An Adaptive Monitoring Service exploiting Data Correlations in Fog Computing

Monica Vitali, Xuesong Peng, and Barbara Pernici

Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, ITALY,
[name].[surname]@polimi.it

Abstract. In smart environments, a big amount of information is generated by sensors and monitoring devices. Moving data from the edge where they are generated to the cloud might introduce delays with the growth of data volume. We propose an adaptive monitoring service, able to dynamically reduce the amount of data moved in a fog environment, exploiting the dependencies among the monitored variables dynamically assessed through correlation analysis. The adaptive monitoring service enables the identification of dependent variables that can be transmitted at a highly reduced rate and the training of prediction models that allow deriving the values of dependent variables from other correlated variables. The approach is demonstrated in a smart city scenario.

1 Introduction

Monitoring Data are generated by sensors used to monitor an environment of interest, that are intended to be utilized by different applications deployed across edge/IoT, fog, and cloud layers. In a smart city, data collected from scattered, different places, converge into a unified monitoring data service used by different applications. The volume of data collected by IoT and sensors makes it time consuming to move all data from the edge where they are generated to the cloud for analysis, likely introducing critical delays. It is important to reduce the size of the data to be moved in order to make this task more agile [1]. As described in Section 2, existing approaches focus on the definition of possibly adaptive sampling rates for each variable. As illustrated in [2, 3], variables collected by a monitoring system may be not independent. In this paper, we propose a service oriented approach to reduce the data volume by exploiting hidden relations among data, distinguishing between regressor variables and dependent variables, for which it is possible to significantly reduce the volume of transmitted data. The paper is organized as follows. Sect. 2 analyzes the state of the art. Sect. 3 describes the overall approach and Sect. 4 illustrates the monitoring reduction service. In Sect. 5 we apply the framework to a smart city scenario.

2 State of the Art

Data reduction in Big Data systems generally reduces either data storage, in-network data transmissions, or data redundancy [4]. The reduction methods

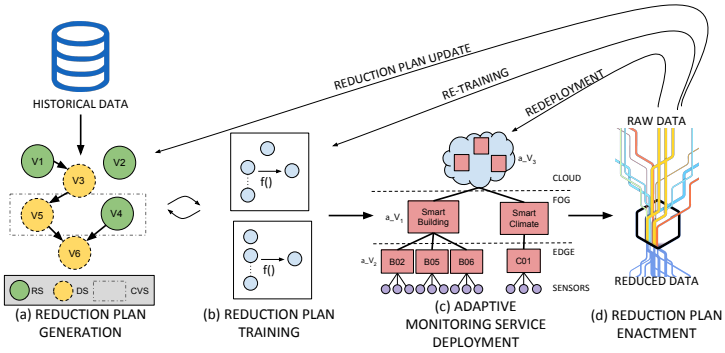

Fig. 1. Adaptive Monitoring Service

include compressing raw data, decreasing the data sampling rate, and reducing the overall data according to the network topology. The work of [5] proposes a knowledge-driven data sharing framework in IoT-based Big Data systems, transmitting knowledge patterns instead of raw data. In our approach we propose to dynamically derive the relationships among variables, based on the actual data, rather than on predefined knowledge patterns. The problem of adaptive monitoring has been widely discussed in the literature [6]. The authors review monitoring tools and techniques for Fog Computing, and consider the reduction in the amount of network traffic as one of the challenges in current monitoring systems. A solution is proposed in [7], where a lightweight adaptive monitoring framework suitable for IoT devices is proposed. The authors reduce the data volume considering an adaptive sampling and (ii) an adaptive filtering. Similarly to [8], the focus is on the adjustment and reduction of single signals generated by sensors, without considering the possible dependencies existing between them.

Smart cities and smart homes are typical applications of the fog computing technology. In [9, 10], architectures for optimizing near real-time services for prediction analytics are discussed. In order to show the performance of our approach, we applied the AMS to a real dataset representing a smart city scenario.

3 Adaptive Monitoring Service

The proposed Adaptive Monitoring Service (Fig. 1) has the goal to reduce monitoring data deriving and exploiting correlations among monitored variables (i.e., sensor-generated data). The first step is the *Reduction Plan Generation (a)*, where historical data are analyzed to discover relations among variables and to generate a Reduction Plan, which indicates the variables that must be collected (regressor variables - RV) and the ones (dependent variables - DV) that can be reconstructed from the values of the collected ones. Each variable can therefore be monitored in three different modalities: (i) $EMPTY$ means no data are transmitted; (ii) $NORMAL$ means all the produced data are transmitted; and (iii) $REDUCED$ means a highly reduced down-sampled set of data is transmitted,

used only for validation purposes. To derive the prediction models, i.e. functions to reconstruct dependent variables from regressor variables, a *Reduction Plan Training (b)* phase is performed. An *Adaptive Monitoring Service Deployment (c)* phase decides how to deploy the services that are needed to enact the Reduction Plan on a fog hierarchical infrastructure, trying to reduce the overall volume of data traveling from the edge of the network to the cloud. The optimization of this step will be analysed in future work. The *Reduction Plan Enactment (d)* transforms the raw data produced by sensors into a reduced dataset. Monitoring can be enhanced considering the variability of the environment[8]. During the collection of reduced data, several events might occur that require adjustments. Therefore, the AMS execution requires a continuous validation phase, where data of dependent variables are collected at a highly reduced frequency only to verify the validity of the prediction functions. Minor events might require a refinement of the prediction models (re-training). However, since the observed environment is dynamic and relations among variables might change, the derivation of a new Reduction Plan might also be needed in some cases.

4 Monitoring Data Reduction

Dependencies between variables are represented through a Direct Acyclic Graph (DAG), derived from the *correlation matrix*, i.e., the matrix obtained by computing the correlation between each pair of variables collected by the monitoring system, orienting the edges by discovering causal relations, using the technique described in [2]. The approach discovers relations between variables by computing the Pearson correlation coefficient between each couple of variables, applying a threshold to filter weak correlations, and deriving causal dependencies through a heuristic search algorithm. Once dependencies are detected, prediction can be provided by building a regression formula able to properly combine all the concurring variables to reconstruct a missing signal. In this way, some of the data produced by sensors can be omitted and reconstructed after the transmission if needed, thus reducing the volume of the data to move.

Before going into the details, we introduce some terminology. We denote the set of all monitoring variables as U, which is split into two sets: (i) *Regressor Variable Set* (RS), composed of all the variables that cannot be derived from other information (independent variables); (ii) *Dependent Variable Set* (DS), composed of variables derivable from other monitored information. According to this, $U = RS \bigcup DS$. Each variable in DS depends on the value of other variables - referred to as *Correlated Variables Set* (CVS) - and can be reconstructed using a **regression function**. Variables in CVS can be both regressor variables $rv \in RS$ and dependent variables $dv \in DS$. In Fig. 1(a) we show an example with six variables and their dependencies. Variable v_6 is depending both on v_4 and v_5, therefore the CVS for v_6 is $\{v_4, v_5\}$. In the figure, we also see that v_4 is a regressor variable while $v5$ is a dependent variable, depending on $v3$. So in this case we have $RS = \{v_1, v_2, v_4\}$ and $DS = \{v_3, v_5, v_6\}$. The correlated variable sets are $cvs(v_3) = \{v1\}$, $cvs(v_5) = \{v_3\}$, $cvs(v_6) = \{v_4, v_5\}$.

As described in [2], the dependencies between monitoring variables are not static (e.g., a sensor might stop working for a period of time, an existing sensor might be moved from a location to another, a new sensor might be installed). The CVS used to predict a variable in DS can change accordingly. Thus, we model an element of CVS as a variable $cvs_{t,k} \subseteq U$ dependent on timestamp t. In the reduced monitoring data, variables in RS keep all raw samples, since they cannot be derived from other variables. Variables in DS, instead, are collected at a reduced sampling rate. Samples are used for continuously validating the reliability of the prediction.

The **Reduction Plan** is the key element of the AMS and the basis for the service to enact data reduction. It gives information on which variables to reduce and on how to reconstruct their value from their correlate variables. It consists of the following parts: (i) *RS/DS partition*: the set of variables U is partitioned into the two subsets RS and DS. The partition at timestamp t is denoted as $< RS_t, DS_t >$; (ii) *CVS*: for each variable in DS the set of correlated variables $cvs_{t,k}$ is used to train the prediction function of $dv_k \in DS$ at time t; (iii) *Prediction parameters*: the prediction parameters describe the quantitative relation between a variable dv_k and its correlated variables $cvs_{t,k}$. A reduction plan is represented as a labeled-DAG (LDAG), sub-graph of the DAG of the dependencies. An edge from v_i to v_j indicates that v_j is reduced and rebuilt starting from the values of v_i. Since the reduction plan can evolve, we denote the reduction plan used at time t as $LDAG_t$ (Eq. 1):

$$LDAG_t = [Nodes_t, Edges_t, Labels_t] \tag{1}$$

Given the Reduction Plan, for each dv_k, the service provides the parameters of the model for enacting the prediction. To capture the correlations between variables collected by the monitoring system, a regression analysis is performed on a training dataset. In this paper we have applied Linear Regression as the regression method, due to its low complexity and reduced execution time given the need to build the model on edge and fog devices with limited resources and to quickly rebuild the model when needed. We assume the CVS of $v_k \in DS$ contains N variables $X = \{x_1, x_2, \cdots, x_N\}$ and the training dataset comprises samples of P timestamps. The linear regression method assumes that the relationships between X and $f_{t,k}$ are linear, as depicted in Eq. 2 at timestamp t:

$$f_{t,k}(X) = \beta_0 1 + \beta_1 x_{t,1} + \cdots + \beta_n x_{t,N} + \epsilon_t \tag{2}$$

In this work, we adopt the Ordinary Least Squares (OLS) method [11] to estimate the parameters values β, as described in [12]. This approach is only used as a proof of concept and alternative methods can be adopted. As an example, we are also investigating alternative solutions such as neural networks.

5 Validation

We applied the AMS to the REFIT Smart Home dataset[1], which includes sensor measurements of smart buildings and climate data recorded at a nearby weather

[1] https://lboro.figshare.com/articles/REFIT_Smart_Home_dataset/2070091

Table 1. Reduction performance scoring of B05 and B06 variables

DV	CVS	corr.
B06_HW1_Temp	B06_S1_Temp	0.88
B05_LR1_Temp	B05_K1_Temp	0.86
B06_BR1_Temp	B06_K1_Temp	0.85
B05_BR2_Temp	B05_S1_Temp	0.89
B05_BR3_Temp	B05_BR2_Temp B05_S2_Temp	0.96
B05_BR1_Temp	B05_S1_Temp	0.85
B06_BT2_Temp	B06_BR3_Temp	0.89
B06_BT3_Temp	B06_LR1_Temp B06_BR1_Temp	0.90
B06_LD1_Temp	B06_BT3_Temp	0.80
B05_BR4_Temp	B05_BR1_Temp B05_BR3_Temp	0.98
B06_K1_Temp	B06_LR1_Temp B06_S1_Temp	0.88

station. Each building is connected to an edge device, collecting the information before sending them to be stored in the cloud.

We used the data collected in 80 days, from 2014-02-05 to 2014-05-05 at a fixed sampling interval of 30 minutes. We used 14 days of data to train the Reduction Plan, then we tested the performance of data reduction with the data left (59 days). Applying the proposed methodology, we found 31 regressor variables rv and 43 dependent variables dv to be predicted. For 17 of these dvs, the AMS reduces the sensor data of more than 40% while maintaining a reasonable accuracy. Tab. 1 shows a subset of the selected reductions, focusing on buildings $B05$ and $B06$. The first column represents the DS discovered while column 2 represents the CVS for each dv. The correlation value of the relation is shown in column 3. As it can be observed, strong relations are discovered between variables of the same kind in the same building, and most of all between temperatures of different rooms. The reduction ratio for the whole dataset of 74 variables in 59 days is 15.95%. This is a good achievement considering also that 31 variables are not reduced and that a portion of the 43 dvs are collected as raw data during the validation and re-training phases. The average reduction ratio considering only the 43 dvs is 27%.

6 Concluding remarks

The Adaptive Monitoring Service proposed in this paper aims to identify a new systematic reduction of sensor data transmitted in a fog architecture. The relationships among the variables are exploited to reduce the data flow between the layers of a fog environment. The implications on service deployment have been discussed and an example based on a smart city scenario has been presented.

In future work we are going to refine the proposed methodology by focusing on the service deployment. We aim to propose an optimised deployment strategy considering the heterogeneity of the monitoring services and of the nodes in

which they can be executed. We will also introduce latency for evaluating the effectiveness of the reduction plans when dealing with high data volumes.

Acknowledgments

This work is supported by European Commission H2020 Programme through the DITAS (Data-intensive applications Improvement by moving daTA and computation in mixed cloud/fog environmentS) Project no. 731945.

References

1. P. Plebani, D. García-Pérez, M. Anderson, D. Bermbach, C. Cappiello, R. I. Kat, F. Pallas, B. Pernici, S. Tai, and M. Vitali, "Information logistics and fog computing: The DITAS approach," in *Proceedings of CAiSE Forum 2017, Essen, Germany, June 12-16, 2017*, 2017, pp. 129–136.
2. M. Vitali, B. Pernici, and U.-M. O'Reilly, "Learning a goal-oriented model for energy efficient adaptive applications in data centers," *Information Sciences*, vol. 319, pp. 152–170, 2015.
3. C. G. Carvalho, D. G. Gomes, N. Agoulmine, and J. N. de Souza, "Improving prediction accuracy for WSN data reduction by applying multivariate spatio-temporal correlation," *Sensors*, vol. 11, no. 11, pp. 10 010–10 037, 2011.
4. M. H. U. Rehman, C. S. Liew, A. Abbas, P. P. Jayaraman, T. Y. Wah, and S. U. Khan, "Big data reduction methods: A survey," *Data Science and Engineering*, vol. 1, no. 4, pp. 265–284, 2016.
5. M. H. U. Rehman, V. Chang, A. Batool, and T. Y. Wah, "Big data reduction framework for value creation in sustainable enterprises," *Int J. Information Management*, vol. 36, no. 6, pp. 917–928, 2016.
6. S. Taherizadeh, A. C. Jones, I. Taylor, Z. Zhao, and V. Stankovski, "Monitoring self-adaptive applications within edge computing frameworks: A state-of-the-art review," *Journal of Systems and Software*, vol. 136, pp. 19–38, 2018.
7. D. Trihinas, G. Pallis, and M. Dikaiakos, "Low-cost adaptive monitoring techniques for the internet of things," *IEEE Transactions on Services Computing*, 2018.
8. M. Andreolini, M. Colajanni, M. Pietri, and S. Tosi, "Adaptive, scalable and reliable monitoring of big data on clouds," *J. Parallel Distrib. Comput.*, vol. 79-80, pp. 67–79, 2015. [Online]. Available: https://doi.org/10.1016/j.jpdc.2014.08.007
9. A. Yassine, S. Singh, M. S. Hossain, and G. Muhammad, "IoT big data analytics for smart homes with fog and cloud computing," *Future Generation Computer Systems*, vol. 91, pp. 563–573, 2019.
10. M. Aazam, S. Zeadally, and K. A. Harras, "Fog computing architecture, evaluation, and future research directions," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 46–52, 2018.
11. F. Hayashi, *Econometrics*. Princeton Univ. Press, 2000.
12. X. Peng and B. Pernici, "Correlation-model-based reduction of monitoring data in data centers," in *SMARTGREENS 2016 - Proceedings of the 5th International Conference on Smart Cities and Green ICT Systems, Rome, Italy, April 23-25, 2016.*, 2016, pp. 395–405.