Estimating Missing Environmental Information by Contextual Data Cooperation

Davide Guastella, Valérie Camps, Marie-Pierre Gleizes

Estimating Missing Environmental Information by Contextual Data Cooperation

Davide Andrea Guastella[1,2(✉)], Valérie Camps[1], and Marie-Pierre Gleizes[1]

[1] Institut de Recherche en Informatique de Toulouse, Université de Toulouse III - Paul Sabatier, Toulouse, France
{davide.guastella,camps,gleizes}@irit.fr
[2] Università degli Studi di Palermo, Palermo, Italy

Abstract. The quality of life of users and energy consumption could be optimized by a complex network of sensors. Nevertheless, smart environments depend on their size, so it is expensive to provide enough sensors at low cost to monitor each part of the environment. We propose a cooperative multi-agent solution to estimate missing environmental information in smart environment when no *ad-hoc* sensors are available. We evaluated our proposal on a real dataset and compared the results to standard state-of-the-art solutions.

Keywords: Smart city · Cooperative multi-agent systems · Missing information estimation

1 Introduction

The concept of *Smart City* emerged in recent years as a way to exploit Information and Communication Technology (ICT) for improving services offered by a city and reducing its ecological footprint. Smart city initiatives are implemented by coupling *Ambient Intelligence* (AmI) and *Internet Of Things* (IoT). The idea behind AmI is to provide an environment with an interconnected network of coordinated IoT devices where the boundary between software and society blends and often disappears. As such, the environment is enriched with artificial intelligence to support humans in their everyday life [3,9]. The computational power of IoT devices coupled with widespread connectivity has increased significantly the development of initiatives to support smart cities. Such initiatives usually implement a monitoring activity of the environment in order to act on it in order to improve the energetic consumption, that is constantly increasing in the recent years [10], and ensure comfort to users. Nevertheless, the necessary devices can be intermittent, so they cannot guarantee continuous operability. In this case, it is necessary to provide accurate estimation of the values that ambient devices would provide if they were available. These estimations must be provided at real-time so that users can access to the information at any time [7]. In fact, a

https://doi.org/10.1007/978-3-030-33792-6_37

continuous monitoring of the environment can be useful when conceiving system to support smart city initiatives [6].

In this paper we propose a solution to estimate environmental information where ad hoc sensors are not available by using mobile and intermittent devices recurrent as well as historical data. Our proposal addresses the following challenges: (*i*) *intermittent data*: we exploit agents to provide accurate estimations when intermittent devices are used; (*ii*) *distributed processing*: each agent has its own local view of the environment, so that the estimation processes in different parts of the environment are independent and (*iii*) *online learning*: agents are capable of learning the dynamic of the environment at real-time without pre-processing data.

2 Proposition

To better understand the addressed problematic of estimating missing information in smart environments, let us consider a dataset of temperatures perceived by an *ad-hoc* device. If the device cannot provide information due to a malfunction at time i, the historical data perceived from the same device and other nearby devices can be correlated in order to provide an accurate estimation for the missing information. In this manner the system is able to evaluate an information that the device would provide if it worked.

Our proposal is based on a cooperative multi-agent system where each agent exploits data windows containing consecutive information in time, called *Ambient Context Windows* (ACW), in order to find recurrent dynamics in the historical data. ACWs are used to provide accurate estimations for missing information.

The rest of the section is organized as follow. In Sect. 2.1 we provide the definitions of the elements composing the proposed system, then in Sect. 2.2 we describe the general steps of our proposal.

2.1 Definitions

Definition 1 (Ambient Context Window). *An* Ambient Context Window *(ACW)* C_i *contains homogeneous environmental information perceived by an agent in a time window* $T = [t_k, t_i]$, $k < i$. *An ACW has* $|C_i| = |T|$ *homogeneous context entries, one for each time instant.*

Definition 2 (Context Entry). *A* Context Entry $E_t^i \in \mathbb{R}$ *is a punctual information perceived at time* $t \in T$, *where* T *is the time window of the ACW* C_i. *The value of a context entry can be any type of environmental information such as temperature, humidity, lightness etc.*

Definition 3 (ACW Distance). *The distance between two ACWs is defined as the absolute difference in time between the context entries divided by the number of entries* γ *of the two ACWs. The smaller the difference is, the more similar two ACWs are. The context distance between two ACWs* C_i *and* C_k *is defined by to the following formula:*

$$d\left(C_i, C_k\right) = \frac{\sum_{l \in [1,\gamma]} |E_l^i - E_l^k|}{\gamma}$$

where $\gamma = |C_i| = |C_k|$.

The distance d *satisfies the following properties: (i)* $d(C_i, C_k) \geq 0$, *(ii)* $d(C_i, C_k) = 0 \iff C_i = C_k$, *(iii)* $d(C_i, C_k) = d(C_k, C_i)$, *(iv)* $d(C_i, C_p) \leq d(C_i, C_k) + d(C_k, C_p)$ *where* C_i, C_k, C_p *are ACWs for information times* i, k, p *respectively. Therefore* d *is a metric.*

Definition 4 (Ambient Context Agent). *An* Ambient Context Agent *(*ACA_i*) identifies an ACW related to the information* i *in the dataset. Its goal is to provide environmental information. A cooperative behavior allows ACAs to provide environmental information even if a real device is unavailable.*

2.2 HybridIoT System Overview

In the proposed HybridIoT system, we suppose that data perceived by ambient devices are stored in a database and that the unavailability of a real device generates an exception as the ACAs is not able to provide an information. This exception is solved by exploiting the *Adaptive Multi-Agent System Approach* (AMAS) [5]. In this approach, an exception is considered as a *Non-Cooperative Situation* (NCS) that has to be solved in a local and cooperative way. In our problem, an incompetence NCS occurs when an ACA us unable to provide an environmental information (because no *ad hoc* sensor is available or it encountered a problem).

The main steps of the solution we propose are depicted in Fig. 1.

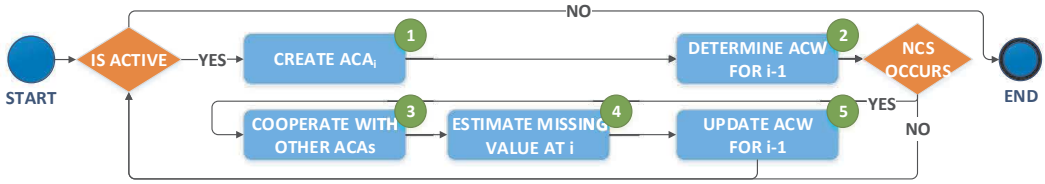

Fig. 1. The main steps of the proposed technique.

When data are available, ACAs are created on the fly and associated to the available information (step (1)). Then the agent determines a context window that is representative of the information perceived (step (2)). When data are not available due to the unavailability of the device or even a missing device, an exception we denote as *Non-Cooperative Situation* occurs; in this case the ACA cooperates with other agents (step (3)) in order to provide an accurate estimation for the missing information (step (4)). Once the information has been estimated, the ACW is being updated by the agent (step (5)).

When the information at time i is not available, the ACA_i has to cooperate with other ACAs by comparing their ACWs in order to determine an accurate estimation. These ACAs are chosen according to the distance between their ACWs and the ACW related to the agent that encountered a NCS. In this way the estimation is evaluated using the ACWs that are the most similar to ACA_i. The set ξ contains the ACWs that minimize the distance from the ACW that contains an information to be estimated.

When the set ξ of ACAs has been evaluated, a weight w is computed by a cooperative process between the ACA_i (that encountered a NCS) and the other ACAs by using each related ACW $C_k \in \xi = \{C_n, ..., C_p\}$, $n < p < i$, for which the distance $d(C_i, C_k)$ is minimized. The weight w is computed as follow:

$$w = \frac{\sum_{C_j \in \xi} \left(E_k^j - E_{k-1}^j \right) \cdot d\left(C_i, C_j\right)}{\sum_{C_j \in \xi} d\left(C_i, C_k\right)}$$

where C_i is the ACW containing the information to be estimated at time i, $C_j \in \xi$, $|\xi| = 10$, is the j-th most similar context window to C_i for which the distance $d(C_i, C_j)$ is minimized and E_k^j and E_{k-1}^j are respectively the k^{th} and $(k-1)^{th}$ context entries of the ACW $C_j \in \xi$ where $k = |C_j|$. Finally, let C_k be the ACW containing an information to be estimated; the estimated context entry E_i^k at time i is computed as follows:

$$E_i^k = E_{i-1}^k + w.$$

Once the information at time i has been estimated, the ACA_i evaluates a dynamic ACW, containing a number of information that is not specified *a priori* (step (5)). The relevance of dynamic size context windows is motivated by the fact that their use allows to obtain accurate estimations for missing information with respect to fixed size windows. More precisely, an ACW_i of dynamic size has a number of context entries that influences the capability of the related agent to make an accurate estimation at time i.

Consider two temperature datasets, one containing daily data and the other data perceived every 30 s. In the first case, the variance could be high. Contrary, in the second case the difference between each sample is relatively low, so as the variance. In this case the ACW of an ACA_i contains many samples while still providing a good estimation of the value at time i.

For an information at time i our solution creates a set $\Lambda_i = \{C_{i,0}, C_{i,1}, ..., C_{i,\lambda-2}\}$ of ACWs, where $|C_{i,k}| = k + 2$. We have fixed $\lambda = 16$, thus Λ_i contains a maximum of 15 contextual windows. Each ACW in Λ_i has at least 2 entries, that is the minimum number of information that can be used in the estimation process. The process of evaluation of dynamic size ACWs gives as output a context window $C_{i,k} \in \Lambda_i$ that minimizes the variance between the information. We verified through experiments that using 15 context windows is sufficient in order to find an ACW that best represent the information to which it is related.

Estimating accurate values for missing information depends on the evaluation of appropriate ACWs of variable size that better describe the information with which they are associated. Moreover, the evaluation of dynamic size ACWs depends on the availability of data, whether they are estimated or real. For this reason, our proposal is divided into two interdependent and coupled subsystems.

3 Experimental Results

The proposed framework has been evaluated using a dataset of 196 real temperature samples from 80 weather stations located in the region of Emilia Romagna, Italy, provided by the ARPAE service [4]. Data from this dataset are acquired daily by weather stations.

To evaluate our method, we applied a k-fold cross validation, whose partitions the original sample in k subsamples. Among the k subsamples, a single subsample is retained as the validation data for testing the classifier, and the remaining $k-1$ subsamples are used as training data. During the training phase, agents assemble the contexts windows for each information. The test phase is then repeated k times, with each of the k subsamples used exactly once as the test data. The k results from the folds are then averaged to produce a single performance estimation [8]. In our experiments, we used a k value of 5, 10 and 15.

The proposed solution has been coded in Java and the experiments were carried out on a computer equipped with i7 − 7820HQ, 32 GB RAM and Windows 10. The estimation of a missing value is practically instantaneous and the evaluation of the solution using cross-validation requires about one second for each station.

The solution has been coded without considering any particular optimization technique. Since the proposed solution has been tested on a single machine, we did not consider any computational overhead of agents such as communication. Moreover, in our proposal the communication between agents is asynchronous and we consider the communication costs as unitary, thus irrelevant for the estimations of missing values. Also, we did not use any specific agent-based technology and our solution is based on a cooperative resolution process between agents, which is technology independent. This allows us to prove the effectiveness of the proposed estimation technique rather than focusing on a specific agent-based architecture to address the estimation problem.

Figure 2 shows the mean error and standard deviation for the regional dataset. The mean error among the considered stations is $-0.092°$, the mean standard deviation is $1.3043°$.

3.1 Comparison to Standard Solutions

We compared the obtained results to different state-of-the-art solutions by using the KNIME analytic platform, a modular environment which enables easy visual assembly and interactive execution of a data pipeline [1].

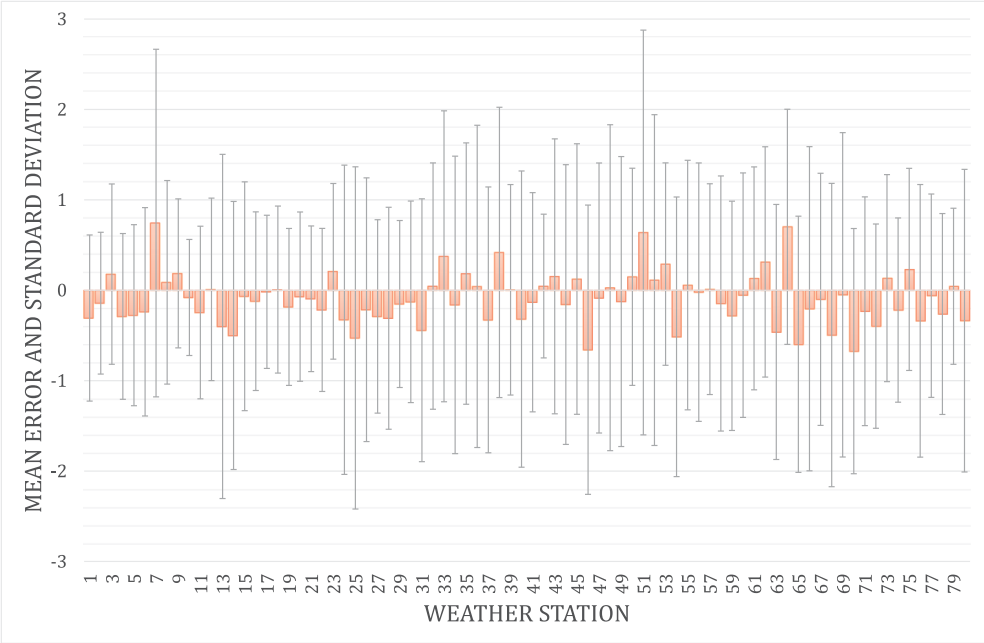

Fig. 2. Mean error bar and standard deviation of temperatures (degree Celsius) for the regional dataset.

A k-fold cross validation has been applied, as a specific node is available in KNIME, using 5, 10 and 15 validation iterations. We verified that when using such number of validation iterations the test set contains enough variation with respect to the training set.

For linear regression, the related node does not provide any configuration. For polynomial regression, we used a maximum polynomial degree of 2. For fuzzy rules, we used the *Best Guess* method to handle missing values [2], which computes the optimal replacement value by projecting the fuzzy rule (with missing value(s)) onto the missing dimension of all other rules. Also, we used *Product Norm* as rule to combine the membership values of each fuzzy interval for one rule and compute a final output across all rules and *Volumn Border Based* as shrink method to reduce rules in order to avoid conflicts between rules of different classes; this shrink method applies the volume loss in terms of the support or core region borders. These parameters gave us good results for the used dataset.

In order to compare to the state-of-art we used the four stations that gave the worst results using our method. The results of the comparison for the regional dataset are shown in Fig. 3. For the four stations considered, our proposal outperforms the results obtained by the state-of-the-art solutions.

In order to evaluate the effectiveness of dynamic ACWs, we used the same 4 stations shown in Fig. 3. We applied a k-fold cross validation using dynamic size ACWs, then we used fixed size ACWs containing a maximum of 3, 7, 10 entries respectively for each experiment. Figure 4 shows the results of the experiments using fixed size and dynamic size ACWs.

Even if using fixed size ACWs our proposal is able to make sufficiently good estimations; the system outperforms the results when using dynamic size ACWs.

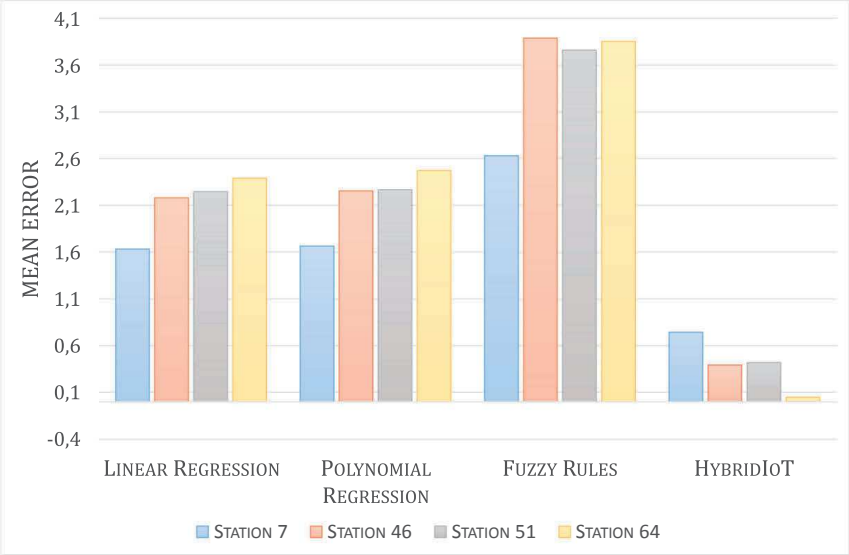

Fig. 3. Comparison of the mean error (degree Celsius) for our solution and standard techniques obtained from the regional dataset.

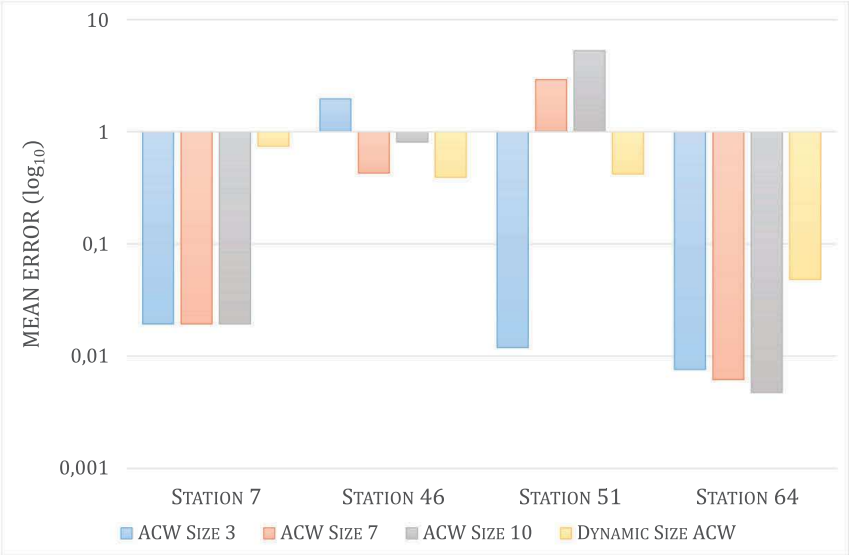

Fig. 4. Comparison of the mean error (degree Celsius) from 4 stations of the regional dataset using ACWs of fixed size (3, 7 and 10) and dynamic size ACWs. The mean error axis uses a logarithmic scale (base 10) as the error is significantly low.

Dynamic size ACWs have a twofold advantage. As we said in the previous section, they are able to better describe the dynamics of the information. Furthermore, when using dynamic size ACWs it is not necessary to specify the size of the context windows. In fact, the system must operate even when considering devices that have different frequency rate at which they perceive information. This is important in order to ensure that the system can operate in large-scale, open environments. In this manner our proposal is able to self-calibrate, thus it

does not require any parameter that depends on specific device configuration, making the system suitable for deployment at large-scale.

4 Conclusion and Perspectives

This paper proposes a cooperative multi-agent system using ACWs to estimate missing data from environmental devices whenever no real sensor is available in a smart environment. Our solution does not require any parameter and is capable of providing accurate estimation at runtime through a cooperative resolution process between agents. Our proposal has several advantages over the state-of-the-art solutions: (*i*) the system can be deployed at large scale thanks to the distributed computation of the agents, enabling a seamless integration in smart cities; (*ii*) agents have a partial view of the surrounding environment, so their computation does not interfere with agents which are located in different and delocalized environments; (*iii*) the cooperation allows to estimate accurate information at real-time. Contrary to classical solutions in which data are a passive entity, in our proposal, data become an active part of the system as the agents identify them. As such, agents are able to cope with non-availability of real sensors; (*iv*) although we considered only a temperatures dataset, the system is generic enough to work with any kind of environmental information without any modification.

In our future works we aim at improving the cooperation process between ACAs by involving ACAs that perceive heterogeneous information.

References

1. Berthold, M.R., et al.: KNIME - the konstanz information miner: version 2.0 and beyond. ACM SIGKDD Explor. Newslett. **11**(1), 26–31 (2009). https://doi.org/10.1145/1656274.1656280
2. Berthold, M.R., Huber, K.P.: Missing values and learning of fuzzy rules **06**(2), 171–178. https://doi.org/10.1142/S021848859800015X
3. Bikakis, A., Antoniou, G.: Defeasible contextual reasoning with arguments in ambient intelligence. IEEE Trans. Knowl. Data Eng. **22**(11), 1492–1506 (2010). https://doi.org/10.1109/TKDE.2010.37
4. Bressan, L., Valentini, A., Paccagnella, T., Montani, A., Marsigli, C., Tesini, M.: Sensitivity of sea-level forecasting to the horizontal resolution and sea surface forcing for different configurations of an oceanographic model of the Adriatic Sea. Adv. Sci. Res. (Copernicus) **14**, 77–84 (2017). https://doi.org/10.5194/asr-14-77-2017
5. Georgé, J.P., Gleizes, M.P., Camps, V.: Cooperation. In: Di Marzo Serugendo, G., Gleizes, M.P., Karageorgos, A. (eds.) Self-Organising Software: From Natural to Artificial Adaptation, pp. 193–226. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-17348-6
6. Guastella, D., Camps, V., Gleizes, M.P.: Multi-agent systems for estimating missing information in smart cities. In: Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART), pp. 214–223. SciTePress (2019). https://doi.org/10.5220/0007381902140223

7. Guastella, D.A., Valenti, C.: Estimating missing information by cluster analysis and normalized convolution. In: 2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI), pp. 1–6. IEEE, September 2018. https://doi.org/10.1109/RTSI.2018.8548454
8. Moreno-Torres, J., Sáez, J.A., Herrera, F.: Study on the impact of partition-induced dataset shift on k-fold cross-validation. IEEE Trans. Neural Netw. Learn. Syst. **23**, 1304–1312 (2012). https://doi.org/10.1109/TNNLS.2012.2199516
9. Sabatucci, L., Seidita, V., Cossentino, M.: The four types of self-adaptive systems: a metamodel. In: De Pietro, G., Gallo, L., Howlett, R.J., Jain, L.C. (eds.) KES-IIMSS 2017. SIST, vol. 76, pp. 440–450. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-59480-4_44
10. Tomazzoli, C., Cristani, M., Karafili, E., Olivieri, F.: Non-monotonic reasoning rules for energy efficiency. J. Ambient Intell. Smart Environ. **9**(3), 345–360.https://doi.org/10.3233/AIS-170434. (IOS Press)