IoT Time-Series Missing Value Imputation - Comparison of Machine Learning Methods

Xudong Chen[1], Bin Sun[1(✉)], Shuhui Bi[1], Jiafeng Yang[2], and Youling Wang[2]

[1] School of EE, University of Jinan, Jinan 250000, China
cse_sunb@ujn.edu.cn
[2] Jinan Lingsheng Info Tech. Co., Ltd., Huaiyin District, Jinan 250000, China

Abstract. Data about time series has been researched for ages in various fields. In past few years, with the advancements of the Internet of Things (IoT) and the use of data acquisition devices, more and more time series data are being provided. However, due to the failure of the data acquisition equipment, some data is lost, and these lost data may contain important information. In order to deal with these lost data, many different machine learning algorithms have appeared, such as K-NN, CNN, random forest, etc.

The purpose of this work is to compare the effects of two diverse models, K-NN and Random Forest on missing values imputation which is in traffic data, and to evaluate the two models, the root mean square error (RSTM) [1] index is adopted.

Keywords: Time series · Missing values · Machine learning · Imputation

1 Introduction

With the extensive application of 5G technology and the progress of Internet of Things (IoT) technology, as far as data in people's real life is collected, recorded and analyzed by a variety of smart devices and sensors, which is called time-series data. Mining and analyzing the hidden information behind the collected time series data is of great significance for future traffic prediction, global positioning and even intelligent medical applications. However, due to some unstable factors of data acquisition equipment, a good deal of data is missed, and these collected data is often incomplete, consequently the deep mining of the data is inhibited. Therefore, it is considerable to manage missing values.

2 Backgrounds and Related Works

Many scholars at home and abroad carry out extensive research on the attribution problem of missing values in the present anomalies.

Some scholars attribute missing value imputation to a special case of time series prediction. We consulted a large amount of literature on studying time series prediction

B. Wang et al. (Eds.): ICMTEL 2023, LNICST 535, pp. 406–411, 2024.
https://doi.org/10.1007/978-3-031-50580-5_37

[2]. B. Sun [3, 4] proposes to improve vicious traffic capability and malware defender performance by examining the characteristics of session datasets and finding matching characteristics to fuse into a dataset.

Wang's laboratory result adequately shows that the missing value imputation based on edge calculation has the best effect among other filling methods, and extremely reduces the energy damage in AIoT [5]. In Fatlawi's work, an adaptive classification model based on machine learning is proposed to handle missing values [6]. Using linear, vector of support regression to make time series prediction about missing values is Yen's work [7]. Velasco introduces a new framework that combines first-order Markov chains with a multivariate imputation method based on comparative approaches to multiple deep learning and time series prediction models, greatly improving the capacity of data entry techniques [8].

On the problem of how to deal with missing values in data set, scholars have carried out a lot of experiments using different methods based on machine learning or deep learning, and they all achieved good results. In Hussain's work, a novel neural network based on hybrid CNN-LSTM is proposed in order to predict large missing value gaps, which greatly reduces the missing value gap in Internet of Things applications [9]. Baggag's team developed a regular decomposition method based on machine learning to fill the missing values from the data segment which is from power fault due to technical malfunction. The method is accurate and efficient, and has achieved good results in other traffic data sets [10]. Che developed a new deep learning models named GRU-D in "Recurrent Neural Networks for Multivariate Time Series with Missing Values" [11] that get well performance in dealing with real-world datasets' missing values in time series. Bergmeir's team uses a blocking mode with cross validation for time series evaluation is "On the use of cross-validation for time series predictor evaluation" [12]. Rahman's [13] team gave some models such as RNN and deep-NN are proposed to predict the hourly electricity consumption of public safety buildings and the total electricity consumption of residential buildings, and good results have been obtained from its experiment. B. Sun [14, 15] and Geng used a dynamic graph structure to work with Graph Neural Network (GNN) algorithm to predicting traffic data.

A method of "tensor mode as matrix extension" into traffic data modeling for the first time, which solves the traffic data missing problem caused by detector and communication faults commendably which is from Tan's team [16]. Habtemichael's team presents a non-parametric and data-driven short-term traffic prediction method "based on identifying similar traffic patterns using the enhanced KNN algorithm" that can accurately predict the evolution of traffic in practical traffic conservancies and controlling applications [17]. In particular, Li [18] propose a multiple scales processing method to measure the missing values of traffic relevant anomalies, and the consequence shows that the model they came up with is superior to other criterion, in particular for missing block models with outstanding low missing rates.

3 Results and Analysis

We describe the benchmark results between the two models and then compare Random Forest (RF) and KNN in more detail. Table 1 shows the results of RMSE with different missing ratio by using RF model and KNN model.

Table 1. RMSE in different missing ratio

Missing Ratio	RMSE (Random Forest model)	RMSE (K-NN model)
1.9% (Origin)	10.7910	10.7910
10%	10.8319	10.8023
20%	10.7988	10.8122
30%	10.8035	10.8236
40%	10.8101	10.8315
50%	10.8157	10.8281
60%	10.8181	10.8710
70%	10.8211	11.4211
80%	10.8197	12.3217

In this experiment, Fig. 1 shows that we increased the missing ratio of the dataset from 1.9% to 80%, and in this process, the value of RMSE remained within the range of 10.7 to 10.9. When the proportion of missing values increased from 1.9% to 10%, RMSE showed the largest increase with a range of +0.0408. Then, the missing proportion continued to increase to 20%, and the value of RMSE showed a significant decline, with a range of −0.0330. When the proportion of missing value increased from 20% to 70%, the value of RMSE gradually increased, but the overall change range was 54.5%.

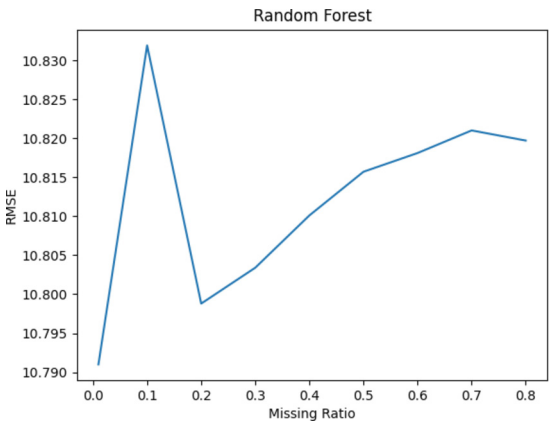

Fig. 1. The RMSE in Random Forest model

We carried out the same missing processing on the same data set, as shown in Fig. 2. KNN model was used to accomplish missing values imputation of the dataset. As the missing ratio increased from 1.9% to 80%, the RMSE value changed from 10.79 to 12.33. When the data missing ratio is within the range of 1.9% to 60%, the value of

RMSE gradually increases from 10.79 to 10.87, with a change range of 5.23%. Visible effect of KNN model imputation is relatively stable in this process. However, when the data missing ratio increased from 60% to 80%, the RMSE value increased sharply from 10.8710 to 12.3217, with a change range of 94.8%.

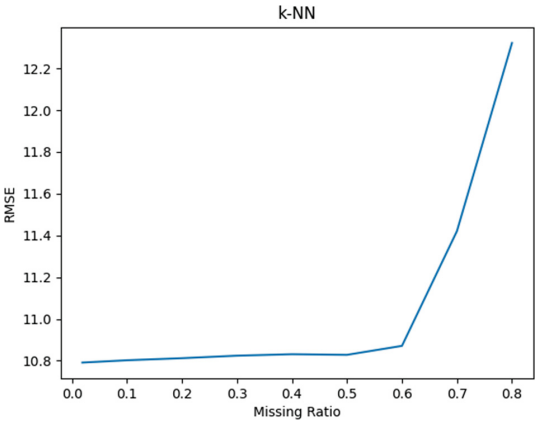

Fig. 2. The RMSE in KNN model

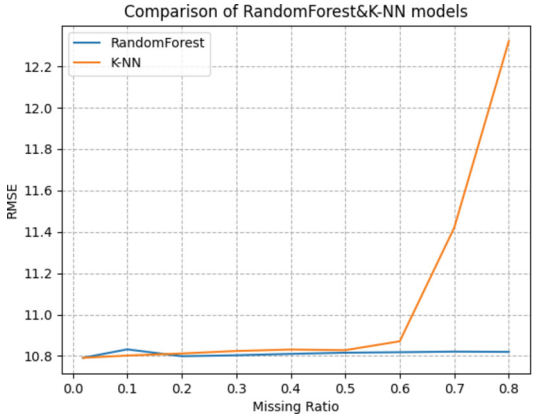

Fig. 3. Comparison of Random Forest and K-NN models

4 Conclusion and Future Work

By analyzing the effects of K-NN and Random Forest models on the missing values imputation in this dataset, as Fig. 3 shows, when the missing values' ratio of the dataset is in low level, the KNN model is better than the RF model for the filling of missing values. As the proportion of missing values increased from 20% to 80%, RF model

was significantly better than KNN model in filling the missing values. In addition, the running time of the Random Forest model is much less than that of the KNN model in the experimental process, which also means that the KNN model has a lot of room for improvement.

The machine learning or deep learning algorithm is used to fill in the missing values of time series [19], the essence of which is the same as that of time series prediction. In this experiment, LSTM, BiLSTM, GNN and other deep learning algorithms were envisioned. The above algorithm can be implemented when missing values are regularly distributed in a data set. However, the distribution of missing values in the real world dataset is mostly irregular [20], and when LSTM, BiLSTM and other algorithms are used, it will be impossible to learn. If GNN algorithm is used, different data features in the data set need to be transformed into picture structure, which is a complicated process. In the future, we will try to use deep learning algorithm to accomplish the missing values imputation of time series.

References

1. Willmott, C.J., Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. **30**(1), 79–82 (2005)
2. Sun, B., et al.: Correcting and complementing freeway traffic accident data using mahalanobis distance based outlier detection. Techn. Gaz. **24**(5), 1597–1607 (2017)
3. Sun, B., et al.: Securing 6G-enabled IoT/IoV networks by machine learning and data fusion. EURASIP J. Wirel. Commun. Netw. **1**, 1–17 (2022)
4. Sun, B., Ma, L., Shen, T., Geng, R., Zhou, Y., Tian, Y.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. Wirel. Commun. Mob. Comput. (WCMC) **2021**(6692390), 1–11 (2021)
5. Wang, T., et al.: Missing value filling based on the collaboration of cloud and edge in artificial Intelligence of Things. IEEE Trans. Ind. Informat. **18**(8), 5394–5402 (2021)
6. Fatlawi, H.K., Kiss, A.: An adaptive classification model for predicting epileptic seizures using cloud computing service architecture. Appl. Sci. **12**(7), 3408 (2022)
7. Yen, N.Y., et al.: Analysis of interpolation algorithms for the missing values in IoT time series: a case of air quality in Taiwan. J. Supercomput. **76**(8), 6475–6500 (2020)
8. Velasco-Gallego, C., Lazakis, I.: A novel framework for imputing large gaps of missing values from time series sensor data of marine machinery systems. Ships Offshore Struct. **17**(8), 1802–1811 (2022)
9. Bogaerts, T., et al.: A graph CNN-LSTM neural network for short and long-term traffic forecasting based on trajectory data. Transp. Res. Part C Emerg. Technol. **112**, 62–77 (2020)
10. Baggag, A., et al.: Learning spatiotemporal latent factors of traffic via regularized tensor factorization: imputing missing values and forecasting. IEEE Trans. Knowl. Data Eng. **33**(6), 2573–2587 (2019)
11. Che, Z., et al.: Recurrent neural networks for multivariate time series with missing values. Sci. Rep. **8**(1), 1–12 (2018)
12. Bergmeir, C., Benítez, J.M.: On the use of cross-validation for time series predictor evaluation. Inf. Sci. **191**, 192–213 (2012)
13. Rahman, A., Srikumar, V., Smith, A.D.: Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. Appl. Energy **212**, 372–385 (2018)

14. Sun, B., et al.: Dynamic emergency transit forecasting with IoT sequential data. Mob. Netw. Appl., 1–15 (2022)
15. Sun, B., et al.: Prediction of emergency mobility under diverse IoT availability. EAI Endorsed Trans. Pervasive Health Technol. **8**(4), e2 (2022)
16. Tan, H., et al.: A tensor-based method for missing traffic data completion. Transp. Res. Part C Emerg. Technol. **28**, 15–27 (2013)
17. Habtemichael, F.G., Cetin, M.: Short-term traffic flow rate forecasting based on identifying similar traffic patterns. Transp. Res. Part C Emerg. Technol. **66**, 61–78 (2016)
18. Li, L., et al.: Missing value imputation for traffic-related time series data based on a multi-view learning method. IEEE Trans. Intell. Transp. Syst. **20**(8), 2933–2943 (2018)
19. Xiao, Y., Shao, H., Han, S., Huo, Z., Wan, J.: Novel joint transfer network for unsupervised bearing fault diagnosis from simulation domain to experimental domain. IEEE-ASME Trans. Mechatron. **27**(6), 5254–5263 (2022)
20. Chen, M., Shao, H., Dou, H., Li, W., Liu, B.: Data augmentation and intelligent fault diagnosis of planetary gearbox using ILoFGAN under extremely limited samples. IEEE Trans. Reliab. (2022)