Data Reduction in Monitored Data

Xuesong Peng[1,*]

Supervisor: Barbara Pernici[1]

[1]Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milan, Italy
xuesong.peng@polimi.it

Abstract. Nowadays, increasing time series data has brought new challenges in many domains because their massive instances, dimensions, speed and complexity. In order to solve this problem, data reduction techniques are becoming an integral part of future systems, especially for Monitoring System. Different from most data reduction techniques which are based on information theory, this paper is planning to explore a novel model-based method, which tries to build piecewise regression models only for correlated data with guidance of priori knowledge, avoiding unnecessary computation between unrelated data streams. Until now, we have designed a primal data reduction framework, corresponding experiments and evaluation criteria.

Keywords: data reduction; time series; model-based; monitored data.

1 Introduction

In the era of Big Data, numerous new data streams are generated every day, resulting extremely high-volume data from multiple sources in many domains, such as electricity grid, data center, smart building, etc. Those data streams are mainly based on sensing systems of physical environments and computing systems, i.e. Monitoring Systems, and most of Monitored Data can be described as time-series data, which is the representation of a collection of values obtained from sequential measurements over time [1].

Modern information systems have to deal with a large amount of information, which always lies in massive data and requires many times of transmission, storing and extraction of data. Moreover, some of them are in continuous data stream, which make these problems even harder to solve. This cost-expensive process now becomes a bottleneck of modern information system, so how to reduce the quantity of stored data while maintaining the ability to derive informative values becomes a key problem to research. My Ph.D. research will be carried on the phenomena of data reduction process in monitoring system, especial streaming data, with two key properties of

interest, namely, performance of process and reserved information values. I will mainly focus on improving data reduction process by exploiting powers of knowledge models, which is a general term of models built on learnt knowledge. They can be simple correlation models describing correlations between data dimensions, as we used in CMBDR framework currently, or even logic-based models like some semantic models which describing complicate relations such as generalization, classification, etc. During my research, I will try to answer the following research question: Is it possible for knowledge models to help information systems improve their ability to do data reduction and information values retrieval? Currently, I'm assuming models are based on a priori knowledge, but it could also be an optional direction to investigate how to build models of monitored data. In this paper, some related algorithms and tools are studied, and based on their drawbacks, the idea of model-based data reduction is proposed. And in order to evaluate this idea, a possible initial approach named CMBDR (a novel model-based framework building recursive regression models to obtain a reduced representation of raw data) is designed and will be validated in future. And obviously, new algorithms of data reduction should be developed and combined with existing ones to implement this idea, which would be my future work.

The rest of paper is organized as follows. In section 2, we discuss the state of the art, explaining some common used data reduction techniques and their disadvantages. In section 3, as a candidate solution to the problem, we present CMBDR framework and mechanisms to implement. In section 4, evaluation mechanisms and criteria are defined to measure framework performance. Section 5 lists some interesting open issues related to CMBDR. We conclude the paper and introduce future work in section 6.

2 State of the art

A variety of techniques have been developed for data reduction in time series in the past few decades, and most of them are from the perspective of information theory. PCA (Principal Component Analysis) implements orthogonal transformation to convert possibly correlated variables to some linearly uncorrelated variables (principal components) based on a few observations [2,3]. PAA (Piecewise Aggregate Approximation) [4] is a simple dimensionality reduction method for time series, which reduce dimensionality with mean values of equal sized frames of original data. Some other common used reduction techniques are SVD (Single Value Decomposition), DFT (Discrete Fourier Transform) and DWT (Discrete Wavelet Transform).

Recently, some new methods are also introduced to implement data reduction in large systems. Cypress [5] is a new method supporting archiving and querying for massive time series data, it firstly transforms the single data stream into several substreams (called trickles) and then directly use some trickles to answer common queries (trends, histograms, and correlations), which doesn't need to reconstruct original data. Those trickles are generated by implementing filtering, down sampling, thresholding and random projection on original time series stream. Cypress framework carries out lossy compression to achieve high compression ratio, and at the same time it

remains to be effective to answer queries (reserving spikes, trends of original data) on compressed data, without reconstruction of original data.

YADING [6] is an end-to-end clustering algorithm which works on large-scale time series with fast performance and quality results employing some data reduction techniques. YADING provides theoretical proof on the lower and upper bounds of the size of the reduced dataset firstly, and then operates random sampling and dimensionality reduction (PAA) on original datasets.

In general, data reduction methods are all built based on data redundancy. In time series case, most of data redundancy lies in temporal repeatability and correlation between each series. And those correlations can be ubiquitous in variant systems in terms of monitored data, in which remarkably, data correlations are always resulted from real-world relations between monitored objects, including spatial relation and also logical relation. For instance, correlation between neighboring sensors placed in Smart Buildings, and collaborating servers in a data center.

As explained above, current data reduction methods are all based on information theory, which means before data reduction phase, they need to analyze data first to understand the repeatability and correlations. But all these methods neglect the semantic meanings of data in their process, causing much aimless computation which lower program performance. And that's also the biggest barriers hinder their implementation in big systems. So this paper goes in the opposite direction, by exploiting properties which can be apparent or easily inducted from existing data and knowledge to avoid aimless computation between uncorrelated data, proposes a new method to do data reduction in monitored data which exploit correlations known from predefined models.

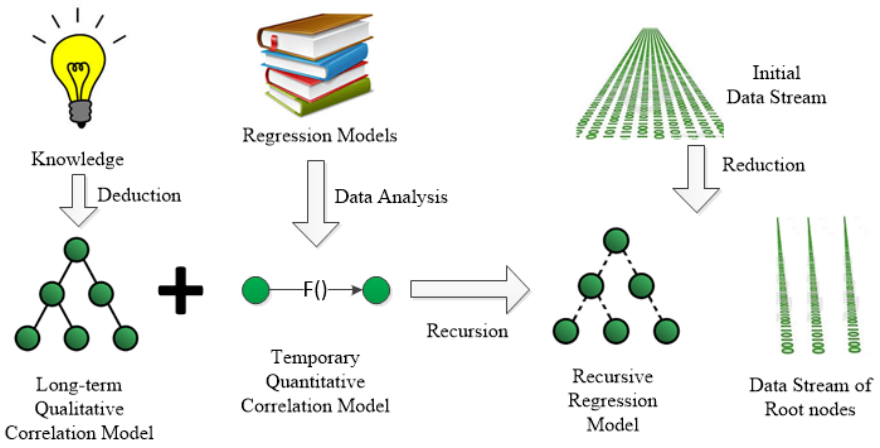

Fig. 1. CMBDR framework

3 CMBDR Framework

As an initial approach, CMBDR exploits correlation models because they can always give direct guidance on reduction process and can be easily obtained in many cases. Learning from the idea of replacing correlated time series data with regression models [7], we propose a new framework for time series data reduction in monitored data, namely, Correlation-Model Based Data Reduction (CMBDR), as shown in **Fig. 1**. It exploits the power of a priori knowledge to find possible correlations between data streams, from which we can build several Correlation Models to describe the qualitative relations between data streams in long term. Then for data output in a short period (called compression window), multiple regression models can be applied to find a matching quantitative temporary relation (function F), so that data is compressed because any values of one data stream can be predicted by the other. Moreover, based on recursion of these quantitative relations on correlation model, the recursive regression model could be built so that data of any nodes on the correlation model could be predicted with raw data stream of the root nodes. In this way, CMBDR saves time by only conducting analysis on correlated data whose relations are specified in the input correlation model.

Compared to other solutions, two highlights of CMBDR lie in the process-guiding correlation model and also recursion of regression which makes scalability possible. In this paper, CMBDR can exploit not only the rigorous correlation models which are derived from empirical data research (e.g. Bayesian Network of indicator correlations [8]), but also some simple models deduced from general experience (e.g. correlation between light, temperature and CO2 level in a greenhouse). In addition, not every arbitrary topology is qualified for the correlation model, the minimum requirement is Directed Acyclic Graph shown in **Fig. 2**, which specifies unambiguous correlation and provides only one recursion choice. When all nodes in the model rely on at most one other node, the model will become a tree-like model shown in **Fig. 1**. Similar to the definition of depth for nodes in a tree, we defined depth for nodes in the Directed Acyclic Graph as the maximum number of edges from the node to the graph's root nodes. Another thing needs to remind is, the so-called long-term correlation model is not permanent.

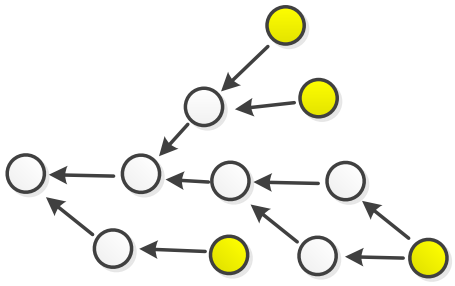

Fig. 2. Directed Acyclic Graph and its root nodes

Currently, CMBDR only exploits segmented dimensionality reduction method (piecewise regression) which assumed the correlation model is true, so the performance is highly dependent on the quality of input model. Although the compression ratio of CMBDR is probably less than some techniques above because of their exhaustive information theory analysis, the pre-defined model can make this framework avoid much more unnecessary computations.

In order to implement CMBDR and evaluate its performance, we propose the solution divided into several steps, listed as follows:

1. Initialize the input correlation model, testing dataset (multi-variable data streams)
2. Initialize parameters, namely, length of compression window, regression threshold (compared with matching degree to validate a regression model), etc.
3. Segment time series data stream into compression windows
4. In a compression window, carry out regression analysis for each pair of streams which has relations in the Correlation Model (in a Width-First-Search order)
5. If the best fitting regression model exceeds the regression threshold, compress data according to this regression model; otherwise, keep raw data
6. Go to step 3 until no more compression windows
7. Output regression model parameters and raw data of root nodes
8. Recover data with regression models and raw data of root nodes (in a Width-First-Search order)
9. Evaluate results

Since CMBDR framework is aimed to quickly find fitting regression models to quantify relationship between two variables which are in a stream manner, the computation complexity of regression analysis must be limited. So in this paper, only the following simple regression models are considered and the framework should give preference to the model with the least time complexity during regression analysis.

- Simple linear regression with time complexity of $O(m)$
- Multiple linear regression with time complexity of $O(m^2 n)$
- Low polynomial regression

4 Evaluation Criteria

CMBDR framework conducts model-based regression analysis on monitored data to achieve data reduction with raw data of root nodes and the recursive regression model. In order to evaluate CMBDR, validation will be conducted with abundant experiments on specific monitored datasets:

- sensors outputs of ventilation facades in a smart building
- software and hardware information of a data center monitoring system

Aimed to find its ability to achieve less execution time, higher compression ratio, and better accuracy compared to some mainstream techniques (PAA, PCA, Rainmon [9]), following evaluation metrics are built:

- Compression ratio to demonstrate compressing performance of data reduction
- Total execution time to show processing speed on data streams
- Accuracy to reflect the informative values remained in compressed data

In specific to the accuracy aspects, instead of one specific criterion, we propose multiple criteria to meet different possible requirements from different applications, the criteria are as following:

- Root-mean-square deviation to measure total accuracy of compression, it represents standard deviation of the differences between raw data y and compressed data f, as shown in equation (1)

$$\mathrm{RMSD} = \sqrt{E((f-y)^2)} \tag{1}$$

- Coefficient of determination, denoted as R^2 to measure total goodness of fit of the compressed model, as shown in equation (2), y_i is raw value of original data and f_i is recovery value of compressed data for time series containing n values

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

- Relative error to weigh informative values decreased by the data reduction method, as shown in equation (3), η is relative error, η is absolute error and υ is range of data

$$\eta = \frac{\varepsilon}{|\upsilon|} \tag{3}$$

- Pearson's Coefficient to measure correlation between data streams so that we can compare and find correlation variations after data reduction, as shown in equation (4), γ is Pearson's Coefficient, X_i and Y_i are sample values for time series X and Y containing n values, $\bar{X}$ and $\bar{Y}$ are mean values of X and Y

$$\gamma = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{4}$$

- Performance of some applications running on both raw data and compressed data (e.g. clustering) to validate practicability of CMBDR, and also as important indicators of reserved informative values in compressed data

5 Open issues

CMBDR is a new data stream reduction framework based on correlation models which is from knowledge. Although this paper has discussed some important aspects of this framework, there are still many open issues not covered, and solving these issues would be very helpful to enhance CMBDR. So in this part, we will talk about three crucial open issues, namely, regression abnormal, error prediction and error control.

Regression abnormal implies the moments when there is no regression model matched for the time series data in a compression window in regression analysis. This is mainly because the simple and limited regression models are not always capable of finding a well-fitting regression model, especially when the data vary exponentially. Currently, CMBDR exploits regression threshold to detect these abnormal, and once an abnormal is detected, the framework will directly use raw data. Some extensions could be researched to compress these abnormal windows, such as abnormal behavior detection based on machine learning algorithms.

Error prediction and error control paradigm should also be investigated to help to improve CMBDR performance. Firstly, based on the historical data of errors, confidence could be attached to the edges in the long-term correlation model, making it possible to maintain relations in a dynamic way, so being able to control unnecessary analysis on outdated relations. Moreover, consider the recursive regression model in **Fig. 1**, it is obvious that errors of a child node will be probably larger than his father's in the compressed data. And the deeper a node is, the more information loss it will have. So analysis could be carried out to find relations between errors of nodes on the recursive regression model, upon which we can build error prediction model. In the meanwhile, some error control methods could also be exploited to improve accuracy for those deep nodes. For instance, use raw data for some nodes in the regression model, so that errors of father nodes will not be inherited by their children. Alternatively, some other reduction techniques could also be exploited here, such as down sampling, PCA etc.

6 Conclusion and future work

As an initial approach to solve the monitored data reduction problem, we have presented CMBDR, a framework operating massive multivariate data stream, based on a correlation model. It applies regression analysis to quantify correlations between data streams and exploits recursion on correlation model to achieve reduced representations of raw data. We designed evaluation criteria to measure performance, and also discussed some crucial open issues which can improve CMBDR performance.

Future developments of this paper are to implement CMBDR framework in two scenarios, the first one is ambient environment monitoring with multiple sensors (temperature, humidity, air speed, etc.), and the second one is a data center monitoring system which collects various software and hardware information. Also, in order to extend CMBDR, the paradigm of error prediction and control discussed in section 5 will be researched, mainly focusing on how modeling of information and learning techniques could support each other. Furthermore, collaborations of CMBDR and other data reduction techniques are worthy of investigation.

References

1. Esling, Philippe, and Carlos Agon. "Time-series data mining." ACM Computing Surveys (CSUR) 45.1 (2012): 12.

2. Jolliffe, Ian. Principal component analysis. John Wiley & Sons, Ltd, 2002.
3. Wikipedia contributors. "Principal component analysis." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 21 Mar. 2015. Web. 30 Mar. 2015.
4. Keogh, Eamonn, et al. "Dimensionality reduction for fast similarity search in large time series databases." Knowledge and information Systems 3.3 (2001): 263-286.
5. Reeves, Galen, et al. "Managing massive time series streams with multi-scale compressed trickles." Proceedings of the VLDB Endowment 2.1 (2009): 97-108.
6. Ding, Rui, et al. "YADING: Fast Clustering of Large-Scale Time Series Data." Proceedings of the VLDB Endowment 8.5 (2015).
7. Carvalho, Carlos, et al. "Improving prediction accuracy for WSN data reduction by applying multivariate spatio-temporal correlation." Sensors 11.11 (2011): 10010-10037.
8. Vitali, Monica, Barbara Pernici, and Una-May O'Reilly. "Learning a goal-oriented model for energy efficient adaptive applications in data centers." Information Sciences (2015).
9. Shafer, Ilari, et al. "Rainmon: an integrated approach to mining bursty timeseries monitoring data." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.