ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 6, May 2018

2nd National Conference on "RECENT TRENDS IN COMPUTER SCIENCE & INFORMATION TECHNOLOGY (RTCSIT'18)"

13th-14th March 2018

Organized by

Departments of Computer Science & Information Science Engineering, Sri Krishna Institute of Technology, Bangalore, India

Prediction of Effective Missing Data over Multivariable Time Series on Apache Spark

Raj Solanki[1], Surbhi Pandey[2], Prof. Lokesh H D[3]

U.G Student, Department of Computer Science and Engineering, Sri Krishna Institute of Technology, Bengaluru, India[1]

U.G Student, Department of Computer Science and Engineering, Sri Krishna Institute of Technology, Bengaluru, India[2]

Assistant Professor, Department of Computer Science and Engineering, Sri Krishna Institute of Technology, Bengaluru, India[3]

ABSTRACT: Huge volume of data are generated in many areas the missing of some values in collected data always occurs in practice and challenge extracting maximal value from the large scaled a test we have taken up the challenge of missing data prediction in multivariable time series by employing improved matrix factorization techniques .Approaches are optimally designed to largely utilize both the internal patterns of each time series and the information of time series across multiple sources . Based on the idea, we have imposed three different regularization terms to constrain the objective functions of matrix factorization and built five corresponding models.

KEYWORDS: matrix factorization, missing data prediction, time series, Big Data, Apache Spark.

I. INTRODUCTION

The smoothness of instruments to collect data from multiple sources and the resulting volume of data have grown to an unprecedented level since the era of big data started in recent years .Multivariable time series, one common data format, are ubiquitous in many real-world applications, like electric equipment monitoring, weather or economic forecasting, environment state monitoring, security surveillance and many more. In most applications multiple sensors are employed to generate time series data, and they usually share one common goal. For example, in the power grid system, various diagnostic gases sensors are deployed to monitor the status of the main power transformers and generate multivariable time series by measuring the content of the diagnostic gases over time .In the "Internet of Things", a large number of sensors are used to produce multivariable time series of the external environment, e.g., the air or water quality. In the medical and healthcare systems, multiple sensors can also be equipped within living spaces to monitor the health and general well-being of senior citizens, while also ensuring that proper treatment is being administered and assisting people regaining lost mobility via therapy as well . In this paper, the sensors sharing one common goal are treated as a sensor network. The inevitable data missing necessitates integrated analysis of observed data sets. A large collection of data mining and statistical methods have been proposed to predict the missing values of time series. One simplest solution might be linear interpolation. But it is only feasible to be applied to the case where only a low ratio of collected data are missing and the time series vary very steadily modelling methods, more commonly used solutions, try to discover the underlying patterns and seasonality to predict the missing values using some common sense . Representative modelling approaches include deterministic models, stochastic models, and state space models . For example, Frasconi et al. employed a seasonal kernel to measure the similarity between time-series instances and proposed the seasonal autoregressive integrated moving average model coupled with a Kalman filter achieved excellent performance of missing data prediction .Baraldi et al. proposed a fuzzy method for missing data reconstruction, which involves fuzzy similarity measurement and shows superiority to an auto associative kernel regression method. Besides, Song et al. used matrix factorization to predict traffic matrices and their method showed more effective performance than traditional methods. In this paper, aiming at solving the above problems, we propose to fuse the temporal smoothness of time series and the information across multiple sources into matrix factorization in order to improve the accuracy of missing data prediction in multivariable time series. First, as each time series rarely fluctuate wildly over

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 6, May 2018

2nd National Conference on "RECENT TRENDS IN COMPUTER SCIENCE & INFORMATION TECHNOLOGY (RTCSIT'18)"

13th-14th March 2018

Organized by

Departments of Computer Science & Information Science Engineering, Sri Krishna Institute of Technology, Bangalore, India

time, i.e., time series usually host internal and tangible pattern of temporal smoothness. Thus, we try to take advantage of the characteristic to reduce the prediction error of the missing data prediction in multivariable time series. Concretely, a selective penalty term is employed in the matrix factorization objective function to smooth the time series, i.e., we aim at minimizing the fluctuation of each time series with time. Second, as there exists valuable correlation information across multiple sources in a sensor network, we also attempt to fuse that information into matrix factorization to obtain higher performance. Specifically, the correlation information is incorporated in designing two sensor network regularization terms, i.e., the correlated sensors based regularization (CSR) term and the uncorrelated sensors based regularization (USR) term, to constrain the matrix factorization objective function. We take aim at minimizing the difference between a sensor and its correlated sensors or maximizing the difference between a sensor and its uncorrelated sensors based on the sensor network regularization.

- MFS: Matrix Factorization with Smoothness constraints;
- CSM: Correlated Sensors based Matrix factorization;
- USM: Uncorrelated Sensors based Matrix factorization;
- CSMS: Correlated Sensors based Matrix factorization with Smoothness constraints;
- USMS: Uncorrelated Sensors based Matrix factorization with Smoothness constraints;

II. LITERATURE SURVEY

The algorithms which create causality trees based on temporary and dimensional information of identified congestions. Frequent substructures of these causality trees reveal not only recurring interactions among spatiotemporal congestions, but potential bottlenecks or flaws in the design of existing traffic networks a frequent Sub tree algorithm, inspired by association rule mining, which generates the most frequent sub-structure (sub tree) from all discovered congestion trees. These frequent sub trees reveal recurrent congestion trees in the data and suggest inherent problems in existing road networks. The model can be used for inference and joint distribution estimation with both completed and hidden data. In our method, DBN model is used to support the STC by providing the congestion propagation distribution of a frequent sub tree. For any given sub-structure with congestion .at the root segment, the DBN's joint distribution estimation is solely depended on the observations of the involved segments rather than requiring data from the whole large network [1].

The multiple time series data is often accompanied by some contextual information in the form of networks. Idea is to find the latent factor delegation of the input time series and that of its embedded network simultaneously. In this they propose a dynamic contextual matrix factorization algorithm to recover the missing values in a network of time series; and analyse its effectiveness, complexity and the relationship with the existing tools on mining multiple time series. a dynamic contextual matrix factorization algorithm to simultaneously find the common latent factor in both the input time series and its encapsulated network.

(a)This Model infers the missing values directly from the observed values the correlation among different time series; (b) the contextual information encoded by the contextual network; and (c) the temporal smoothness along the time dimension [2].

(1)Explain how social network information can benefit recommender systems; (2) We interpret the differences between social-based recommender systems and trust-aware recommender systems; (3) We coin the term Social Regularization to show the social constraints on recommender systems, and we systematically illustrate how to design a matrix factorization objective function with social regularization; and (4) The proposed method is quite general, which can be easily extended to incorporate other contextual information, like social tags, etc. The empirical analysis on two large datasets demonstrates that our approaches outperform other state-of-the-art methods.

They will also incorporate users' social network information to improve recommender systems scenario includes two central elements of social recommendation: the friends network and the favours of these friends, which can essentially be modelled by the examples of social. Network graph An efficient and effective approach to recommender systems is to factorize the user-item rating matrix, and utilize the factorized user-specific and item-specific matrices to make further missing data prediction [3].

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 6, May 2018

2nd National Conference on "RECENT TRENDS IN COMPUTER SCIENCE & INFORMATION TECHNOLOGY (RTCSIT'18)"

13th-14th March 2018

Organized by

Departments of Computer Science & Information Science Engineering, Sri Krishna Institute of Technology, Bangalore, India

III. RELATED WORKING

Existing System

Most of the methods have been proposed on this One simplest solution might be linear interpolation and Representative modelling approaches include deterministic models, stochastic models, and state space models. Frasconi et al. employed a seasonal kernel to measure the similarity between time-series instances and proposed the seasonal autoregressive integrated moving average model coupled with a Kalman filter achieved excellent performance of missing data prediction. Baraldi et al. proposed a fuzzy method for missing data reconstruction, which involves fuzzy similarity measurement and shows superiority to an auto associative kernel regression method.

The method of SARMIA aims at predicting the missing values with underlying seasonality, and thus the method might have trouble modelling data that have no strict internal seasonality. Especially, when the amount of instances becomes too large, the seasonal kernel computation would cost too much time. The common matrix factorization method usually needs to incorporate the internal specific characteristic of the time series data, such as the spatial information, which might restrict the method to one specific application.

Disadvantage of the Existing System:

- The amount when of instances becomes too large, the seasonal kernel computation would cost too much time.
- It is only feasible to be applied to the case where only low ratios of collected data are missing and the time series vary very steadily.

Proposed System

We describe the details of the proposed methods for prediction of missing data in multivariable time series. We begin with the baseline solution for the problem. We elaborate the key idea of the proposed methods. We present how to take advantage of the smoothness characteristic to reduce the error of the missing data prediction in multivariable time series. We also elaborate why and how to utilize the valuable correlation information across multiple sources in time series data collected from one sensor network to improve the prediction performance. We carefully design three regularization terms to constrain the matrix factorization and then build five different models.

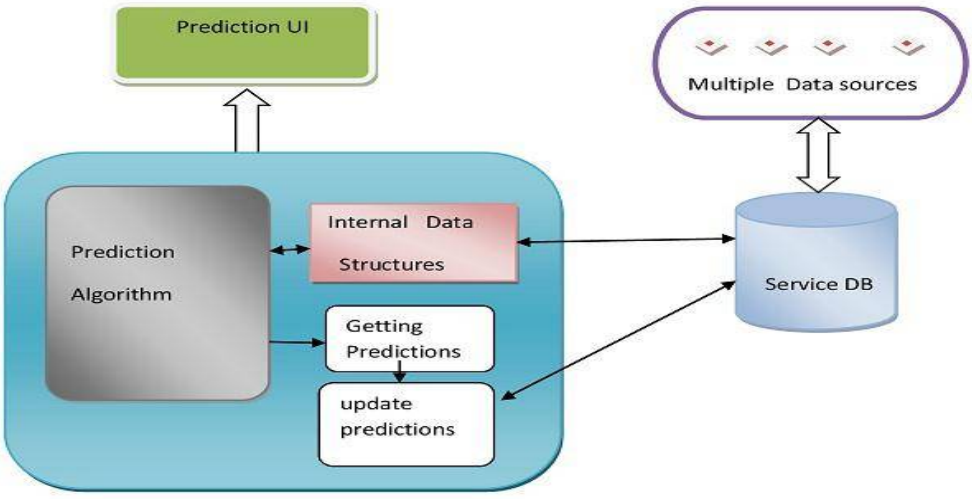

Fig. System Architecture

Experimental Setup

As the time series show special temporal characteristic, we randomly split the data sets. For fair comparison with the baseline methods, the experiments are conducted with the same parameters when we evaluate the performance of the proposed method with different missing ratios. In addition, when we conduct the experiments on one specific

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 7, Special Issue 6, May 2018

2nd National Conference on "RECENT TRENDS IN COMPUTER SCIENCE & INFORMATION TECHNOLOGY (RTCSIT'18)"

13th-14th March 2018

Organized by

Departments of Computer Science & Information Science Engineering, Sri Krishna Institute of Technology, Bangalore, India

parameter, the other parameters remain unchanged and the missing ratio is equal to 0.1.To evaluate all the methods fairly; we incrementally simulate the data missing of the four data sets with an increasing missing ratio. For example, to increase the missing ratio from 0.80 to 0.90, we randomly move 10% of the total data from the observed data set to the missing data set. In this way, the subsequent missing data set always contains the missing data of the previous one. The missing data prediction of testing data sets is based on the known 10% of available values in the training set.

Parallel computing experiments are carried out in a cluster of four working machines based on the same experimental setup given above. As there is no reasonably significance in implementing parallel experiments with small or medium data sets, these experiments are only conducted based on the two large scale data sets, i.e., GSA and SYN data sets. The working nodes are virtual machines (VM) and each of them has two cores with an Intel 2400 CPU and 4G memory. The operating system for the cluster is Cent OS7, while the version of Apache Spark platform is 1.4.0 and the Hadoop platform is version 2.6.0.

- **Data Set Description:** The data sets consist of two small scale data sets and two large scale data sets.
- **Motes Data Set:** The motes data set contains temperature time series collected from 54 sensors deployed in the Intel Berkeley Research lab in about one month. Each time series are collected once every 31 seconds. In the experiment, the length of each time series is 14000.
- **Sea-Surface Temperature Data Set:** The Sea-Surface Temperature (SST) data set consists of hourly temperature measurements from Tropical Atmosphere Ocean Project. In the experiment, the length of each time series is 18000.

IV. CONCLUSION

We have proposed novel methods to constrain the matrix factorization for predicting the missing data in the time series from multiple sources, which achieve satisfactory performance of missing data prediction and high computing efficiency. The methods aim at fusing the smoothness characteristic of each time series and valuable correlation information across multiple sources in a sensor network into matrix factorization. Correspondingly, the methods incorporate smoothness, CSR and USR constraints to optimize the solution of matrix factorization. Based on the idea, we proposed five effective models. The prominent superiority of MFS, CSM and USM reveals the effectiveness of latent factors extraction in the process of matrix factorization after incorporating the constraints. Furthermore, the combination of information ex-traction across multiple sources and temporal smoothness of each time series demonstrate the effectiveness of the proposed methods. Even when the missing ratio is as high as 90%, the RMSE of the proposed methods is still within reasonable range. Finally, the experiments under parallel environment reveal that the USMS model can be executed effectively. We conclude that the proposed methods are alternative models for predicting the missing values in large scale multivariable time series.

REFRENCES

[1].Y. N. Dauphin and Y. Bengio, ``Big neural networks waste capacity," Dept. d'informatique Recherche Opérationnelle, Univ. MontréalMontréal, QC, Canada, Tech. Rep., Mar. 2013. [Online]. Available: http://arxiv.org/pdf/1301.3583.pdf
[2].A. d'Avila Garcez et al., ``Neural-symbolic learning and reasoning:Contributions and challenges," in Proc. AAAI Spring Symp. Knowl.Represent. Reason., Integr. Symbolic Neural Approaches, Mar. 2015
[3].A. Graves, G.Wayne, and I. Danihelka, ``Neural turing machines," GoogleDeepMind, London, U.K., Tech. Rep., Dec. 2014. [Online]. Available:http://arxiv.org/pdf/1410.5401.pdf
[4].Q. V. Le et al., ``Building high-level features using large scale unsupervised learning," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2013, pp. 8595_8598.
[5].J. G. Wolff and V. Palade, ``Proposal for the creation of a research facility for the development of the SP machine," Cognition Research Menai Bridge, U.K., Tech. Rep. 2015-11-16, 2015. [Online]. Available: http://bit.ly/1zZjj
[6].A. d'Avila Garcez et al., ``Neural-symbolic learning and reasoning: Contributions and challenges," in Proc. AAAI Spring Knowledge. Represent. Reason., Integral Symbolic Neural Approaches, Mar. 2015
[7].E. Davis and G. Marcus, ``Common sense reasoning and common sense knowledge in artificial intelligence," Commun. ACM, vol. 58, no. 9
[8].H. L. Dreyfus, What Computers Still Can't Do a Critique of Artificial Reason. New York, NY, USA: MIT Press, 1992.