [2]. Subsequently, we have concluded that our premise is valid, hence we directed our efforts to the second stage our work.

2) Federated prototype-based learning applied to water leakage detection: In the second part of our work, we have investigated the task of detecting leakages in WDNs by proposing a distributed approach to PBMs (see Fig. 1). In particular, we extended our initial formulation by designing a solution grounded on FL. Moreover, we further extended the previous analysis by including water flow measurements. Since we have utilized a federated modeling paradigm, we have considered that every device is responsible for processing their collected samples and generating local models. Meanwhile, the server only processes local models to build an efficient global model. Consequently, we have investigated the last research question, 1(iv), and the remaining research problems, 2, 3, and 4, during this stage of the work.

Similar to the first stage of the work, we created realistic and compact sets by using a reduced number of prototypes for generating representative samples for fault detection/classification of the monitored WDN. Specifically, firstly, we trained the prototypes to represent the observed hydraulic data into comprehensible subgroups. Then, we used the trained prototypes to process operational condition predictions. Finally, we compared the performance between the distributed and centralized approaches concerning the sum of the squared errors minimization, the clustering cohesion, and the analysis of the generated Voronoi regions.

In the context of the Federated PBMs (FPBMs), we proposed the distributed algorithm federated averaging (FedAvg) by extending the conventional PBM winner-takes-all (WTA) and presenting a device-oriented learning rate. We have investigated this assumption and submitted our findings to [3] and [4].

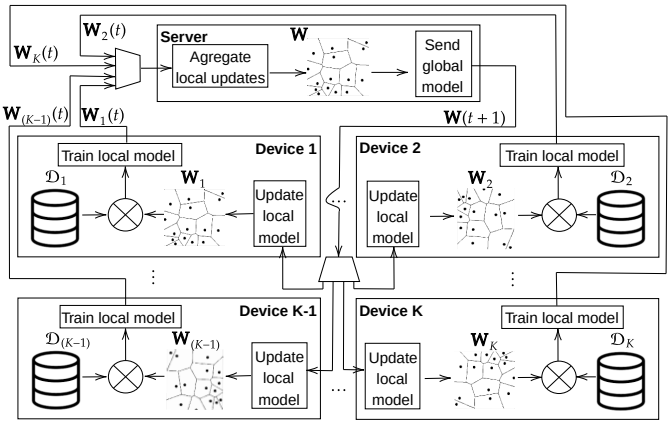

Fig. 1. Fed-WTA scenario with K devices and a server

III. Conclusions and future directions

In this thesis, different topics of ISs are studied. In particular, different designs of learning models for machine learning (centralized and distributed learning) have been explored to address the water leakage detection problem in WDNs. Nevertheless, the proposed solutions can be adjusted to be applied to diverse problems in anomaly detection. Therefore, in this final chapter, we summarize our key findings, as well the future directions for further research.

A. Conclusions

In this thesis, we explored low complexity machine learning models for water leakage detection in WDNs through the analysis of observed hydraulic data by means of emerging machine learning strategies. In particular, we proposed a distributed algorithmic solution for leakage detection. Specifically, our methodology used techniques from prototype-based learning and FL paradigms.

To evaluate our proposed solution, we considered real world data from water pressure and flow measurements from pumps in a residential district-metered area (DMA) of the WDN of Stockholm, Sweden. An important aspect to highlight is the required amount of data to properly generate the predictive system. We acknowledge that scenarios with non-sufficient data for training could lead to significantly misleading outcomes when anomalous behaviours in DMAs are analyzed. Therefore, we analyzed the total amount of collected data (15 months) that the SVOA company has shared with us to conduct our machine-learning strategies. From the water utility side, the company may continuously collect new observations to increase the reliability of the predictive system.

We divided our investigation into two parts. Firstly, we investigated the task of detecting water leakage by proposing a modeling solution grounded on prototype-based models, which are categorized as centralized learning. Then, we extended our

initial predictive model by proposing a distributed approach that can be applied to every prototype-based models.

On the concern of the experiments conducted along the first part of the thesis, we evaluated the potential benefits of adopting strategies grounded on prototype-based learning (PBL) for leakage detection of monitored WDNs. In particular, we obtained interesting classification rates in scenarios extremely limited of resources and, consequently, preset low computational cost. In addition, we analyzed the significance level of each pumping station in the regard of the predictive model.

Then, in the context of the second part of this thesis, the numerical findings showed the viability and potential benefits of combining PBL and FL. Although our analysis are constrained to FedAvg, we hope our insights can inspire future work established on other FPBMs.

Moreover, since we considered only one standard learning rate for the traditional model in the comparative analysis and our proposed learning rate for the distributed model is optimized to manage the presence of non-IID data, we obtained better purity results for all pumping stations in the federated learning than the centralized learning scheme. In this regard, we hope our findings can encourage future research on other learning rate strategies that can improve both PBMs and FPBMs.

To conclude, we highlight the main contributions of this thesis: (i) efficient leakage detection on water distribution networks; and (ii) mathematical formulation of FPBMs. From the last contribution, it is valuable to emphasize that the proposed learning rate is device-oriented and can be used in the modeling of distributed frameworks of PBMs.

B. Future directions

There are still have many interesting ideas, problems and challenges that remain to be investigated for this topic. Some relevant ones are highlighted in the following.

1) **Design improvement of the study case:**
   - For future works, we aim to evaluate our proposed FL in diverse sampling periods. Such analysis has high importance when dealing with anomaly detection tasks, such as the required acquisition time before prediction;
   - When a higher level description of the WDN is available, it is possible to use such knowledge about the network architecture to apply clustering methods aiming to divide the DMA and reduce the search area for the localization of the predicted leakages. In addition, it is also possible to label each DMA and analyze the entire WDN by extending our solution to a multiclass leakage detection problem. Therefore, our machine learning strategies can be extended and support solutions formulated by hydraulic modeling.
2) **Future work on federated prototype-based learning:**
   - Firstly, we are evaluating in the convergence of our proposed federated solution into more complex scenarios, including partial device participation;
   - Moreover, we are extending our proposed formulation of the FedAvg algorithm to other PBMs, such as the LVQ variants we used in the first part of this thesis. To this end, we are utilizing the device-oriented learning rate we proposed for the FedAvg algorithm to set the learning rate of those more complex variations of PBMs.
   - Furthermore, inspired by the developed methodology during the first part of our work, we target to integrate clustering validation techniques into our proposed distributed algorithmic solution. This corresponding analysis has high relevance, since this improvement can adjust the number of prototypes during the training of the FPBM.
   - Lastly, our proposed federated solution and the future directions presented above can also be applied to the WDNs available on the BattleDIM platform.

REFERENCES

[1] D. P. Sousa, R. Du, B. da Silva Jr, J. Mairton, C. C. Cavalcante, and C. Fischione, "Leakage detection in water distribution networks: Efficient training by data clustering," in *IWA World Water Congress & Exhibition, 11-15 September 2022 Bella Center— Copenhagen, Denmark*. IWA Publishing, 2022.

[2] D. P. Sousa, R. Du, J. Mairton Barros da Silva Jr, C. C. Cavalcante, and C. Fischione, "Leakage detection in water distribution networks using machine-learning strategies," *Water Supply*, vol. 23, no. 3, pp. 1115–1126, 02 2023. [Online]. Available: https://doi.org/10.2166/ws.2023.054

[3] D. P. Sousa, J. Mairton Barros da Silva Jr, C. C. Cavalcante, and C. Fischione, *A Federated Prototype-Based Model for IoT Systems: A Study Case for Leakage Detection in a Real Water Distribution Network*. John Wiley & Sons, 2025.

[4] D. P. Sousa, B. da Silva Jr, J. Mairton, C. C. Cavalcante, and C. Fischione, "Federated learning for water leakage detection using prototype-based models," in *International Conference Innovations for the Blue Planet, 23-25 April 2024 — Stockholm, Sweden*, 2024.