A Bayesian approach for an efficient data reduction in IoT

Cristanel Razafimandimby, Valeria Loscri, Anna Maria Vegni, Driss Aourir, Alessandro Neri

A Bayesian approach for an efficient data reduction in IoT

Cristanel Razafimandimby[1], Valeria Loscrí[1], Anna Maria Vegni[2], Driss Aourir[1], and Alessandro Neri[2]

[1] Inria Lille - Nord Europe, Lille, France
{jean.razafimandimby_anjalalaina,valeria.loscri,driss.aourir}@inria.fr
[2] Department of Engineering, Roma Tre University
COMLAB Telecommunication Laboratory,Rome, Italy
{annamaria.vegni,alessandro.neri}@uniroma3.it

Summary. Nowadays, Internet of Things (IoT) coupled with cloud computing begins to take an important place in economic systems and in society daily life. It has got a large success in several application areas, ranging from smart city applications to smart grids. Despite the apparent success, one major challenge that should be addressed is the huge amount of data generated by the sensing devices. The transmission of these huge amount of data to the network may affect the energy consumption of sensing devices, and can also cause network congestion issues.
To face this challenge, we present a Bayesian Inference Approach (BIA), which allows avoiding the transmission of high spatio-temporal correlated data. In this paper, BIA is based on a hierarchical architecture with smart nodes, smart gateways and data centers. Belief Propagation algorithm has been chosen for performing an approximate inference on our model in order to reconstruct the missing sensing data. BIA is evaluated based on the data collected from the M3 sensors deployed in the FIT IoT-LAB platform and according to different scenarios. The results show that our proposed approach reduces drastically the number of transmitted data and the energy consumption, while maintaining an acceptable level of data prediction accuracy.

Key words: IoT, Belief Propagation, Markov Random Fields, Cloud, Smart Node.

1 Introduction

Despite of the large success of IoT, it raises yet many challenges and one of them is the management of massive amount of data generated by sensing devices. Locally storing this big data will not be possible any more. Therefore, harnessing cloud computing capacity is needed [3], but unfortunately this is not enough. However, it was observed that, with the increase of sensor density, data generated by IoT devices tend to be highly redundant. Thus, uploading raw data to the cloud can become extremely inefficient due to the waste of memory and network overloading.

To address this issue, we proposed in [6] and [7] an efficient Bayesian Inference Approach (BIA) in the IoT context for indoor and outdoor environments. For this aim, real data collected from sensors deployed in the Intel Berkeley Research lab [5] and in the PEACH project [9] have been used. Although these data allowed simulating the efficiency of our proposed approach, the lack of access to the deployed sensors did not allow us to experiment our Bayesian approach directly on the sensors. In this paper, in order to validate the scalability of our BIA approach and filter the raw data directly in the sensing nodes, we run experimentation on our FIT IoT-LAB platform [1]

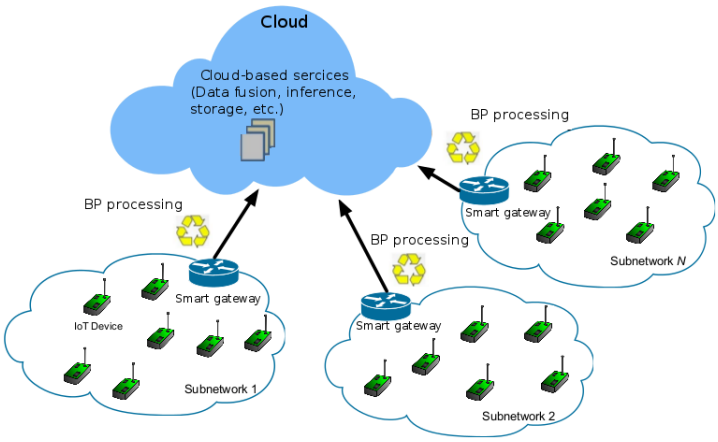


Fig. 1: A cloud-based IoT network model.

which is a very large scale infrastructure facility suitable for testing small wireless sensor devices and heterogeneous communicating objects over large scale.

The main contributions of this paper can be summarized as follows:

- Adoption of a Bayesian Inference Approach that allows avoiding the transmission of high spatio-temporal correlated in heterogeneous IoT networks. The Pearl's Belief Propagation (BP) algorithm [10] has been chosen to infer the missing data;
- Use of smart node and gateway in order to decrease the estimation error and increase the network lifetime. Smart in the sense that the node and gateway know exactly when to send or not the data;
- Performance assessment based on data collected from the M3 sensors deployed in the FIT IoT-LAB platform.

This paper is organized as follows. Section 2 describes our reference cloud-based architecture for the IoT scenario. In Section 3, we present the proposed Bayesian Inference Approach based on BP algorithm for data sharing in an IoT scenario. Section 4 provides the experimental results for the assessment of the proposed BIA technique in different real scenarios. Finally, conclusions are drawn at the end of the paper.

2 Network model

As depicted in Fig. 1, in this paper we propose a BP approach in a cloud-based architecture consisting of M3 nodes, smart gateways and data centers. Each entity in our architecture plays a different role w.r.t the functionalities, the computational and communication capabilities. Our IoT network model may include multiple subnets associated with different applications. In this paper, each subnet corresponds to one site of the FIT IoT-Lab testbed and is composed of IoT devices connected to each others for data sharing, and a smart gateway that relays the data flows to the cloud. The cloud in turn is responsible of data storage and all the cloud-based services.

3 Bayesian Inference Approach

In this section, we describe our BIA technique. As mentioned before, our main goal is to avoid sending useless data, while keeping an acceptable level of data content accuracy. For this aim, BIA is based on Pearl's BP algorithm that will be described below.

As a starting point before any inference procedure, the design of a graphical model should be provided. Graphical models are schematic representations of probability distributions. They consist of nodes connected by either directed or undirected edges. Each node represents a random variable, and the edges represent probabilistic relationships among variables. Models which are comprised of directed edges are known as *Bayesian networks*, whilst models that are composed of undirected edges are known as *Markov Random Fields* (MRF) [8]. In this paper, we present an inference approach under the hypothesis of MRF, modeled by means of Factor Graphs. It follows that our goal is to estimate the state X of the sensed environment starting from the sets of data collected by each sensor node. Based on the remarkable Hammersley-Clifford theorem, the joint distribution $P_X(x)$ of an MRF model is given by the product of all the potential functions *i.e.*,

$$P_X(x) = \frac{1}{Z} \prod_i \psi_i(x_i) \quad \prod_{i,j \in E} \psi_{ij}(x_i, x_j), \tag{1}$$

where Z is the normalization factor, $\psi_i(x_i)$ is the evidence function, E is the set of edges encoding the statistical dependencies between two nodes i and j, and $\psi_{ij}(\cdot)$ represents the potential function. Note that the graphical model parameters (*i.e.*, ψ_i and ψ_{ij}) can be estimated from the observed data by using a learning algorithm like in [4] and [2]..

For simplicity, in this paper, we consider widely used pairwise MRF, *i.e.*, MRF with the maximum clique [1] of two nodes.

One of the main goals when dealing with graphical models is the marginal distribution computation, as shown in Eq. (2). They are used to predict the most probable assignment for a variable node. For notation convenience, let us assume that X and Y are two distinct multivariate random variables with assignments $x \in \mathcal{X}^m$ and $y \in \mathcal{Y}^n$. The nodes in Y are called hidden nodes and those in X are the observed ones. So, given the i-th device in our network, x_i will be the observation of the phenomenon we intend to share (*e.g.*, pressure) and y_i will be associate to the phenomenon we want to infer, (*e.g.*, temperature)

$$p(y_v|x) = \sum_{y_1} \sum_{y_2} ... \sum_{y_n} p(y_1, y_2, y_3, ..., y_n|x). \tag{2}$$

Obviously, using (2), the complexity of a complete enumeration of all possible assignments to the whole graph is $O(|\mathcal{Y}|^{n-1})$, which is intractable for most choices of n. Therefore, we need a faster algorithm like Belief Propagation (BP) [2] [10] for computing the marginal probability. BP is a well known algorithm for performing inference on graphical models [10].

Let $p(y_i)$ represents the marginal distribution of i-th node, and BP allows the computation of $p(y_i)$ at each node i by means of a message passing algorithm. The message from the i-th to the j-th node related to the local information y_i is defined as:

$$m_{ji}(y_i) \propto \int \psi_{ji}(y_j, y_i)\psi_j(y_j) \prod_{u \in \Gamma(j), u \neq i} m_{uj}(y_j) dy_j, \tag{3}$$

[1] A clique is defined as a fully connected subset of nodes in the graph.
[2] Only take linear time.

where $\Gamma(j)$ denotes the neighbors of node j and the incoming messages from previous iteration are represented by m_{uj}. Notice that (3) will be performed between all nodes in the model until the convergence or if a maximum number of iterations I_{max} will be reached. Thus, the prediction *i.e.*, the belief at the i-th node, is computed through all the incoming messages from the neighboring nodes and the local belief, *i.e.*:

$$\hat{y_i} = belief(y_i) = k \cdot \psi_i(y_i) \prod_{u \in \Gamma(i)} m_{ui}(y_i), \tag{4}$$

where k is a normalization constant. Finally, it is worth to mentioning that the BP is able to compute the exact marginalization in the case of tree-structured graphical models.

4 Experimental results

In this section we provide the experimental results of our BIA approach using the FIT IoT-LAB testbed [1]. Ten nodes from Lille site and ten nodes from Grenoble site were used for the data collection. Nodes were of the M3 type [1], which are equipped with an 32-bit ARM Cortex-M3 MCU, 64 kB of RAM, 256 kB of ROM, an IEEE 802.15.4 2.4 GHz radio transceiver and four different sensors (light, accelerometer, gyroscope, pressure & temperature). Data collected from all the M3 nodes has been used to build the BIA model. Each data collection has been performed every 15 minutes and the collected data consists of 2.5 days of readings.

During the 2.5 days of reading, we noticed that there is a good correlation between pressure and temperature data (it is about -0.7720841). Hence, we can infer the temperature data from pressure data and vice versa. In this paper, we decided to infer temperature from pressure. The temperature is in degrees Celsius, whilst the pressure is in mbar.

We assess our approach w.r.t. (*i*) the number of transmitted data, (*ii*) average value of the estimation error (ER), (*iii*) average value of the distortion level as a Mean squared Error (MSE), and (*iv*) the energy consumption (EC).

The number of transmitted data represents the total number of data transmission performed by all the M3 nodes during the 2.5 days . In our energy consumption evaluations, we assume that the energy cost for sending each temperature and pressure value is 14 mW.

Furthermore, all of our assessments are based on three different scenarios (*i.e.*, s_1, s_2, and s_3). In scenario s_1, the M3 node sends to the gateway all the temperature and pressure data it receives. This means that the gateway does not perform any inference (*i.e.*, no inference). In the second scenario s_2, the M3 nodes sends only the pressure data to the gateway, and the gateway in turn infers the corresponding temperature data by using the BP algorithm. Finally, in the scenario s_3, we consider that the M3 nodes are "smart" devices, meaning that before sending their data to the gateway, they first compute the probability $\Pr(e|T,P)$ of making an inference error e on the gateway given the temperature data T, and the pressure data P. If there is a strong probability that the error magnitude *i.e.*, $|e|$, exceeds a predefined threshold *i.e.*, $|e|_{Max}$, the M3 node sends both pressure and temperature data to the gateway, else the M3 node sends only the pressure data, and the temperature value will be inferred in the gateway using the BP algorithm. This can be expressed mathematically as the inference error probability higher than a maximum allowed value $|e|_{Max}$, and conditioned to the temperature and pressure measurements *i.e.*, T and h, is lower or at least equal to a given threshold P_e^{Max}, that is:

Scenario	*#Transmitted data*	*EC* (kJ)	*MSE*	*ER*
s1	10440	1716.64	-	-
s2	5220	858.32	1.43	0.55
s3	5829	958.46	0.43	0.43

Table 1: Results obtained during the two days and half of readings.

$$\Pr\{|e| > |e|_{Max} | T, P\} \leq P_e^{Max}, \tag{5}$$

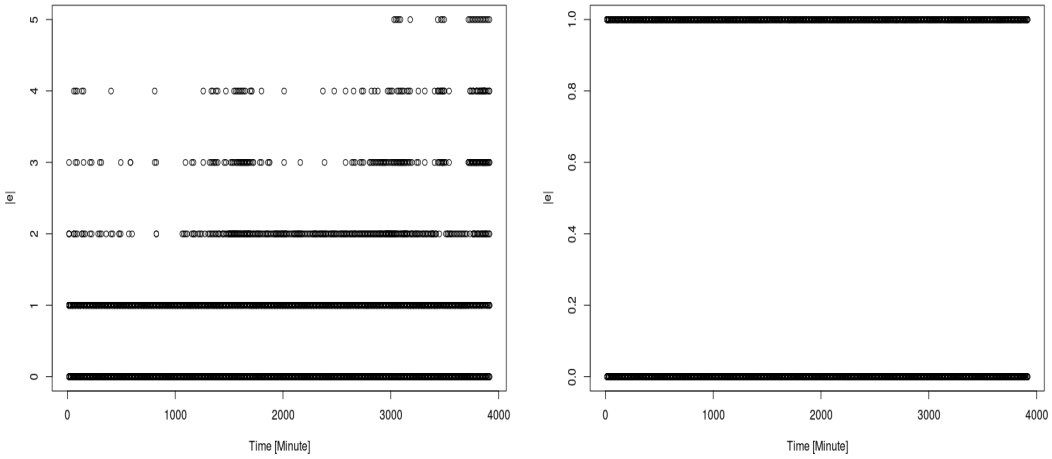
where the computation of $\Pr(e|T, P)$ is done by means of the BP algorithm. It should be noted that this computation requires the knowledge of the a priori probability of inference error *i.e.*, $\Pr(e)$. Also, the value of the threshold $|e|_{Max}$ strictly depends on the application context. In our case, we set this value equal to 1 but later we will see how the choice of this value may influence our results. A similar consideration can be applied to the probability threshold P_e^{Max}, which has been set to 0.5.

Table 1 illustrates the obtained results during 2.5 days of readings, for different simulated scenarios. We can notice that our Bayesian inference approach drastically reduces the number of transmitted data and the energy consumption, while maintaining an acceptable level of prediction accuracy and information quality. We can notice also that we decrease considerably the estimation error by using the scenario s_3. Indeed, the M3 nodes are smarter in this case *i.e.*, by computing the a posteriori probability of the inference error, the M3 nodes will be able to estimate the right moment and the data type to send in the gateway. However, this increases the number of transmitted data (and hence the energy consumption), as compared to scenario s_2. This is due to the fact that in s_2, the M3 node send only the pressure data without worrying of the risk of inference error in the gateway. It is important to say that we have a good quality of information in the scenario s_3 despite the fact that we have an inference error of 43%. This is due to the fact that we allow only a maximum error of one unit (i.e $|e|_{Max} = 1$)

Fig. 2 shows the variation of $|e|$ during the 2.5 days of reading using s_2 and s_3 [3], where $|e|$ is the difference between the true value and the inferred one of temperature data *i.e.*, $|e| = |\hat{y}_i - y_i|$. This metric illustrates therefore the inference error of our BIA approach during all the readings. No inference error occurs for $|e| = 0$, *i.e.*, when $\hat{y}_i = y_i$. In s_2, for the majority of time we notice no inference error *i.e.*, the probability of having a null inference error is $\Pr(|e| = 0) = 45.13\%$, while we have $\Pr(|e| = 1) = 41.83\%$, $\Pr(|e| = 2) = 6.91\%$, $\Pr(|e| = 3) = 4.04\%$, $\Pr(|e| = 4) = 1.60\%$, $\Pr(|e| = 5) = 0.45\%$,and $\Pr(|e| = 6) = 0\%$. Best performances are for scenario s_3, where we observe no error for the 57.32% of time, while we have $\Pr(|e| = 1) = 42.68\%$ for the remaining time.

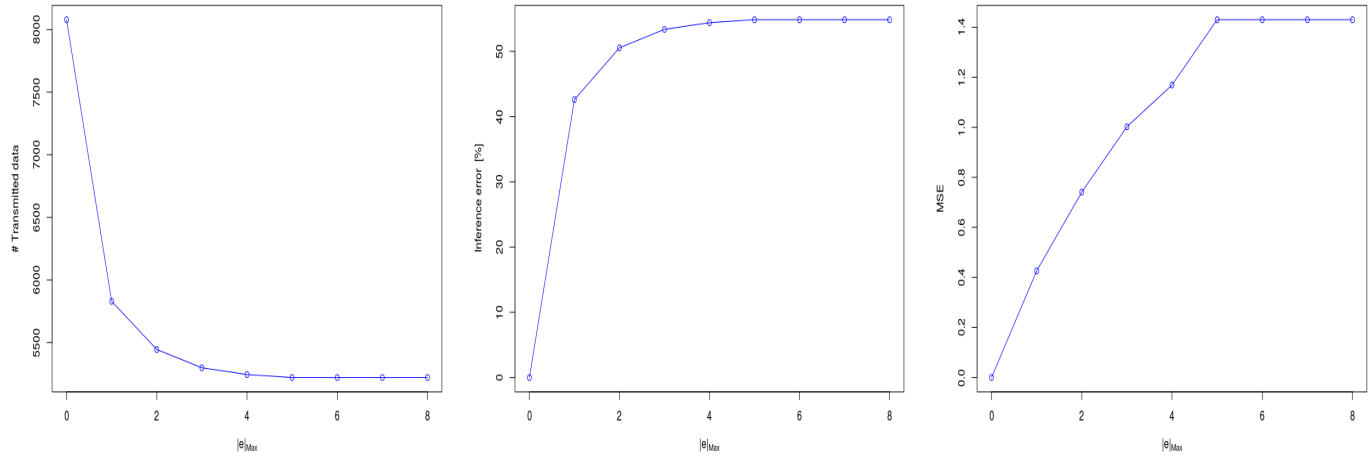
As we stated before, the value of the threshold $|e|_{Max}$ strictly depends on the application context. Its choice has a non-negligible impact on the final results. From Fig. 3, for example, we can say that the more we use a higher threshold, the less we send data but also the more we get an inference error and the more we lose in information quality.

[3] Of course, in Fig. 2 we did not consider the scenario s_1 since it does not use the proposed inference approach.



(a) (b)

Fig. 2: Variation of $|e|$ in scenario (a) s_2, and (b) s_3 versus 2.5 days collection time.



(a) (b) (c)

Fig. 3: Variation of (a) the transmitted data, (b) the estimation error and (c) MSE according the value of the threshold $|e|_{Max}$.

5 Conclusions

In this paper, we have presented an inference-based approach for avoiding transmitting high correlated data in an heterogeneous IoT network. A good correlation between data was taken into account for this study. Indeed, It is important to have a good data correlation to avoid a very high error rate. Through experimentation on FIT IoT-LAB platform using the M3 nodes, we have showed that our Bayesian inference approach is scalable and reduces considerably the number of transmitted data and the energy consumption, while keeping an acceptable level of estimation error and information quality. We have also shown that the use of smart node decreases the inference error.

Acknowledgment

This work was partially supported by a grant from CPER Nord-Pas-de-Calais/FEDER Campus Intelligence Ambiante.

References

1. Cedric Adjih, Emmanuel Baccelli, Eric Fleury, Gaetan Harter, Nathalie Mitton, Thomas Noel, Roger Pissard-Gibollet, Frederic Saint-Marcel, Guillaume Schreiner, Julien Vandaele, et al. Fit iot-lab: A large scale open experimental iot testbed. In *Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on*, pages 459–464. IEEE, 2015.
2. Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
3. Giancarlo Fortino, Antonio Guerrieri, Wilma Russo, and Claudio Savaglio. Integration of agent-based and cloud computing for the smart objects-oriented iot. In *Computer Supported Cooperative Work in Design (CSCWD), Proceedings of the 2014 IEEE 18th International Conference on*, pages 493–498. IEEE, 2014.
4. Zoubin Ghahramani. Graphical models: parameter learning. *Handbook of brain theory and neural networks*, 2:486–490, 2002.
5. W. Hong S. Madden M. Paskin P. Bodik, C. Guestrin and R. Thibaux. Intel lab data. `http://www.select.cs.cmu.edu/data/labapp3/index.html`. Accessed July 20, 2016.
6. Cristanel Razafimandimby, Valeria Loscri, Anna Maria Vegni, and Alessandro Neri. A bayesian and smart gateway based communication for noisy iot scenario. In *International Conference on Computing, Networking and Communications*, 2017.
7. Cristanel Razafimandimby, Valeria Loscri, Anna Maria Vegni, and Alessandro Neri. Efficient Bayesian Communication Approach For Smart Agriculture Applications. In *2017 IEEE 86th Vehicular Technology Conference*, Toronto, Canada, September 2017.
8. Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Computer Vision and Image Understanding*, 117(11):1610–1627, 2013.
9. Thomas Watteyne, Ana Laura Diedrichs, Keoma Brun-Laguna, Javier Emilio Chaar, Diego Dujovne, Juan Carlos Taffernaberry, and Gustavo Mercado. Peach: Predicting frost events in peach orchards using iot technology. *EAI Endorsed Transactions on the Internet of Things*, 2016.
10. Jonathan S Yedidia, William T Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8:236–239, 2003.