DOI 10.5162/6EuNetAir2016/10

Machine learning for future intelligent air quality networks

S. De Vito[1], E. Esposito, M. Salvato, G. Fattoruso and G. Di Francia
[1] ENEA - Agenzia per le Nuove Tecnologie, l' Energia e lo Sviluppo Economico Sostenibile.
Energy Technologies Dept., C.R. Portici, P.le E. Fermi, 1 - 80055 Portici (NA) - Italy

Abstract
During the last few years, machine learning emerged as a very effective tool for data analysis and sematic value extraction from the large amount of data generated from deployed chemical multisensors devices. Many works have now highlighted the potential impact on multisensor device calibration, drift counteraction, data assimilation, optimal deployment of these classes of algorithms. Unlike 5 years ago, the huge amount of available data make possible to confirm this potential on real-world long-term deployments. This work analyze the literature produced by EuNetAir partners extracting the lessons cooperatively learnt about their impact and propose a novel architecture for future intelligent air quality networks based on the machine learning emerging paradigm.

Key words: Machine learning, chemical multisensor devices, distributed algorithms.

Introduction

Air quality in city landscapes is one of the most concerning environmental factors affecting our health and quality of life. Assessing and forecasting air quality is now paramount for our society in order to limit the issues, for ensuring awareness and ultimately to contribute to ameliorate the overall air quality conditions. Unfortunately, air quality assessment in complex scenarios like city landscapes is hampered by costs and dimension factor when relying on bulky but certified analyzers. This cause the development of sparse measurement matrices that cannot cope with intrinsic spatial variability of air quality conditions due to difference in emissions and fluidynamic transport problems. EU regulatory framework and, specifically, the 2008 directive on air quality begun to address the issue of defining data quality objectives for so called “indicative” measurements designed to complement the static and coarse grained conventional analyzer networks [1]. Small, low cost and wearable/portable gas microsensors devices are targeted as the next revolution in this field allowing to build high density air quality networks and personal air quality monitoring systems. However, precision and accuracy of these systems is severely negatively affected by technological limits such as non-linearity, low specificity and sensibility to environmental conditions as well as low stability commonly known as the drift problem. Slow dynamic is also known to affect their capability to promptly react to rapid transient occurring when operating at street level in fixed or mobile settings (i.e. when crossing a gas plume emitted by a truck, while cycling beyond a car, while operating at a traffic stop sign, etc.). These ultimately prevent them to reach DQO levels as described by EU/2008 Air quality directive for indicative measurements. During the last few years, within the chemical sensing community and specifically in the EuNetAir community, researchers have conducted several studies highlighting the potential impact of machine learning techniques in this realm. Machine learning (ML) is a complex framework oriented to the development of techniques that make computing machines capable to deal with problems without being specifically programmed to and ultimately being able to generalize their knowledge to unseen situations. Their usage has proven useful to deal with several sensors limitations like low selectivity, concept and sensor drifts, slow dynamics and so on. This paper try to summarize, at the end of the action, the lesson that the community has learned through these years proposing a novel paradigm for truly intelligent air quality monitoring networks.

On Field Calibration

In the late 2000, several works have highlighted the impact of on field deployment on lab calibrated multisensor devices. The different and continuously changing environmental conditions and the complexities of the real world chemical mixtures make their lab calibration suffering from significant inaccuracies. Low selectivity and interfering gases influence was identified as the primary issue leading to the introduction of on field multivariate calibration techniques, i.e. calibration extraction processes based on the use of the response of all the sensor array to the gases encountered during a real world deployment. A co-located conventional analyzer is used as a provider of true concentration levels [2]. In this framework, by using techniques like neural networks (NN) these researchers managed to use non selectivity to their advantage improving the overall performances of chemical multisensor devices facing air quality monitoring tasks. These was obtained without the explicit needs to model the sensor response to complex chemical mixtures of several pollutants and without the need to setup long instrument measurement campaigns in certified laboratories. All sensors responses RSens_i was functionally related with the actual pollutants concentrations C_j for all gases for which a certified reading is available (eq. 1):

$$C_j = \Psi(\mathrm{RSens}_i) \qquad (1)$$

The functional relationship was then learnt by example by training a machine learning tool like a neural network. These early results was then confirmed for multiple species.

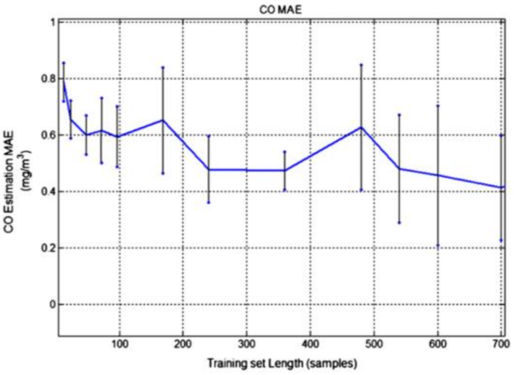

Fig. 1. Mean Absolute error on CO estimation obtained by a NN regressor as a function of the number or samples (hours) used for training phase. The Pirelli-ENEA dataset was used [8].

Performance is linked to the number of available sample for implementing the training process and one week seems the minimum amount for reasonable outcomes (see Ref. [2] and Fig. 1). Recently, these techniques emerged as the best performing tool for chemical multisensor devices in a well-known works series by JRC [3]. Scalability, though, remains an issue that can be currently solved only with ad-hoc systems for parallel recording of multiple multisensory unit responses together with the collocated conventional certified analyzers. Another issue to mention is the possible ”locality” effect that could limit the optimal performance when overall conditions that are significantly different from those faced during on field calibration are met by the deployed systems. Currently no information is available on the real extent of these potential issues.

Drift Counteraction

While non linearity and cross sensitivity have been confirmed to be positively affected by approaches like neural networks, drift counteraction has still to find an efficient solution ultimately affecting the possibility to reach the needed long term unmanned operation capability. However, multiple solutions have been proposed basing on machine learning techniques that seems to provide interesting results on a limited set of dataset with which they have tested. Here we would like to review two techniques. The first technique was proposed by De Vito et al. [4]. Based on semi-supervised learning, it seems to provide an interesting feature helping systems to learn also by unlabeled examples, i.e. sensors data for which a true concentration value is not available. As cheap as they are to obtain, unlabeled samples can be used to extend the knowledge of the machine learning tool obtained on a limited set of costly labeled samples that can only be obtained by sensors co-location with a conventional analyzer. This may actually reduce the number of labeled samples to use to obtain a sufficiently accurate calibration or to adapt the obtained knowledge to new situations arising from sensor poisoning or just seasonal environmental effects. These specific results have been obtained by the cited publication by using a one year long dataset recorded using a MOX based multisensor device on the field. The dataset is available on the UCI repository [8]. The second significant approach to long-term drift counteraction may exploit continuous or active learning ML paradigms. In this case new knowledge represented by couples of sensors reading and true concentration levels may be added when available to the knowledge set of the machine. In this way, the tool may adapt his knowledge to the arising but slow drift effects. One of the way to achieve this capability was initially shown by Tsujita et al. in the early 2000 [5]. He and his co-authors shown that, in particular circumstances, the response of the sensors can be corrected from baseline drift. Practically, when the regional background was near to zero, this situation, detected by a network of analyzers and a simple underlying model, can be used to reset the zero response of the sensor. This is a simple and neat example in which a simplified spatial model may be used to counteract the drift effect in the sensors that feed it. However, adaptive learning can be also based on the actual temporary co-location of a multisensor system with a peer or conventional/certified analyzer. In this way, a higher accuracy estimation of the pollutant concentration level may be exchanged by the on-board machine learning algorithms. Eventually, the involved actors may ameliorate their calibration accuracy continuously adapting it to the continuously changing sensors health status and environmental conditions. A relevant example of the possibility opened by this technique, has been shown by L. Thiele and coauthors partially using simulated data [6].

Overcoming Dynamic Issues

Slow dynamic affecting all chemical sensors significantly affects the overall accuracy with enhanced effects arising when sensors are subjected to rapid concentrations transients. These transients may become recurrent in mobile deployments where spatial gradients may be transformed in temporal gradients as well as near-to-road fixed deployments. These effects have been highlighted by a study of University of Cambridge and the authors by using a dataset recorded in the city of Cambridge focusing on electrochemical sensors systems [7]. In this work, the authors propose the use of dynamic machine learning architectures and specifically tapped delay neural networks. These architectures search for a functional relationship between the recent history of the sensors response and the actual concentration value. In this way they appear to be able to learn using the dynamic behavior of the sensor as an input actually learning to represent sensors inner dynamics (see fig. 2). As stated previously also in this case the capability of the network to learn this complex relationship may be limited by the conditions encountered during calibration period showing a “locality” effect.

Intelligent Air Quality Networks

Recent outcomes of the significant researchers effort in these fields can be now combined in order to propose a novel complete architecture that could represent the future of air quality monitoring mesh.

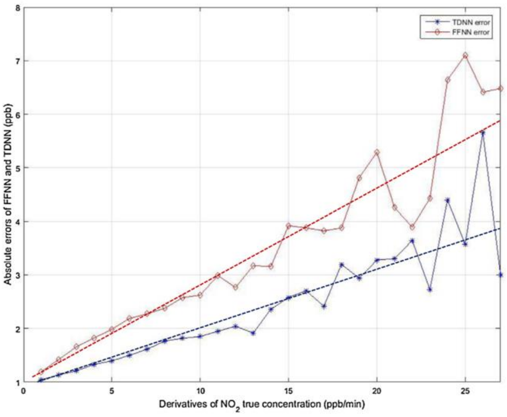

Fig. 2. Accuracy improvement obtained by dynamic machine learning approach with respect to a static machine learning tool while estimating NO2 true concentration at different derivative rate in a real world setting.

In particular, we propose a pyramidal architecture combining different interacting systems each showing different and time varying performance levels cooperating to reconstruct a high time and space resolution image of the pollution in the complex city landscape. Basically, we envisage the cooperation of a hierarchy of networks of heterogeneous systems providing different precision and accuracy while measuring different air quality parameters simultaneously in different positions and while moving (see fig. 3 and 4). Each system is intelligent in the sense that it is equipped with adaptive machine learning systems, calibrating their raw sensors response in a stochastic concentration estimation. This data can be used at different levels to contribute to the general assessment, for example, by contributing location of the measurement together with the concentration as well as uncertainty levels. In this sense, citizen science framework, sported by several recent projects, have paved a way for more pervasive and massive experimentations. Contributed data can be integrated with model based ad-hoc sensor fusion algorithm to reconstruct assessments and forecasting maps. Maps could then be used to provide accurate readings and assess and forecast exposure for people not having their own personal exposure monitor. Mobile systems, in particular move themselves (carried out by moving cars or just people) throughout the cities eventually meeting each other or coming close to conventional analyzers. Their calibration is known to loose accuracy and precision while sensor drift set in. In this sense their calibration is ageing but could be revamped by the use of fresh precise on field calibrated data coming from a more accurate assessor either a fixed conventional analyzer, a freshly calibrated (or recalibrated) multisensor device or eventually by assessment coming from model based pollution maps. As we have seen, many of the needed technological tools for the described architecture have indeed been already proposed with different aims in the literature.

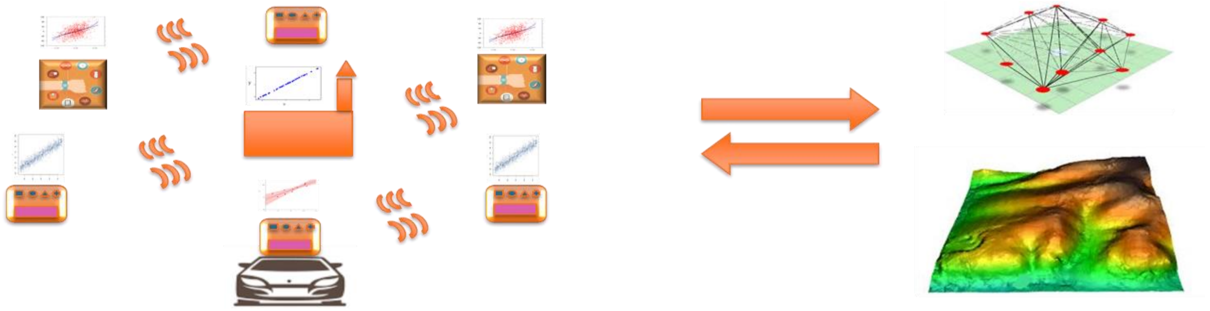

Fig. 3. Proposed architecture for future intelligent air quality monitoring networks integrating wearable, fixed and certified sensing devices. An accuracy mediated interaction among the different accuracy sensing modules, ensure the continuous learning and upgrade of the on board calibration rules. Model based interpolations provide real time upgraded high spatial accuracy pictures of the city air pollution levels using a data assimilation scheme. Model outcomes may be also used to upgrade calibrations.

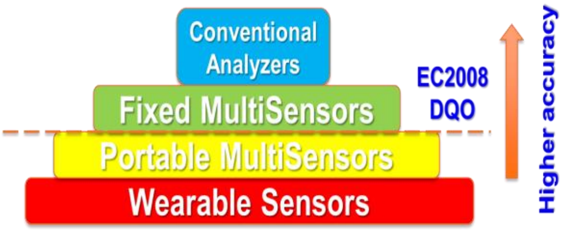

Fig. 4. An example of the accuracy/numerosity pyramid in pervasive air quality networks.

Conclusions

In this work we briefly review the recent literature highlighting the possible transformative impacts of machine learning techniques on air quality monitoring techniques involving the use of chemical multisensors units. In order to achieve truly spatially dense and accurate assessments, we believe that a significant innovation is needed at architectural level to correctly assimilate data from pervasive and mobile deployments of heterogeneous systems in geospatial models. To this purpose we briefly introduce an uncertainty mediated architecture integrating and exploiting the new possibilities opened by machine learning systems.

Acknowledgements

Authors would like to acknowledge the partial funding from COST Action TD1105 EuNetAir.

References

[1] DIRECTIVE 2008/50/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL.

[2] S. De Vito et al., On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sens. and Act. B: Chemical 129(2): 750-757, http://dx.doi.org/ 10.1016/j.snb.2007.09.060.

[3] L. Spinelle, et al., Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: O_3 and NO_2, Sens. and Act. B: Chemical 215:249-257, http://dx.doi.org/10.1016/j.snb.2015.03.031.

[4] S. De Vito et al., Semi-Supervised Learning Techniques in Artificial Olfaction: A Novel Approach to Classification Problems and Drift Counteraction, IEEE Sens. Journ. 12, (11), pp. 3215-3224, 2012. doi: 10.1109/JSEN.2012.2192425.

[5] Tsujita W. et al., Gas sensor networks for air pollution monitoring, Sensors and Actuators B Chemical 110(2):304-311.

[6] D. Hasenfratz et al., On-the-Fly Calibration of Low-Cost Gas Sensors, Volume 7158 of the series Lecture Notes in Computer Science pp 228-244.

[7] E. Esposito, et al., Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems, Sens. and Act. B: Chem. 231: 701-713, http://dx.doi.org/ 10.1016/j.snb.2016.03.038.

[8] ENEA-Pirelli “Air Quality” Dataset - https:// archive.ics.uci.edu/ml/datasets/Air+Quality.