International Conference on Recent Innovations in Science, Engineering & Management
R.R. INSTITUTE OF TECHNOLOGY, Bengaluru (India)
10th - 11th May 2019 www.conferenceworld.in ISBN : 978-93-87793-87-3

Successful Prediction of Missing Data over Multivariate Time Series on Apache Spark

[1]Afsha Firdose, [2]M J Prasanna Kumar, [3]Dr. Ravikumar G K
[1]M.Tech pursuing, Dept of CS&E, BGSIT, B.G Nagar
[2]Assistant Professor, Dept of CS&E, BGSIT, B.G Nagar
[3] Professor, Dept of CS&E, BGSIT, B.G Nagar

Abstract— *Now a days incredible amount of data are generated in many areas than ever before. However, from the collected data the loss of some values may occur and many methods were proposed in order to overcome data loss. But still, in multivariable-time series, most of the existing-methods are either might be unreal or could be inefficient to predict the missing data. In the proposed method first we check the reason for data loss. The data loss can occur due to two reasons: the first is hacker can alter the data and second way is due to environmental disaster or the hardware failure. In our paper, we have used improved matrix-factorization techniques in order to solve the problem of missing-data forecasting in multivariable-time series. We have also proposed five models to solve problem. Our proposed system is optimally designed to predict the information loss across multiple sources.*

1. INTRODUCTION

The advancement of devices to collect data from multiple sources and the resulting volume of data have grown to an unrivalled level since the period of big data started in recent years. Multivariable time series one regular data design are universal in many real world applications, like electric-equipment monitoring, weather-forecasting or economic-forecasting, environment state monitoring, security-surveillance and many more. In most applications, multiple sensors are enrolled to generate time-series data, and they usually share one common goal. For example, in the power-grid system, various diagnostic gases sensors are enrolled to monitor the status of the main power transformers and also generate multivariate-time series by measuring the content of the diagnostic gases over time. Let consider the example of "Internet of Things" devices used to check the air-quality or water-quality, a large number of sensors are used to produce multivariate-time series of the external-environment, e.g.,. In the medical and health-care systems, numerous sensors can likewise be used to monitor the health and general prosperity of old persons, while additionally ensure that the proper treatment is being directed and helping people regaining lost mobility via therapy as well. In this paper, the sensors sharing one common goal are treated as a sensor network.

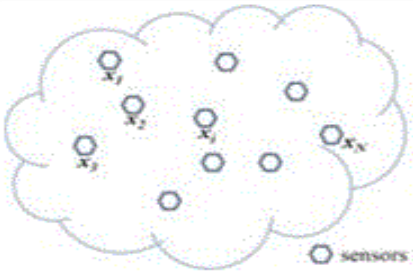

Fig.1: Illustration of a sensor network

Unfortunately, due to the environmental-disaster or harsh-working conditions or uncontrollable factors, such as the severe-weather condition, equipment or device failure or the change of communication-signals, usually causes the missing of values. For example, during the service or working time, missing values in power grid surveillance systems can occur for various reasons: such as quick evaporation of acetylene, the existence of contamination on the surface of the platinum alloy of a gas meter, etc. Yet, in practice, sensors-failure and communication-failures are more common factors that causes the missing of data in many applications. And still worse, immediate fixation of these practical problems is rarely possible and even might cost too much.

The unavoidable 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 to solve this problem 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. Modeling-method is one of the most commonly used solutions, to forecast the missing- values using some common sense. Representative-modeling approaches include deterministic-models, stochastic-models, and state-space-models. Nevertheless, these methods either focus on forecasting the missing-data in the time-series from one single source or could not effectively handle data-loss from multiple sources.

1. Proposed system

In this proposed paper, we combine 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 multivariate-time series. First, as each time series rarely fluctuate wildly over time. Thus, we try to take advantage of the characteristic to reduce the prediction error of the missing data-prediction in multivariable-time series. 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.

Clearly, 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.

Moreover, to treat the correlated-sensors or uncorrelated-sensors differently, we further improve the sensor network regularization terms of the objective function by incorporating similarity functions. By taking advantage of the knowledge of time-series across multiple sources, five models are built in the paper:

(1) MFS: Matrix Factorization with Smoothness constraints;

(2) CSM: Correlated Sensors based Matrix factorization;

(3) USM: Uncorrelated Sensors based Matrix factorization;

(4) CSMS: Correlated Sensors based Matrix factorization with Smoothness constraints;

(5) USMS: Uncorrelated Sensors based Matrix factorization with Smoothness constraints;

If we consider big data, more massive volume of time series data are generated nowadays than ever before. Besides, analyzing big data is a complex and time-consuming task, which needs more efficient and specific analysis tool than traditional ones. Thus parallel versions of matrix-factorization have become of great interest Apache-Spark is a large scale distributed data processing environment that builds on the principles of scalability and fault tolerance that are instrumental in the success of Hadoop and MapReduce. Here, we implement our proposed approaches using the Apache-Spark platform.

The above proposed CSM, USM and MFS models aim at taking advantage of either the information across multiple-

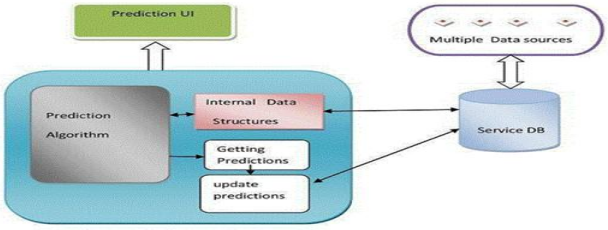

Fig 2: System Architectur

sources or the temporal-smoothness of time-series. To evaluate the performance of the proposed method, root mean squared error (RMSE) is used to measure the prediction quality RMSE is defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i,j}(1 - W_{ij})(X_{ij} - \hat{X}_{ij})^2}{\sum_{i,j}(1 - W_{ij})}},$$

where Xij is the raw time series matrix and ^ Xij is the corresponding predicted value. W is the indicator value.

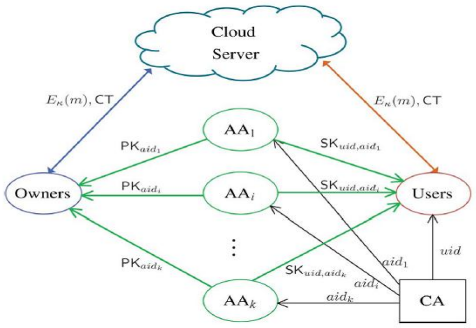

Fig 3: missing data prediction implemented

The above fig shows the working procedure of missing data-prediction on apache spark. In the existing system the data loss can be happen in two ways: one is data can be loss at source due to device problem or environmental disaster, the second way is due to hacker attack. After data get transferred from source it may travel through multiple network layers in order to reach the destination. During the data traveling time hackers can alter the source data and data loss can occur. To avoid this problem a prediction-method will be used. In the proposed system we have developed owner module, user module, cloud(server) module. The data generated at owner module. The owner will send the data to the cloud, during data transfer if data get corrupted by the hacker then the notification about data loss will be automatically generated and sent to the owner directly instead of sending it to the cloud. So that the owner resend the data and also can identify the details of hacker and block the hacker id etc. by doing like this the data loss can be minimized. To identify data loss by hacker and to get the data loss notification we have proposed below prediction-algorithm. If the data loss done due to

environmental problem or device problem then that information will be send to the owner. The detailed reason for data loss will be send to the owner directly. The data-loss information due to hardware-failure or environmental-disaster problem a trigger methodology will be used.

repeat
  γ = computing the best step size
  for $i = 1$ to N **do**
    $S_i = S_i - \gamma \frac{\partial \mathcal{L}_V}{\partial S_i}$
  end for
  for $j = 1$ to M **do**
    $V_j = V_j - \gamma \frac{\partial \mathcal{L}_V}{\partial V_j}$
  end for
until Convergence
Predicted $\hat{X} = SV^T$

Algorithm 1:Missing Data-Prediction in Multivariable-Time Series

2. Conclusion:

In our paper, we have proposed novel-methods to constrain the matrix-factorization for forecasting the missing-data in the time-series from multiple sources. 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. The proposed
method is best for predicting missing data.

References:

[1] S. F. Wu, C.-Y Chang and S. –J. Lee "Time series forecasting with missing values" in Proc. 1st Int. Conf. Ind. Netw. Intell. Syst., 2015.
[2]151-156 N. Meger, C. Rigotti and C. Pothier "Swap Randomization of bases of sequences for mining satellite image time series" in Proc. Eur. Conf.Mach Learn.Knowl. Discovery Databases, 2015.