IoT 시스템을 위한 Tensor Factorization 기반 누락 데이터 복구 알고리즘

악터샤티, 윤석훈*
울산대학교

erittrashathee@gmail.com, *seokhoonyoon@ulsan.ac.kr

A Missing Data Recovery Algorithm for IoT Systems based on Tensor Factorization Technique

Shathee Akter, Seokhoon Yoon*
Ulsan Univ.

Abstract

To make life more automatic and easier, nowadays, the internet of things (IoT) is being applied in numerous fields, e.g., smart city, intelligent transportation, and logistic. In order to provide good quality services, IoT systems need to have high-quality sensor data. However, sensing data in IoT systems are often missing owing to various reasons (e.g., mechanical faults, loss of communications, measurement, and synchronization errors), which affect the data processing and reduces the reliability and quality of the services and applications. Therefore, missing data imputation is crucial in IoT systems. This paper proposes a data imputation method based on the tensor factorization and Bayesian approach to estimate the missing values. In addition, it takes spatial-temporal correlation between data into account and aims to minimize the error between actual and estimated sensing data.

Ⅰ. Introduction

With the recent technological advancements, such as computational power, memory capacity, sensing technology, and wireless communication, Internet of things (IoT)-based applications have become applicable in a wide range of domain, e.g., healthcare, logistic, transportation, monitoring, and smart building [1]. IoT collects a large amount of data through seamless connections of a huge number of sensors and actuators with various objectives and analyzes data by utilizing fog, edge, and cloud computing to provide services to end-users [2]. However, in reality, data collection may not always be successful, i.e., data points can be missing due to various reasons (such as unreliable sensor devices, unstable network communication, synchronization problems, environmental and mechanical errors), which may decrease the reliability of the services and performance of the applications [3]. Therefore, in this paper, we aim to approximate the missing values as closely as possible to the actual values. We propose a missing data imputation framework based on the tensor factorization technique and probabilistic approach to minimize the error between predicted and actual value.

Ⅱ. Tensor factorization-based data recovery

In this section, we introduce the considered system model and proposed missing data recovery framework for IoT systems.

Assume that there are K sensing tasks, and the location of each task from where data need to be collected is denoted by (l_x, l_y), where l_x and l_y are the x-coordinate and y-coordinate, respectively. The tasks are performed periodically, i.e., the whole sensing period is divided into Z number of timeslots, and in each timeslot, the sensing data are collected. The sensing data collected at timeslot $z(1 \le z \le Z)$ and location (l_x, l_y) is denoted by S_{xyz}. Then, let's define an observed data tensor $S \in R^{K \times K \times Z}$ that contains S_{xyz} and $S = S \circ I$, where $I \in R^{K \times K \times Z}$ is a binary tensor that indicates which sensing task is incomplete. Our objective is to find a tensor S' by deducing the missing value based on the observed data in S such that the error between S' and G (G is a tensor, where no data is missing) is minimum. To estimate S', we propose a data recovery framework based on the probabilistic perspective of the tensor factorization technique [4] in this paper.

According to tensor factorization [5], S can be approximated by using the outer product of the three low-rank matrices $A \in R^{D \times K}$, $B = \in R^{D \times K}$, and $C \in R^{D \times Z}$. Therefore, by estimating these latent feature matrices,

the missing data points can be recovered. We apply the Bayesian inference approach, i.e., maximum a posteriori probability (MAP) estimate, to find the probabilistic estimation of A, B, and C. The prior distribution of A, B, and C is a gaussian distribution with mean μ and standard deviation σ. The conditional distribution over observed entry S_{xyz} is given by the normal distribution with mean μ_0 and variance σ_0^2. μ_0 is the sum over the elementwise multiplication of x, y, and z^{th} columns of A, B, and C. Then, the approximated features of A, B, and C are obtained by computing MAP estimate using numerical optimization, i.e., gradient descent algorithm, with the objective of minimizing the error between approximated tensor S' (composed using estimated A, B, and C) and S. However, the varying observed data scale may lead to a long execution time; thus, we scale data the data between 0 and n, where n is the upper bound.

Ⅲ. Result and Analysis

In this section, the proposed approach, namely tensor factorization-based data recovery using probabilistic approach (TF-DRP), is compared with the existing method K-nearest neighbor-spatio-temporal (KNN-ST) [6] using real-world dataset [7] under the effect of the different number of sensing tasks [10, 15, 20] to verify the performance, where the performance matrix considered is the root means square error (RMSE) between predicted and actual value. To simulate the environment, we extract weather data, i.e., information about light, of one day from 1 am to 7 am from the Intel Lab dataset [6], where the interval between each data is 30 minutes. Furthermore, we assume that 70 % of values are missing from the dataset. μ, μ_0, and σ_0^2, σ^2 are set to the mean and variance of the observed data.

Table 1. shows the performance of the TF-DRP and KNN-ST when the number of tasks changes from 10 to 20. It can be seen that the error rate is lower in TF-DRP in all cases than KNN-ST because TF-DRP can capture the spatial-temporal correlation well using latent feature matrices. Furthermore, when the number sensing tasks increases, the error decreases as there are more samples for training the gradient descent in TF-DRP and for interpolation in KNN-ST.

Table 1: Effect of different number of sensing tasks

Number of sensing tasks	RMSE	
	TF-DRP	KNN-ST
10	0.213	0.450
15	0.142	0.474
20	0.132	0.408

ACKNOWLEDGMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2019R1F1A1058147).

References

[1] L. Atzori, A. Iera, and G. Morabito, “The internet of things: a survey,” Computer Networks, vol. 54, no. 15, pp. 2787- 2805, 2010.

[2] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari and M. Ayyash, "Internet of Things: A survey on enabling technologies protocols and applications", IEEE Commun. Surveys Tuts., vol. 17, no. 4, pp. 2347-2376, 4th Quart. 2015

[3] X Yan, W Xiong, L Hu, F Wang and K Zhao, "Missing value imputation based on Gaussian mixture model for the internet of things", Mathematical Problems in Engineering, pp. 1-8, 2015.

[4] L. Xiong et al., “Temporal Collaborative Filtering with Bayesian Probabilistic Tensor Factorization”, Proc. 2010 SIAM International Conference on Data Mining, pp. 211-222, Dec. 2010.

[5] F. L. Hitchcock, “The Expression of a Tensor or a Polyadic as a Sum of Products”, Journal of Mathematics and Physics, vol. 6, no. 1, pp. 164- 189, Apr. 1927.

[6] N. Marchang and R. Tripathi, “KNN-ST: Exploiting Spatio-Temporal Correlation for Missing Data Inference in Environmental Crowd Sensing”, IEEE Sensors Journal, vol. 21, no. 3, pp. 3429-3436, Sept. 2020.

[7] Intel Berkeley Research Lab Data, (http://db.csail.mit.edu/labdata/labdata.html).