Time Series Data Imputation Using Expectation-Maximization with Principal Component Analysis

Renkang Geng, Jing Cao(✉), Qinjun Zhao(✉), and Yujie Wang

School of Electrical Engineering, University of Jinan, Jinan 250022, Shandong, China
caojingjn@163.com, cse_zhaoqj@ujn.edu.cn

Abstract. Data quality is the basis of data analysis and determines the effect and depth of data analysis. Missing values are a very important factor affecting data quality. Since machine learning algorithms have been used to process data, the processing of missing values has become an important field of machine learning. For the vast majority of data, the first step of data analysis is often to complete the missing values of data. With the increasing complexity of current social traffic conditions and the increasingly serious urban road congestion, traffic data, as the most intuitive reflection of urban road conditions, has great application value and application potential. The quality of traffic data is directly related to whether we can accurately predict the traffic flow and judge the road traffic conditions so as to effectively govern the urban traffic problems. Therefore, it is very important to choose appropriate algorithms and methods to fill in the missing values in traffic data quickly and accurately. In this paper, taking traffic flow data as an example, we create different missing value ratios for the data and use the PCA-EM algorithm to fill in the missing value. Through the experimental results, we have a preliminary evaluation of the comprehensive performance of the PCA-EM algorithm when it is used to fill in the missing value of traffic data.

Keywords: Traffic flow data · Missing value · PCA-EM

1 Introduction

It is of great significance to study traffic data for improving urban congestion and reducing traffic anomalies [1, 2]. If the traffic data can be accurate processing and analysed, can effectively predict the future traffic flow, and carries on the anomaly detection of a traffic incident, so that the relevant traffic department, the traffic in an orderly, the guidance of the timely handling of traffic accidents, channel distribution vehicles, improve urban congestion and facilitate public travel, enhance people's quality of life, To ensure the safety of citizens [3, 4]. For the analysis of traffic data, the key step lies in whether the missing values can be accurately filled, which seriously affects the accuracy of our analysis [5].

S. Wang et al. (Eds.): IoTCare 2021, LNICST 415, pp. 352–357, 2022.
https://doi.org/10.1007/978-3-030-94182-6_26

Since the beginning of the last century, people have begun to pay attention to the problem of data quality, so the lack of data has deeply affected many ongoing studies. Lack of data will affect task analysis and make it more difficult, resulting in deviation of results, which will greatly reduce the work efficiency of statistical work [6]. Especially when there is a systematic difference between complete observation and incomplete observation, there will be great limitations and deviations in the conclusions made by using mathematical methods and conventional statistical methods on incomplete data sets [7, 8]. Such structure cannot replace the original data. The progress of time series missing value preprocessing methods enables us to use better methods for research. Missing data filling is one of the important research contents in data clarity and data mining [9].

The generation of data missing value will cause problems in many ways, for data development and in the field of data mining aspects of the work, the existence of data missing value, will produce a lot of influence, data sets to produce missing value can cause a lot of useful information is missing [10, 11], and some programs only in view of the complete data sets, which reduces the work efficiency, also, for instance, For KNN algorithm, it is very excellent in terms of classification accuracy, and the accuracy of decision tree classification is much higher than that of the largest category classification. However, missing values have a very important impact on the classification ability of KNN [12, 13]. In the KNN algorithm, data sets containing missing values are classified, and the classification accuracy is completely dependent on the size of the missing proportion.

The larger the missing proportion, the smaller the classification accuracy. If the data is almost completely missing [14], the classification accuracy can reach 0. The uncertainty in the data set may cause confusion of the data distribution and the reduction of the uncertainty in the data set. This situation will be more obvious, and the certainty in the data set will be more difficult to grasp. The inclusion of missing values in a data set can cause confusion in the data mining process and add a lot of unreliability to the data output [15, 16].

2 Related Work

Due to the different reasons for data missing, there are different types of data missing. In order to deal with all kinds of data missing problems more conveniently, it is necessary to classify data missing. In this way, we can find out the solution to the problem more efficiently and specifically, so that our data filling is closer to the real value and the result is more accurate [17, 18]. We can classify the missing data according to the relationship between the complete data and the missing data. This classification method is called missing mechanism classification, which is generally divided into three categories: 1) Completely random missing [19]: Completely random loss refers to the distribution of the missing value in a data set is random, no rules to follow, this has to do with his own properties, and there is no relationship between adjacent data, for example, when we study student's result, the lack of a course grade has nothing to do with other results, so we can't find a rule to fill the missing value from other grades. 2) Random missing: Random deletion index data deletion is only related to the value of the complete attribute.

3) Non-random missing: this kind of missing data is related to its own attributes and other data in the data set. Generally speaking, such missing data cannot be deleted directly [20].

Probabilistic PCA-EM filling [21]: Probabilistic PCA-EM is an algorithm that uses a variational EM algorithm to process missing data. Essentially, it is a method that considers the probability distribution of each variable. After determining the probability function of the principal element and error, the model is established by the EM algorithm. Firstly, the implicit variable X and the probability distribution of the observed data were obtained by standard centralized processing of the original training sample data. Then, the expectation was calculated and the maximum likelihood estimate of the hidden variable was calculated by using the existing estimate of the hidden variable. Then maximize the maximum likelihood value to calculate the value of the parameter. The objective function of the Probabilistic PCA-EM algorithm is the lower bound of the logarithmic likelihood of data. In high dimensional space, the computation required for each iteration of the PCA-EM algorithm is much smaller than that of traditional algorithms.

3 Methods

The specific steps of this experiment are as follows: First reads raw traffic data and will be treated as raw data branch according to each day, each column data corresponding to the same period under different days of traffic flow data, that is to say, each column data into a set of common attributes of the data, then we can make different missing value in each column proportion, finally, use PCA - EM algorithm to fill each column of missing value.

In this experiment, we used the control variable method to create missing values of different proportions for the same data set and then used the PCA-EM algorithm to fill in the missing values. Finally, the filling effect was compared and evaluated. In this way, we can effectively compare the filling effect of the PCA-EM algorithm when the proportion of missing values is different, and get the best filling effect of the PCA-EM algorithm when facing the traffic flow data under what condition the missing values are in. In this experiment, the missing values artificially created accounted for 10%, 30%, 50%, 70% and 90%, respectively.

In order to compare the distribution of real data and filled data, we draw a Q-Q graph and draw the baseline of Y = X. The essence of the Q-Q graph is that, firstly, real data and filled data are arranged according to the quantile and corresponding real data and a filled data under the same quantile are taken to draw a coordinate point as the coordinate (x, y), and then the whole Q-Q graph is drawn through iteration. In addition to the Q-Q diagram, in order to accurately quantify the relationship between real data and filled data, we also used MSE, RMSE, MAE, SMAPE and other indicators to intuitively analyze the effect of completion. The indices are formulated as follows:

$$MSE = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}$$

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|(y_i - \hat{y}_i)\right|$$

$$SMAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{\left|\hat{y}_i - y_i\right|}{\left(\left|\hat{y}_i\right| + |y_i|\right)/2}$$

4 Results

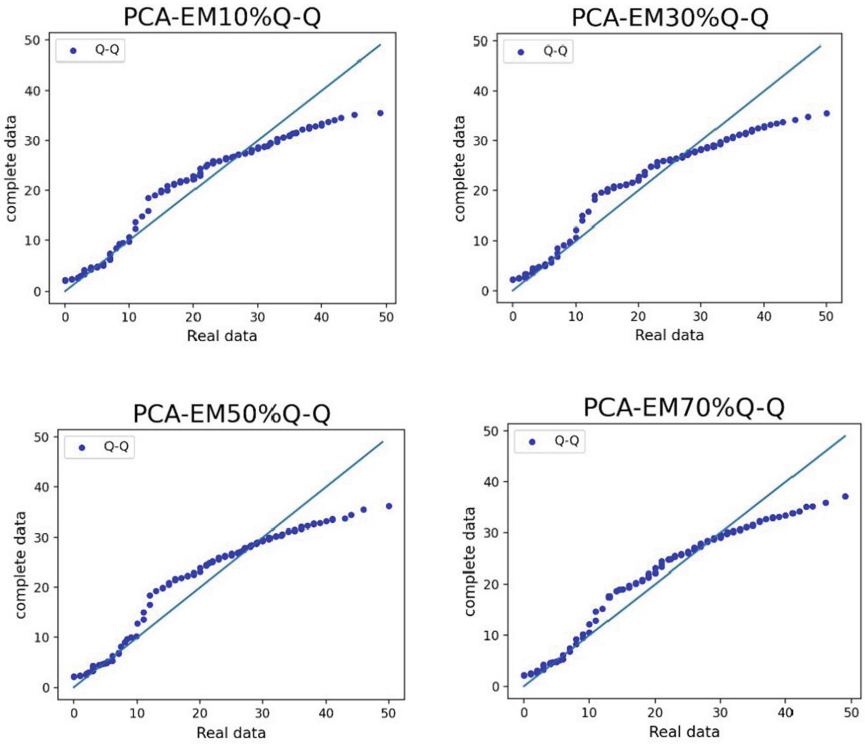

Fig. 1. The figure shows the Q-Q diagram completed by the PCA-EM algorithm after we have created different missing value ratios for the data. The abscissa is the real data, and the ordinate is the completed data. It can be seen that the complete effects in the four cases are generally similar. When the real data is less than 25, the complete data is larger than the real data, while when the real data is greater than 25, the complete data is smaller.

Table 1. Under different missing value ratio, the value of each indicator

Index	10%	30%	50%	70%
MSE	56.6	56.7	56.0	57.0
RMSE	7.5	7.5	7.5	7.5
MAE	5.4	5.4	5.3	5.4
SMAPE (%)	33.4	33.0	33.2	33.7

From all the data in the table, we can find that when the missing value accounts for 30%, the PCA-EM algorithm has the best filling effect, and SMAPE is 33.0%. In the numerical analysis of MSE, when the missing value accounted for 50%, the value of MSE was the lowest, only 56.0. In general, the overall filling effect of the PCA-EM algorithm is good. In the case of different missing values, the filling accuracy of the PCA-EM algorithm is slightly different, and the difference is small, indicating that the overall optimization of the PCA-EM algorithm is good when it is used to fill the traffic flow data. We finally proved that the PCA-EM algorithm is most suitable for filling traffic flow data with missing values accounting for about 30% (Table 1 and Fig. 1).

5 Conclusion

In this paper, we first made a general introduction to data missing and summarized several categories of data missing. Then it describes the influence of missing values in traffic data on data analysis, and the importance of studying the missing values of traffic data for improving urban congestion and reducing traffic accidents. During the experiment, we selected the traffic flow data and adjusted the proportion of missing values. PCA-EM method was used to fill the traffic flow data with different proportion of missing values successively, and the corresponding filling effect was obtained. Finally, we have a preliminary evaluation of the comprehensive performance of the PCA-EM algorithm when it is used to fill in the missing values of traffic data. In the future, I will continue to make further research in this field.

References

1. Jie, L., Van Zuylen, H.J.: Road traffic in China. Procedia Soc. Behav. Sci. **111**, 107–116 (2014)
2. Liu, Z., Yue, X., Zhao, R.: The cause of urban traffic congestion and countermeasures in China. Urban Stud. **11**, 90–96 (2011)
3. Sun, B., Cheng, W., Goswami, P., Bai, G.: An overview of parameter and data strategies for k-nearest neighbours based short-term traffic prediction. In: ACM International Conference Proceeding Series 2017, pp. 68–74. ACM (2017)
4. Sun, B., Cheng, W., Bai, G., Goswami, P.: Correcting and complementing freeway traffic accident data using Mahalanobis distance based outlier detection. Tehnicki Vjesnik-Technical Gazette **24**(5), 1597–1607 (2017)

5. Wen, H., Sun, J., Zhang, X.: Study on traffic congestion patterns of large city in China taking Beijing as an example. Procedia Soc. Behav. Sci. **138**, 482–491 (2014)
6. Li, J., Walker, J.L., Srinivasan, S., et al.: Modeling private car ownership in China: investigation of urban form impact across megacities. Transp. Res. Rec. **2193**(1), 76–84 (2010)
7. Sun, B., Ma, L., Shen, T., et al.: A robust data-driven method for muti-seasonal and heteroscedastic IoT time series preprocessing. In: Wireless Communications and Mobile Computing (WCMC), p. 6692390 (2021)
8. Ma, L., Sun, B., Han, C.: Learning decision forest from evidential data: the random training set sampling approach. In: 4th International Conference on Systems and Informatics (ICSAI), Hangzhou, China (2017)
9. Bramer, M.: Principles of Data Mining. Springer, London (2007). https://doi.org/10.1007/978-1-84628-766-4
10. Scheffer, J.: Dealing with missing data (2002)
11. Ahmed, M.S., Cook, A.R.: Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques (1979)
12. Ma, L., Sun, B., Li, Z.: Bagging likelihood-based belief decision trees. In: 20th International Conference on Information Fusion (FUSION), Xi-An, China, pp. 1–6 (2017). http://ieeexplore.ieee.org/abstract/document/8009664/
13. Geng, R., Sun, B., Ma, L., Zhao, Q., Shen, T.: Anomaly-aware in sequence data based on MSM-H with EXPoSE. In: 40th Chinese Control Conference (CCC 2021), Shanghai, China (2021)
14. Zhang, S., Li, X., Zong, M., et al.: Learning k for kNN classification. ACM Trans. Intell. Syst. Technol. (TIST) **8**(3), 1–19 (2017)
15. Liu, P., Lei, L., Wu, N.: A quantitative study of the effect of missing data in classifiers. In: The Fifth International Conference on Computer and Information Technology (CIT 2005), pp. 28–33. IEEE (2005)
16. Sun, B., Cheng, W., Goswami, P., et al.: Short-term traffic forecasting using self-adjusting k-nearest neighbours. IET Intel. Transp. Syst. **12**(1), 41–48 (2018)
17. Sun, B., Cheng, W., Ma, L., Goswami, P.: Anomaly-aware traffic prediction based on automated conditional information fusion. In: International Conference on Information Fusion (FUSION), Cambridge, United Kingdom, pp. 2283–2289. IEEE (2018)
18. Zhang, S., Zhang, C., Yang, Q.: Data preparation for data mining. Appl. Artif. Intell. **17**(5–6), 375–381 (2003)
19. Bhaskaran, K., Smeeth, L.: What is the difference between missing completely at random and missing at random? Int. J. Epidemiol. **43**(4), 1336–1339 (2014)
20. Marlin, B.M., Zemel, R.S.: Collaborative prediction and ranking with non-random missing data. In: Proceedings of the Third ACM Conference on Recommender Systems, pp. 5–12 (2009)
21. Yu, L., Snapp, R.R., Ruiz, T., et al.: Probabilistic principal component analysis with expectation maximization (PPCA-EM) facilitates volume classification and estimates the missing data. J. Struct. Biol. **171**(1), 18–30 (2010)