In **20th Intl Conf on Artificial Neural Networks (ICANN 2010), Lecture Notes in Computer Science 6354**, (K. Diamantaras *et al*, eds.), Thessaloniki, Greece, Springer Verlag, 2010, pp.451-456.

Data mining methods for quality assurance in an environmental monitoring network

Ioannis N. Athanasiadis[1], Andrea-Emilio Rizzoli[1], and Daniel W. Beard[2]

[1] Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland
ioannis@athanasiadis.info ; andrea@idsia.ch
[2] Environmental Protection Department, Saudi Aramco, Dhahran, Saudi Arabia
daniel.beard@aramco.com

Abstract. The paper presents a system architecture that employs data mining techniques for ensuring quality assurance in an environmental monitoring network. We investigate how data mining techniques can be incorporated in the quality assurance decision making process. As prior expert decisions are available, we demonstrate that expert knowledge can be effectively extracted and reused for reproducing human experts decisions on new data. The framework is demonstrated for the Saudi Aramco air quality monitoring network and yields trustworthy behavior on historical data. A variety of data-mining algorithms was evaluated, resulting to an average predictive accuracy of over 80%, while best models reached 90% of correct decisions.

1 Introduction

Sensor network recordings are prone to several types of failures related to noise, polarization, calibration, physical obstacles, humidity, communication latency, variability of meteorological conditions, or other network and sensor faults. The quality assurance process of environmental monitoring networks is critical for supporting both the scientific work and the dissemination of environmental conditions at operational time. Quality assurance and quality control procedure (QA/QC) deals with the major uncertainties that characterize the natural environment. Typically, QA/QC takes place off-line, on batches of measurements, processed by expert scientists. Experts identify erroneous measurements and validate sensor recordings, based on statistical indexes, boundary rules, and more importantly their experience. The process is empirical, thus it can not be fully automatized, as expert knowledge is hard to capture and depends on location, season and type of equipment.

The Saudi Aramco Air Quality Monitoring and Meteorology Network (AMMNET) is no exception to the above situation [1]. Expert scientists are routinely asked to review the quality of the recorded measurements, and restore missing or erroneous ones. The target of AMMNET is air quality, where uncertainties are driven by the atmospheric chemistry and physics, and the stochastic nature of the major air pollutant emission mechanisms (including those of anthropogenic and natural origins). Concerning photochemical pollutants (such as ozone) the emission mechanism is dynamic, accompanied by strong non-linearities in the underlying physical and chemical mechanisms, and therefore it has been always among the major challenges of air quality modeling and

forecasting. These characteristics make quality control a demanding task that occupies significant efforts of expert scientists. Automating the QA/QC process for AMMNET or similar environmental monitoring networks is of extreme importance, as it will relieve scientists from a tedious and laborious task.

Near-real time decision support and surveillance systems requires the identification of erroneous measurements while the monitored conditions still occur. Data uncertainty problems associated with environmental monitoring networks bring forth issues of measurement validation and estimation of missing or erroneous values, which are critical for taking trustworthy decisions in a timely fashion [2]. This vision also drives towards the extension of legacy environmental monitoring infrastructure with modular, flexible software interfaces that provide with inputs to environmental decision support systems for hazard identification and incident forecasting [3]. This paper investigates how the quality assurance process can be improved by adopting data mining methods under both near real time constraints and off-line using historical data from an environmental monitoring network.

2 Towards a data-driven quality assurance system

Key requirements: A semi-automated quality assurance *system* needs to be able to evaluate the quality of the incoming measurements both online, as they are recorded into the system, and offline, as they are grouped in batches. Semi-automated processing is an important factor for near-real time dissemination [4], but also for identification of hazardous events and operational decision making [3].

To achieve this, a **quality index** is attached to each measurement recording at the time it arrives. Assume some sensor captures a measurement. Then, the *system* stores it, and preprocesses it by attaching a measure of quality for this recording, i.e the quality index. To compute the quality index the *system* may use all information available: past recordings from the same sensor, past and concurrent recordings of other sensors (in the same or other locations), and their statistics (i.e minimum, maximum or average values during a given period, e.g. hourly).

The inputs of the system are all raw measurements, as recorded by the sensor network, and the output is the "quality index" of the measurement. This setting was selected to mimic the information that is available to the experts: Experts investigate raw measurements and they are able to indicate the quality of the recordings. We want our data-driven decision making model to be able to do so. The amount of information available to the expert includes the current recording of the target pollutant, but also the concurrent and historical values of other pollutant concentrations and meteorological attributes, with their statistics. By visually inspecting all these time-series an expert can assess the quality of the target pollutant concentration. By presenting past data with the correct decisions attached to a supervised learner, we can train it, using a classification algorithm, to approximate the behavior of the expert. Then, the produced decision making model can be used for future predictions. In this respect the *system* is deployed around a **decision model** that is triggered by incoming recordings and responds with their corresponding quality indices (Fig. 1). The *decision model* functions as a *transformation* function, that utilizes a set of inputs originating from the sensor network in

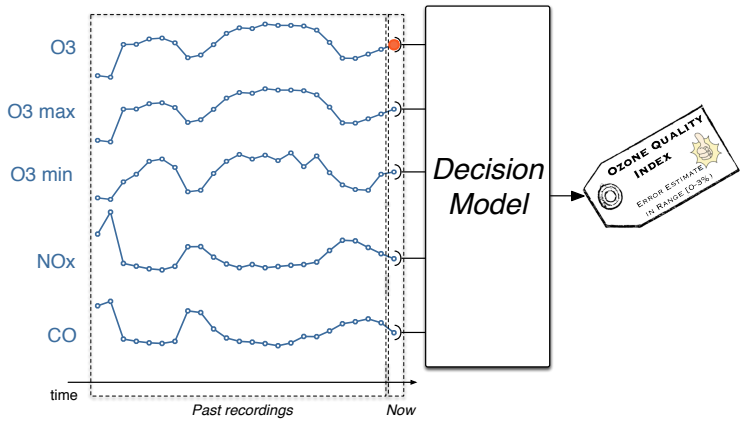


Fig. 1. The proposed abstract architecture
The decision model collects relevant time series data and responds on the quality of the currently arrived ozone measurement.

order to derive the quality index. It can be the result of a supervised learning process, i.e an expert system, a decision tree, a rule set or a neural network.

Another key aspect of the *system* is that it needs two modes of operation: one for the near-real time case and one for the off-line evaluations. In the first case, the decision model employs as inputs only concurrent and past measurements, while in the second case it may also make use of measurements recorded later.

Abstract architecture: We have designed a semi-automated quality assurance system for AMMNET, although the architecture is not specific to the network structure or the domain, and it can be deployed for other types of monitoring networks. Figure 1 illustrates the abstract architecture of the system that serves as a mediating layer between a sensor network and any end-user applications (i.e. database, monitoring or warning systems). The system mediates in order to assign quality indices to the incoming measurements. In Fig. 1 the system responds on the quality of an ozone measurement, on the basis of past and current measurements of ozone and other pollutants, and assigns a qualitative quality index (i.e. the ozone quality index is in the range of 0-3 percent).

The quality indices and problem formulation: The quality of a recording of the monitoring network can be measured using the relative error of the raw value with respect to the correct one, as a quantitative index. In practice, we can consider the correct value to be the one that is proposed by the quality assurance expert. Let $p(t)$ be the raw value recorded at time t and $\hat{p}(t)$ the correct value, then the relative difference of the two values defines the *quantitative quality indicator*, as:

$$e_p(t) = \left| \frac{p(t) - \hat{p}(t)}{\hat{p}(t)} \right| \times 100\% \tag{1}$$

The *quantitative quality indicator* values can be classified into crisp *qualitative quality bands*, which in turn serve as a figure of merit of the raw data quality.

In the envisioned automated system, a set of thresholds $\{s_1 < s_2 < \cdots < s_t\}$ will be used for defining the quality index for pollutant p as:

$$QI_p(t) = \begin{cases} \text{In range } [0, s_1) & \text{if } e_p(t) \leq s_1 \\ \text{In range } [s_1, s_2) & \text{if } s_1 < e_p(t) \leq s_2 \\ \cdots & \\ \text{In range } [s_t, +\infty) & \text{if } e_p(t) > s_t \end{cases} \quad (2)$$

Each range can be associated with some textual information. For example, 'credible' (in range:0–5%), 'suspicious'(in range: 5–15%), and 'distorted' (in range: >15%). The discretization of the quality index into bands with an associated labels allows for a qualitative evaluation of raw measurements. This is particularly interesting for an online system that is able to provide with an automatically computed quality index of the incoming recordings as they arrive. The selection of the appropriate thresholds for the *Quality Index* is a matter of design choice[3]. More importantly, the use of categories as a quality index instead of the real error, allows to reformulate the problem from a function approximation to a classification problem: Instead of approximating the error value, we aim to design a system that approximates the error range.

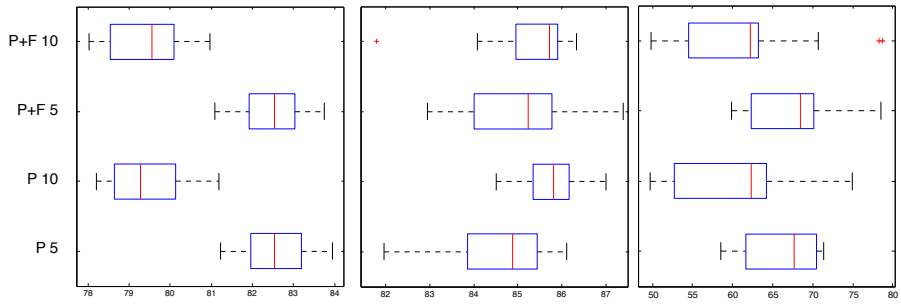
3 System deployment: The case of AMMNET

Design choices: Deploying a data-driven decision making system for AMMNET involves three important design choices. First comes the selection of the appropriate thresholds to define quality bands and indices. The latter serve as the outputs of the decision model. Second is the selection of the appropriate inputs for the decision model. And finally, the selection of suitable algorithms to implement the decision model.

To evaluate the feasibility of the approach, we conducted three rounds of experiments: We started with an **explorative round**, where several alternative system formulations and algorithms were screened, in order to valuate their potential. Secondly, we performed an **extensive evaluation round**, where selected algorithms and system configurations have been thoroughly studied in order to conclude to a statistically credible performance evaluation. Finally, the study concluded with a **blind-test round**, where the decision models induced in the previous round were used for newly collected data, not evaluated yet by experts and the results were presented to experts. The results of our experiments are summarized below in Section 4.

Data available: In our evaluations we focused on the quality assurance of **ozone**, and the goal was to build a data-mining model for the **ozone quality index**. We used historical data from three stations of the AMMNET network (namely Rahimah, Riyadh and Dhahran) covering in the period from January 2005 through June 2007. In each station data from several gases are monitored (namely: O_3, H_2S, SO_2, NO, NO_X, NO_2, CO), along with meteorological attributes (namely: Dew Point, Precipitation, Pressure, Solar Radiation, Temperature (at $2m$ and $10m$), Wind Speed and Direction). For all measurements both raw and quality assured data were available, at hourly intervals accompa-

[3] It can also be defined by legislation, but still a set of arbitrary thresholds



(a) Dhahran (b) Riyadh (c) Rahimah

Fig. 2. Classification accuracy for all 25 training schemes, using 10-cross fold validation. Both hypotheses have been tested (marked as 'P' and 'P+F' for 'past' and 'past and future' data used as inputs), with two single quality thresholds set at 5% and 10%.

nied with sensory statistics (minimum value, maximum value, standard deviation). In our experiments, we used as inputs raw values.

4 Results

Explorative round of experiments: The first part of this study concentrated on the rough evaluation of alternative algorithms and inputs. We investigated a set of four alternative quality indices, 14 model inputs and 12 algorithms (with 161 configurations). In total, 27'048 evaluations were performed. In each one of them, a decision-model was trained with data spanning from January 2005 to December 2006, and the induced decision models were validate against human decision for a period covering January to June 2007. Results for Ryadh and Dhahran were very promising, for the cases where one or two quality thresholds were applied: the average performance was above 80% and the best achieved result were in the range of 95% of correct decisions. In terms of algorithm selection, most credible results were achieved by C4.5 decision trees [5], Bayesian Networks [6], Fuzzy Lattice Reasoning [7], and Multi Layer Perceptrons, implemented with WEKA [8].

Extensive evaluation round: In a second set of experiments, we evaluated thoroughly a set of algorithms in a statistically coherent way, to guard results against data partitioning biases, by applying the **10-fold cross validation** method. Both hypotheses were tested, considering a single quality threshold for ozone set at 5% or at 10%. As inputs we considered gas concentrations and meteorological attributes, accompanied by sensory statistics. We employed C4.5 decision trees with confidence factor pruning (9 configurations) and reduced error pruning (10 configurations), and bayesian networks with a simple hill climbing search algorithm (6 configurations). The classification accuracy for all training configurations are presented in Fig. 4. While the average prediction accuracy ranges above 85% for Dhahran and Riyadh, it is limited to below 75% for Rahimah. However, Rahimah's best results outperform those of the other two stations.

Blind test results: The third round of experiments has evaluated further the best models trained above in a blind test, inspired from the Turing test: We presented a (previously

unknown to the system) data stream to the decision models, and recorded model decisions. Then, model decisions have been presented to AMMNET experts, in order to analyze them and compare them with the in-house QA/QC process. Experts' response for Rahimah was very positive, while this was not the case for Riyadh.

5 Discussion and future work

This study concluded that data-mining algorithms selected and tuned during our experiments have yielded credible performance, and are capable for inclusion in a semi-automated system for quality assurance of a sensor network. However, expert involvement is still needed for model training and selection.

Time and space is a very important factor. The length of calibration period needs further investigation. Similarly holds for the proliferation of the extracted decision models, with respect to local incidents (i.e instrument calibration, scheduled maintenance, etc) and seasonal patterns that occur in each station. Influence across stations (spatial interactions) need further study. Another important issue is the *preferred type of failures*. Any automated system will make wrong decisions. The issue here is what type of wrong decisions we prefer. A stringent system that falsely rejects measurements is certainly preferred against an 'easy' system that falsely accepts wrong measurements.

Finally, semi-automation still remains the goal of a future course of action. An interactive *system* can be designed and deployed, so that it is capable of allowing experts to mark interesting cases that need to be considered, and to reject trivial or false patterns discovered. Such an approach will put the foundations for a win-win cycle, where experts instruct the *system* and review models extracted from data.

References

1. Beard, D.: Saudi Aramco Real-Time Air Quality and Meteorological Monitoring Network. In Information Technologies in Environmental Engineering (ITEE), Shaker (2005) 630–643
2. Athanasiadis, I.N., Mitkas, P.A.: Knowledge discovery for operational decision support in air quality management. Journal of Environmental Informatics **9** (2007) 100–107
3. Athanasiadis, I.N., Milis, M., Mitkas, P.A., Michaelides, S.C.: A multi-agent system for meteorological radar data management and decision support. Environmental Modelling & Software **24** (2009) 1264–1273
4. Athanasiadis, I.N., Mitkas, P.A.: An agent-based intelligent environmental monitoring system. Management of Environmental Quality **15** (2004) 238–249
5. Quinlan, J.R.: C4.5 Programs for Machine Learning. Morgan Kaufmann (1993)
6. Pearl, J.: Probabilistic reasoning in intelligent systems. Morgan Kaufmann (1988)
7. Kaburlasos, V.G., Athanasiadis, I.N., Mitkas, P.A.: Fuzzy Lattice Reasoning (FLR) classifier and its application for ambient ozone estimation. International Journal of Approximate Reasoning **45** (2007) 152–188
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations **11** (2009) 10–18