A complete pipeline for the fusion of multiple heterogeneous redundant sources

Marco S. Reisa,\*, Eugeniu Streleta, Ivan Castillob, You Pengb, Swee-Teng Chinb

aUniv Coimbra, CIEPQPF, Department of Chemical Engineering, Rua Sílvio Lima, Pólo II – Pinhal de Marrocos, 3030-790 Coimbra, Portugal

bThe Dow Chemical Company, Lake Jackson, USA

\* marco@eq.uc.pt

Abstract

In the present context, there is an abundance and diversity of cost-effective measurement systems and data collection methods. These are often able to provide multiple estimates for a given target quantity, property, or key performance indicator (KPI). Such estimates can originate from various sources and have different levels of quality, such as high-quality laboratory measurements, medium-quality online analytical devices, and low-quality IoT readings or soft sensor estimates. While the different sources contain redundant information about the variable of interest, they differ in the acquisition rate, modality, and quality (i.e., uncertainty). Dealing with these multiple aspects simultaneously can be a complex task, usually addressed case-by-case. This paper introduces an integrated data fusion pipeline capable of accommodating the heterogeneity of data from various sources while assessing and monitoring their quality. It is designed to be scalable and adaptable to the diversity of solutions available today and in the foreseeable future. The proposed approach is rigorously tested and compared against other single-source and multi-source estimation alternatives, in a real industrial system that showcases both its flexibility and state-of-the-art estimation accuracy.

**Keywords:** Regularized Bayesian fusion, information quality, redundant sources.

* 1. Introduction

Modern industrial processes are characterized by the availability of multiple sources of information, each with its own modality, associated quality, and sampling rate (Becerra et al., 2021). Integrating all these heterogeneous data sources into a coherent analysis pipeline poses significant challenges (El Faouzi, 1997). For instance, the implementation of machine learning (ML) models often faces difficulties in the face of missing information, independently of the underlying mechanism (multirate sampling, missing observations/blocks at random, etc.). A common solution is employing data imputation schemes, but these approaches need to be selected and fine-tuned case-by-case, increasing the complexity of the analysis pipeline, as well as the risk of introducing artifacts in data during the imputation operations. Furthermore, some of the existing approaches compromise the expected smoothness of the target variable profile over time (Gutiérrez et al., 2022; Qin, 2014). Therefore, an alternative methodology should be adopted, that is scalable to multiple sources and naturally embraces the raw nature of industrial data, and in particular the dimensions referred to above (multimodality, diverse quality, irregular sampling).

A prevalent situation often found in industrial settings, is the existence of high-quality/low-frequency data (e.g., laboratory measurements of the quality parameters, usually by applying standard reference methods), co-existing with medium-quality/medium-frequency measurements obtained from online instrumental methods (such as inferential models based on process analytics technology (PAT) soft sensors), and low-quality/high-frequency data, derived from process data (such as sensor readings of temperature, pressure, flow, located in the process streams and units) (Qin, 2014). However, when integrating such multiple sources of information, one often finds unstable and uneven estimates of the target variable, instead of smoothly time-varying predictions, as expected from the inertial effects created by the large units and phenomena taking place therein. This can be attributed to the different quality of the sources available and the asynchronous pattern of collected data. The sensor fusion pipeline proposed in this work also overcomes this problem in the case of redundant sources. Let us start by clarifying the meaning of this terminology.

The nature of the several information sources with respect to a target response can be classified as (i) redundant, (ii) cooperative, or (iii) complementary, depending on the superposition of the predictive information each one carries regarding the target response. Complementary sources offer distinct pieces of information that are relevant to explain the variability of the response. Cooperative sources provide partially overlapping information that reinforces each other, but also contain some unique components that are not shared and are relevant for prediction of the response. Finally, redundant sources provide independent estimates of the same underlying phenomenon (Castanedo, 2013). Existing data fusion methods often tacitly assume a given relationship for the sources, or use fusion algorithms based on black box modeling, such as deep learning (DL) (Gao et al., 2020), without explicitly considering the inherent relationships between the sources. This can lead to sub-optimal fusion results that hinder the full exploitation of data in industrial applications.

In this paper, a flexible pipeline is proposed that combines all redundant sources of information about a given target property of interest, taking into account their associated quality (which is first estimated and then monitored), and accommodating different sampling rates and modalities, in order to produce updated estimates with the expected smoothness level, compatible with the process dominant dynamic mode.

* 1. Proposed data fusion pipeline: EVE-RegBF-MEMS

To effectively combine redundant information sources, a pipeline was assembled that consists of the following main steps. First, (i) the redundancy assumption is rigorously assessed. Afterward, (ii) the quality of the sources is estimated using EVE (Error Variance Estimation). Then, (iii) a new fusion method, RegBF, is applied, that takes into account such quality assessment. Finally, to secure the reliability of the fused data (and ultimately, of the whole fusion process), the quality of the sources is monitored over time (iv), using the MEMS methodology (Mutual Error Monitoring of Sources). This method identifies and flags any potential faulty or inconsistent source of information.

* + 1. Classification of the sources

The first step is devoted to verifying if all the sources are of the redundant type. For such, a methodology was developed, that compares the predictive subspaces of the sources and assesses if they overlap (and are therefore redundant) or not (see Figure 1). In brief terms, we compute the projections of the reference source measurements onto the two sources under analysis, say and ( and , respectively), and calculate the angle, , between such projections: if , then and are complementary; if is close to , they are redundant, where is a redundancy parameter determined by Monte Carlo simulation. A statistical hypothesis test was also devised to support this decision (based on the bootstrap).

* + 1. Assessing the sources’ quality

The Error Variance Estimation (EVE) method estimates the error variance of different sources based on the differences calculated between them. EVE assumes that the variations of the sources result from the additive effect of state variability and independent Gaussian noise, . The EVE method consists of the following steps: (i) computing the variance for all possible combinations of paired differences between sources, (ii) formulating the system of equations for the observed variances of the paired differences, where the individual noise variances are the unknown parameters, and (iii) solving this system of equations using constrained weighted least squares (WLS). This process yields the individual error variance for each source. It is important to note that the EVE method requires at least three sources to be effective. With fewer than three sources, the system of equations is underdetermined.

* + 1. Regularized Bayesian fusion (RegBF)

RegBF (Figure 2) is a new fusion method that leverages the availability of different sources at time and the corresponding estimated uncertainties provided by EVE (and dynamically updated), to produce an improved estimate of the target response at . The methodology shares some similarity with Bayesian fusion but imposes an additional regularization constraint for the successive estimates that controls their fluctuation or volatility over time. This regularization implies that information from is considered for the purposes of obtaining the estimate at . Upon suitable tuning, the regularization has the effect of smoothing out the fused estimate to the expected dynamic behavior of the process. In other words, the regularization constraint provides an additional degree of freedom to align the estimation framework with the intrinsic dynamics of the system.

|  |  |
| --- | --- |
| *Figure 1 – Classification of the sources.* | Uma imagem com texto, diagrama, desenho, esboço  Descrição gerada automaticamente  *Figure 2 – Regularized Bayesian Fusion (RegBF).* |

* + 1. Monitoring the sources’ quality

The quality of the sources is monitored via a developed approach called mutual error monitoring of sources (MEMS). The main assumptions of the MEMS are that the probability of simultaneous failure of two or more sources is low in comparison to individual failures, and the source’s uncertainty is stable under normal operating conditions (NOCs). MEMS is based on the mutual errors between the sources. Exponentially weighted moving average (EWMA) control charts are used to monitor the paired errors. A key performance indicator (KPI) for each source is determined by counting the number of OOC situations where each source is involved. A KPI=0 means a flowless source, whereas one with a high KPI indicates potential problems, possibly needing maintenance intervention.

* 1. Wastewater treatment plant (WWTP) case study

The proposed fusion pipeline was tested with a real case study. Data was gathered from a Dow Chemical Company wastewater treatment plant (WWTP) where a toxic substance should be eliminated from the effluent. The WWTP's essential processing stages are depicted in Figure 3, encompassing a settling unit, a flotation unit, and a filtering unit. The critical phase for toxin removal occurs within the flotation unit, where additives are introduced into the water stream to create flocculates containing the toxin. Subsequently, these flocculates are separated and removed, resulting in a clarified liquid stream that is finally released into the environment.

|  |
| --- |
| Uma imagem com texto, captura de ecrã, Tipo de letra, design  Descrição gerada automaticamente  *Figure 3 - Schematic representation of the wastewater treatment plant.* |

Operating a WWTP poses significant challenges stemming from the intricate, dynamic nature of biological processes and the characteristics of the data collected. One notable challenge is that the quality parameters of interest are often accessible at lower frequencies compared to the acquisition rates of the installed process sensors. This can significantly reduce the effectiveness of process management and decision-making. Furthermore, first principles models are not available to describe the system with enough accuracy or have too many unknown or unreliably estimated parameters. On the other hand, available data originates from various sources, arriving at different frequencies, and exhibiting distinct structures and varying quality levels. While these sources are complex to handle, they represent the only viable source of information to use for managing and optimizing WWTP operations.

In this case study, multimodal data were gathered from multiple locations within the WWTP, including images, process sensor readings, and laboratory measurements (as illustrated in Figure 3). These data sources were used to develop three soft sensors through Random Forest (RF) regression: (i) RF Pro flotation utilizes process data from the flotation unit, (ii) RF FI flotation leverages image data from the flotation unit, and (iii) RF FI settling relies on image data from the settling unit. All three soft sensors are designed to predict toxin levels at sampling point 2 (Strelet et al., 2021).

The primary objective of this case study is to achieve accurate toxin level predictions at sampling point 2 in real-time, despite the irregular nature of the information sources and their diverse quality. Due to confidentiality reasons, all data from this case study was previously normalized. This does not affect the conclusions but just protects the confidentiality nature of the information.

* 1. Results and discussion

Initially, we confirmed that all information sources within the WWTP were redundant. Figure 4 illustrates one such situation for two sources, namely RF Pro and RF FI flotation. Similar outcomes were observed for the other pairs of sources.

|  |  |
| --- | --- |
| Uma imagem com texto, diagrama, file, círculo  Descrição gerada automaticamente  *Figure 4 – Classification of sources: RF Pro flotation vs RF FI flotation.* | Uma imagem com diagrama, captura de ecrã, texto, file  Descrição gerada automaticamente  *Figure 5 – Accuracy of the soft sensors (RF Pro flotation, RF FI flotation, and RF FI settling) in the test dataset.* |
| Uma imagem com texto, captura de ecrã, diagrama, Gráfico  Descrição gerada automaticamente  *Figure 6 – Time series of RegBF and MEMS results.* | |

The RegBF approach was applied using the outcomes of the EVE method for estimating source uncertainties within a time window spanning 14 units of time. The prediction performance under test conditions of the individual sources and RegBF is presented in Figure 5. Given the unavailability of the true toxin values in this industrial case, the performance was assessed using laboratory measurements as a reference source, as it exhibited the lowest uncertainty () (consequently, this source does not appear in Figure 5).

The analysis of the performance results indicates that RegBF outperforms the information sources based on soft sensors, owing to the several advantageous properties of RegBF. Notably, the fusion weights are dynamically updated through the EVE method (Figure 6 (b)). Additionally, the regularization parameter is dynamically updated based on the stability or changes in the toxin level. Specifically, a higher rate of changes in toxin state leads to a lower regularization weight. This aspect played an important role during the test period where the laboratory measurements presented several spikes in the toxin level (see Figure 6 (a)), contrasting with the remaining sources that were more stable. Importantly, all sources were treated in exactly the same way, without conferring any special status to the reference measurements.

This approach contributes to a stable performance even in a faulty region that coincides with the test domain. As observed in Figure 6 (c), the KPI obtained via MEMS for that region frequently ranks high for all sources when compared to the training domain. This suggests a reduced reliability of these models in the test domain, possibly due to extrapolation issues. In fact, a closer analysis revealed a notable shift between the test and training domains. Given that three out of the four sources are soft sensors, this shift can potentially impact their performance. Furthermore, it is essential to acknowledge that laboratory measurements are also not entirely error-free.

* 1. Conclusions

A new fusion pipeline for redundant sources was proposed and validated. It efficiently integrates diverse sources characterized by differing modalities, acquisition frequencies, and data quality, into an accurate estimation. It is worth emphasizing that, similar to any fusion algorithm, the performance of RegBF is intricately tied to the quality of the information sources. Therefore, the quality was estimated without relying on a reference source, using EVE, and monitored using MEMS, further enhancing the stability of RegBF. In addition to uncertainty estimation, the quality monitoring enabled by MEMS contributes to our understanding of when the sources operate under normal conditions or may be undergoing failure. Due to space limitations, the method descriptions provided here are necessarily concise, but more details will be provided in the follow-up publications.

References

Becerra, M. A., Tobón, C., Castro-Ospina, A. E., & Peluffo-Ordóñez, D. H. (2021). Information Quality Assessment for Data Fusion Systems. Data, 6(6), 60. https://doi.org/10.3390/data6060060

Castanedo, F. (2013). A Review of Data Fusion Techniques. The Scientific World Journal, 2013, 1–19. https://doi.org/10.1155/2013/704504

El Faouzi, N.-E. (1997). Heterogeneous Data Source Fusion for Impedance Indicators. IFAC Proceedings Volumes, 30(8), 1307–1312. https://doi.org/10.1016/s1474-6670(17)44002-x

Gao, J., Li, P., Li, P., Chen, Z., Chen, Z., & Zhang, J. (2020). A Survey on Deep Learning for Multimodal Data Fusion. Neural Computation, 32(5), 829–864. https://doi.org/10.1162/neco\_a\_01273

Gutiérrez, R., Rampérez, V., Paggi, H., Lara, J. A., & Soriano, J. (2022). On the use of information fusion techniques to improve information quality: Taxonomy, opportunities and challenges. Information Fusion, 78, 102–137. https://doi.org/10.1016/j.inffus.2021.09.017

Qin, S. J. (2014). Process data analytics in the era of big data. AIChE Journal, 60(9), 3092–3100. https://doi.org/10.1002/aic.14523