proceedings

Abstract

Concept Drift Mitigation in Low-Cost Air Quality Monitoring Networks †

Gerardo D'Elia [1,2,*], Matteo Ferro [3], Paolo Sommella [2], Sergio Ferlito [1], Saverio De Vito [1] and Girolamo Di Francia [1]

1 TERIN-FSD Division, ENEA CR-Portici, P. le E. Fermi 1, 80055 Portici, Italy; sergio.ferlito@enea.it (S.F.); saverio.devito@enea.it (S.D.V.); girolamo.difrancia@enea.it (G.D.F.)
2 Department of Industrial Engineering (DIIn), University of Salerno, 84084 Fisciano, Italy; psommella@unisa.it
3 Hippocratica Imaging S.r.l., Via Giulio Pastore, 84131 Salerno, Italy; mferro@unisa.it
* Correspondence: gerardo.delia@enea.it; Tel.: +39-081-7723-616
† Presented at the XXXV EUROSENSORS Conference, Lecce, Italy, 10–13 September 2023.

Abstract: Future air quality monitoring networks will include fleets of low-cost gas and particulate matter sensors calibrated using machine learning techniques. Unfortunately, it is well known that concept drift is one of the primary causes of losses in data quality in operational scenarios. This work focuses on addressing a low-cost NO_2 sensor calibration model update triggered via a concept drift detector. This study defines which data are most appropriate for use in the model updating process in order to maintain compliance with the relative expanded uncertainty (REU) limits established by the European Directive, as well as evaluate the potential of general and importance-weighted calibration models in the mitigation of concept drift effects.

Keywords: air quality network; concept drift; general calibration; relative expanded uncertainty

check for updates

Citation: D'Elia, G.; Ferro, M.; Sommella, P.; Ferlito, S.; De Vito, S.; Di Francia, G. Concept Drift Mitigation in Low-Cost Air Quality Monitoring Networks. *Proceedings* **2024**, *97*, 2. https://doi.org/10.3390/proceedings2024097002

Academic Editors: Pietro Siciliano and Luca Francioso

Published: 12 March 2024

1. Introduction

The data quality of low-cost gas sensors calibrated by means of machine learning techniques is still the crucial factor limiting their spread. Recently, the introduction of a concept drift detector in the chain value of data formation, able to provide an alert for the possible retraining or update of the calibration model, has been proposed [1]. At this point, what to do after a re-calibration request has arrived is an open question. The first question is about which data to select for the update and whether the reference data are available or not. An alternative model that we will consider for tackling this issue is the general calibration model [2]. This approach has been introduced as an attempt to reduce calibration costs. It consists of identifying and applying a general calibration model to all co-located nodes, thus avoiding the need for additional ad hoc calibrations, since the inherent variability of each individual sensor is incorporated into a single model. Also, an importance-weighted calibration model has been considered. Its procedure is to "weigh" the samples of the test set in order to "match" the distribution used during the training phase [3].

2. Materials and Methods

The winter 2020 co-location campaign dataset created in Portici (Italy), during which three MONICA devices (AQ6, AQ11, AQ12) were co-located with a reference mobile laboratory for two months and characterized by the presence of concept drift, is used to address concept drift handling in two steps: (i) selecting the appropriate data to use for the calibration model's update and (ii) exploring the general and importance-weighted calibration models' performances as alternative models. The dataset is divided into eight one-week time slots and, after that, the concept drift in T4 is characterized by the worst-quality NO2 estimations for the subsequent time slots [4]. Relative expanded uncertainty

(REU) is the metric adopted. Three options are investigated for data selection: the data preceding the concept drift alert (called "Last"), the data following the concept drift alert ("Next"), or parts of both ("Mixed") [5]. The main idea is to try to mitigate the effects of concept drift by exploiting the information content of the co-location data as much as possible, so the following two models will be explored.

General calibration model: if n sensors are placed in co-location, then the set of the medians of all the single quantities involved in the calibration model's creation constitutes the training set which the general calibration model is trained on.

Importance-weighted calibration model: The importance of a sample (the "weight") is calculated as the ratio between the probability density functions of the test and training sets. Once the weights are obtained, these will be applied in the training process, obtaining a new calibration model.

3. Results and Discussion

The REU charts show that the "Next" approach is to be preferred over the others, but it has a drawback: the node keeps releasing poor-quality data. This amount of invalid data is reduced by applying the "Mixed" approach. However, the data contained in the "Last" batch are not usable since they are not representative of the "Next" operational scenario. The general calibration model works well for the AQ6, suggesting that it is efficient at the mitigation of the concept drift's consequences (the REU plot drops below 25% at 45 $\mu g/m^3$), while for AQ12 devices it matches ad hoc model performance. The AQ11's intrinsic variability makes this instrument too different from the others. The application of the importance-weighted calibration model for AQ6 and AQ12 does not improve their performances compared to the ad hoc model, whilst it works in the T5 time slot for the AQ11 device.

4. Conclusions

The preliminary results of this work make clear the effectiveness of both the proposed methodological approach and the alternative calibration models used, as well as extending the validity of their calibrations. A further investigation is ongoing, aimed at further improving the obtained results through the use of a stacking ensemble which embeds the general calibration model and importance-weighted calibration model as base learners.

Author Contributions: Conceptualization, G.D. and P.S.; investigation G.D., M.F. and P.S.; methodology, S.F. and S.D.V.; data curation, G.D., S.F. and M.F.; validation, G.D., S.F. and P.S.; supervision, G.D.F.; funding acquisition, S.D.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the EU through the UIA Air-Heritage and VIDIS Project Horizon 2020 Research and Innovation program under grant agreement No 952433.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Supporting data are currently available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bagkis, E.; Kassandros, T.; Karatzas, K. Learning Calibration Functions on the Fly: Hybrid Batch Online Stacking Ensembles for the Calibration of Low-Cost Air Quality Sensor Networks in the Presence of Concept Drift. *Atmosphere* **2022**, *13*, 416. [CrossRef]
2. Malings, C.; Tanzer, R.; Hauryliuk, A.; Kumar, S.P.N.; Zimmerman, N.; Kara, L.B.; Presto, A.A.; Subramanian, R. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmos. Meas. Tech.* **2019**, *12*, 903–920. [CrossRef]
3. Sugiyama, M.; Krauledat, M.; Müller, K.R. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **2007**, *8*, 985–1005.

4. D'Elia, G.; Ferro, M.; Sommella, P.; De Vito, S.; Ferlito, S.; D'Auria, P.; Di Francia, G. Influence of Concept Drift on Metrological Performance of Low-Cost NO_2 Sensors. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1004811. [CrossRef]
5. Baier, L.; Reimold, J.; Kühl, N. Handling concept drift for predictions in business process mining. In Proceedings of the 2020 IEEE 22nd Conference on Business Informatics (CBI), Antwerp, Belgium, 22–24 June 2020; pp. 76–83.