*Artificial Intelligence Research and Development*
*A. Cortés et al. (Eds.)*


*doi:10.3233/FAIA220320*

# Optimizing Online Time-Series Data Imputation Through Case-Based Reasoning

Josep Pascual-Pañach[a,b,1] , Miquel Sànchez-Marrè[b], Miquel Àngel Cugueró-Escofet[c]
[a] Consorci Besòs Tordera, Catalonia (*jpascual@besos-tordera.cat*)
[b] *Intelligent Data Science and Artificial Intelligence Research Centre (IDEAI-UPC) Universitat Politècnica de Catalunya (UPC), Catalonia (miquel@cs.upc.edu).*
[c] *Advanced Control Systems Research Group, Universitat Politècnica de Catalunya (UPC), Catalonia (miquel.angel.cuguero@upc.edu).*

**Abstract.** When working with Intelligent Decision Support Systems (IDSS), data quality could compromise decisions and therefore, an undesirable behaviour of the supported system. In this paper, a novel methodology for time-series online data imputation is proposed. A Case-Based Reasoning (CBR) system is used to provide such imputation approach. The CBR principle (i.e., solving the current problem using past solutions to similar problems) may be applied to data imputation, using values from similar past situations to replace incorrect or missing values. To improve the performance of the data imputation process, optimal case feature weights are obtained using genetic algorithms (GA). The proposed methodology is validated with data obtained from a real Waste Water Treatment Plant (WWTP) process.

**Keywords.** Online Data Imputation; Time-series; Case-Based Reasoning; Optimization; Intelligent Decision Support.

## 1. Introduction

Intelligent Decision Support Systems (IDSSs) operate using data obtained from different sources, such as sensors, and often in real time. The quality of these data is a common problem that should be tackled to ensure the good performance of the system. To solve the data imputation problem different machine learning techniques and models can be used. Here, a Case-Based Reasoning (CBR) approach is proposed in order to impute missing values in an online fashion, optimizing feature weights and considering the time through temporal CBR (TCBR). In [1] an imputation method based on a k nearest neighbours' algorithm is proposed and applied to a financial prediction problem. [2] proposes also the use of a reliable k nearest neighbours' (RKNN) algorithm applied to incomplete interval-valued data. In [3] a CBR approach for offline medium-gaps (from 3 to 10 missing values) imputation is proposed and applied to meteorological time series.

[1] Corresponding author: Josep Pascual-Pañach, Consorci Besòs Tordera, Av. Sant Julià, 241, 08403, Granollers, Catalonia; E-mail: jpascual@besos-tordeca.cat.

## 2. Methods

In this section, we present a methodology to calibrate a CBR system to impute missing values from online time-series. Figure 1 shows how the imputation process through CBR is integrated in the classical CBR cycle.

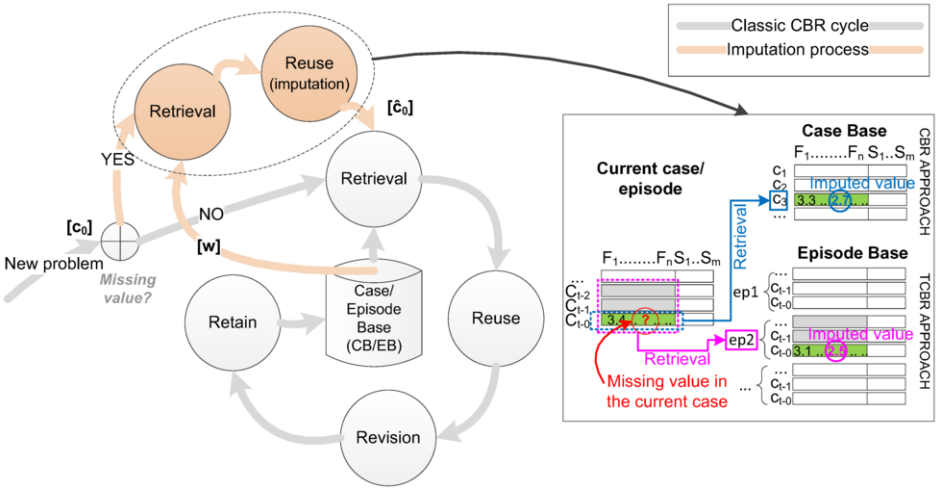


**Figure 1.** Integration of the CBR-based imputation approach in the CBR cycle

When a value from the current case $c_o$ is incorrect or missing, the available part of the case is used to find the most similar ones in the CB. Considering the TCBR approach, sets of consecutive cases (i.e., episodes) are used instead of particular cases in order to take into account data dynamics. This temporal approach is based on the one described in [4], and explained in another work under revision. The retrieval is done using episodes and giving the same importance to all cases. Regarding the imputation (reuse stage), in both CBR and TCBR the procedure is the same. The value of the missing feature is imputed using the corresponding value in the retrieved case or episode, obtaining a new case $\hat{c_0}$. When using episodes, the value from the most recent case in the episode is used (which is the one corresponding to the most recent one in the current episode).
Assuming that all features are numeric, a weighted Euclidean Distance (*wED*) similarity measure is used in the retrieval stage as in Eq. (1).

$$wED(c_0, c_i) = \sqrt{\sum_{n=1}^{N} w_n (f_{0n} - f_{in})^2} \quad (1)$$

where $c_0$ and $c_i$ are two cases, $f_{0n}$ and $f_{in}$ are the feature n values for each case, $w_n$ is the weight for feature n and N is the number of features. Feature weights are calculated in order to minimize the error between the predicted value and the measured value for a particular feature. The metric used is the Root Mean Square Error (RMSE), calculated as in Eq. (2).

$$RMSE = \sqrt{\frac{1}{J} \cdot \sum_{i=1}^{J} \big(y(i) - \hat{y}(i)\big)^2} \quad (2)$$

where $J$ is the number of samples in the dataset, $y$ is the measured data for a particular feature and $\hat{y}$ is the predicted value for the same feature.

## 3. Experimentation

To evaluate the viability of the proposed imputation method, historical data from the real process under study has been used to generate the CB and to calibrate the imputation system by calculating an optimal vector $w$ of feature weights. The optimization problem described in Section 2 has been solved using a Genetic Algorithm (GA). The problem to solve can be described as in Eq. (3):

$$\begin{aligned} &\min_{w} e(w) \ \ subject\ to: \\ &linear\ constraints: [1 \quad \dots \quad 1] \cdot w = 1 \\ &bounds: 0 \ \leq w \ \leq 1 \end{aligned} \tag{3}$$

where $w$ is the weights vector and $e(w)$ is the cost function to be minimized. The cost function integrates Eqs. (1) and (2) to optimize the retrieval process with the aim of minimizing the RMSE between the measured and predicted values.

The method is integrated in an IDSS based on the integration of CBR and Rule-Based Reasoning systems used to set adequate operational set-points to control the biological process in a real Waste Water Treatment Plant (WWTP) [5]. Presented results correspond to the imputation of medium gaps –6 samples (30 minutes) and 12 samples (1 hour) of missing values due to typical real faults, e.g., communication faults or invalid values during the sensor calibration process (a common sensor maintenance periodic procedure in the real facility. Here, faults are simulated in order to have the measured values of the whole dataset to evaluate the performance of the method. Figure 2 shows the performance attained with both CBR approaches using unfaulty data.

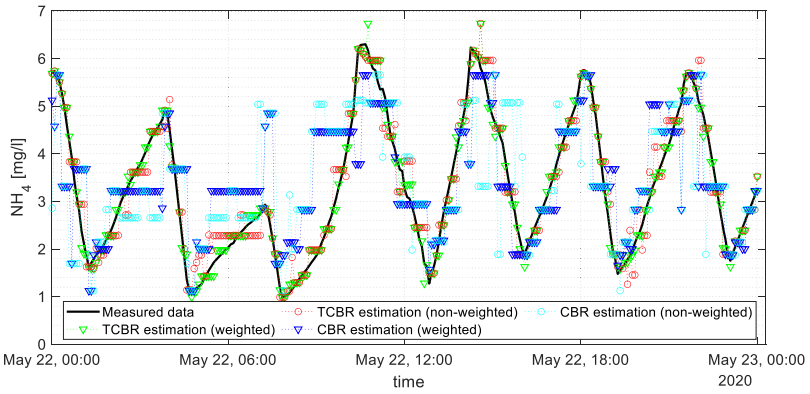


**Figure 2.** Different models are compared with unfaulty data

The best model (weighted TCBR) is validated with faulty data in Figure 3. Episodes have a fixed length of 6 samples, achieving a good trade-off between performance and computing time.

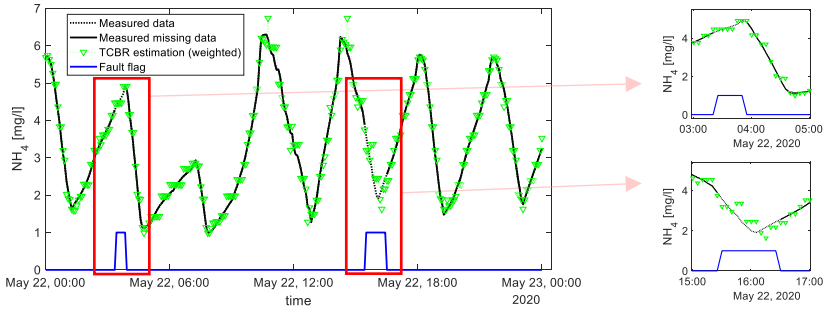


**Figure 3.** TCBR weighted model validated with 2 faults

## 4. Discussion, conclusions and next steps

This paper presents a data imputation method based on a CBR approach. The proposal has been evaluated using real data from a WWTP and considering different realistic medium missing data windows based on real faults in the ammonia sensor, which is a critical variable for the biological process control. An improved performance is obtained when using a calibrated CBR imputation system in comparison with the non-calibrated counterpart. The RMSE of the estimation with weighted features is almost 40% lower than the non-weighted estimation when using TCBR. Regarding the comparison between CBR and TCBR, the TCBR approach provides clearly better performance, with a RMSE about 60% lower than the calibrated CBR approach.

Next steps will consider a more in-depth evaluation of the method's performance with other sensors, different types of faults or multiple missing values, and the comparison with other classical time-series models and machine learning methods [6, 7] or state-of-the-art imputation techniques. Episodes' length will be also addressed.

## References

[1] Cheng, C.-H., Chan, C.-P., Sheu, Y.-J., 2019. A novel purity-based k nearest neighbors imputation method and its application in financial distress prediction. Engineering Applications of Artificial Intelligence 81, 283–299.

[2] Qi, X., Guo, H., Wang, W., 2021. A reliable KNN filling approach for incomplete interval-valued data. Engineering Applications of Artificial Intelligence 100, 104175.

[3] Flores, A., Tito, H., Silva, C. (2019). CBRm: Case based Reasoning Approach for Imputation of Medium Gaps. International Journal of Advanced Computer Science and Applications. 10. 10.14569/IJACSA.2019.0100949.

[4] Sànchez-Marrè, Miquel, Cortés, Ulises, Martínez, Montse, Comas, Joaquim, Rodríguez-Roda, Ignasi. (2005). An Approach for Temporal Case-Based Reasoning: Episode-Based Reasoning. Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). 3620. 465-476. 10.1007/11536406_36.

[5] Pascual-Pañach, Josep, Cugueró-Escofet, Miquel Àngel, Sànchez-Marrè, Miquel, Interoperating data-driven and model-driven techniques for the automated development of intelligent environmental decision support systems, Environmental Modelling & Software, Volume 140, 2021, 105021, ISSN 1364-8152.

[6] Cugueró-Escofet, M. A., García, D., Quevedo, J., Puig, V., Espin, S., Roquet, J., 2016. A methodology and a software tool for sensor data validation/reconstruction: application to the Catalonia regional water network. Control Engineering Practice 49, 159–172.

[7] Pascual-Pañach, J., Sànchez-Marrè, M., Cugueró-Escofet, M.A. (2022). Ensemble model-based method for time series sensors' data validation and imputation applied to a real Waste Water Treatment Plant. *11th International Congress on Environmental Modelling and Software, Brussels, Belgium.*