Journal of Computers Vol. 30 No. 5, 2019, pp. 213-219
doi:10.3966/199115992019103005016

A Slope Based Selection Algorithm of Optimal Window in Lagrange Interpolation

Yue Shen[1], Bo Shen[1*], Ying-Ji Liu[2], Ling-Yun Chi[3]

[1] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
{17120109, bshen}@bjtu.edu.cn

[2] Key Laboratory of Operation Safety Technology on Transport Vehicles, Research Institute of Highway, Ministry of Transport, Beijing 100088 China
Yj.liu@rioh.cn

[3] School of Information and Communication Engineering, University of Electronic Science and Technology, Sichuan 611731, China
chilingyun@uestc.edu.cn

Received 13 January 2019; Revised 5 March 2019; Accepted 12 March 2019

Abstract. Aiming at the problem that the accuracy of Lagrange interpolation with fixed window is not high enough, based on vehicle positioning data, a slope based selection algorithm of optimal window in Lagrange interpolation is proposed. The optimal window size is determined by the slope which is calculated by connecting the Lagrange prediction value and the forward/backward data of the missing data, so as to obtain the optimal interpolation value. Taking the vehicle positioning data as an example to verify the proposed algorithm. The results have shown that the performance of the algorithm proposed is significantly improved compared with the fixed-window Lagrange interpolation algorithm and other methods. At the same time, it has been found that this method is beneficial to solve the problem of data oscillation after multiple interpolation caused by the "Runge" phenomenon in polynomial interpolation.

Keywords: Lagrange interpolation, optimal window, "Runge" phenomenon, slope, vehicle positioning data

1 Introduction

The application of technology in the era of big data has promoted the intelligence of all walks of life in cities. The intelligent state of urban transport is also the primary component of a smart city. Behind the intelligence, the support of big data is very important [1-3]. The quality of each data is related to the entire process of data analysis, storage, fusion and prediction [4-6]. In the process of collection, the dataset can be affected by many aspects, such as device failure, network or signal problems, and processing errors, etc., which will result in different degrees of missing data [7]. Therefore, how to deal with these missing values is a problem that researchers need to face [8].

The commonly used method is to repair rather than delete the missing data. The reason is that although the deletion is the simplest method, it is easy to cause the loss of useful information, which will adversely affect the subsequent data mining and application. The commonly used methods repairing missing data include: forward interpolation, linear interpolation, Lagrange interpolation, and association rule interpolation [9]. However, these methods are faced with the problem that the interpolation accuracy is not high enough, especially for data such as vehicle positioning data that requires relatively high precision [10]. Therefore, based on the Lagrange interpolation method [11], a slope based selection algorithm of optimal window in Lagrange interpolation is proposed. This method can improve the

* Corresponding Author

repairing accuracy of missing values by dynamically considering the laws of change of its adjacent values and selecting the optimal number of neighboring values to help to predict the missing data.

2 Methods

2.1 Lagrange Interpolation Algorithm

According to mathematical knowledge, a (n-1) degree polynomial can be found for n points known on the plane (any two points not on a straight line) [12]:

$$y = a_0 + a_1 x + a_1 x^2 + \ldots + a_{n-1} x^{n-1} \tag{1}$$

Substituting (x_1, y_1), (x_2, y_2) $\ldots (x_n, y_n)$ into the polynomial above,

$$y_1 = a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_{n-1} x_1^{n-1} \tag{2}$$

$$y_2 = a_0 + a_1 x_2 + a_2 x_2^2 + \ldots + a_{n-1} x_2^{n-1} \tag{3}$$

$$\ldots$$

$$y_n = a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_{n-1} x_n^{n-1} \tag{4}$$

So the Lagrange interpolation polynomial is [12]:

$$\begin{aligned} L(x_i) &= y_1 \frac{(x_i - x_2)(x_i - x_3)\ldots(x_i - x_n)}{(x_1 - x_2)(x_1 - x_3)\ldots(x_1 - x_n)} \\ &\quad + y_2 \frac{(x_i - x_2)(x_i - x_3)\ldots(x_i - x_n)}{(x_2 - x_1)(x_2 - x_3)\ldots(x_2 - x_n)} \\ &\quad \ldots \\ &\quad + y_n \frac{(x_i - x_1)(x_i - x_3)\ldots(x_i - x_{n-1})}{(x_n - x_1)(x_n - x_2)\ldots(x_n - x_{n-1})} \\ &= \sum_{p=0}^{n} y_p \prod_{q=0,\, q\neq p}^{n} \frac{x_i - x_q}{x_p - x_q} \end{aligned} \tag{5}$$

Substituting x_i corresponding to the missing value into the interpolation polynomial to obtain the approximate value $L(x_i)$ of the missing value. Note that the number n (equals k in this paper) of data involved in the calculation need to be preset, that means the order of the Lagrange interpolation (n-1) and the window size (k) mentioned in this paper should be preset.

2.2 The algorithm proposed

When using Lagrange interpolation, the window size of the interpolation is generally fixed (5 usually) [13]. But in fact, the window size that makes the optimal interpolation for each missing value is often not same. Considering that if the interpolation value is a suitable one to use, it must not deviate too far from the original data, that is, its slope calculated by connecting it with the adjacent data will not be too large. Inspired by this idea, we propose a slope based selection algorithm of optimal window to make Lagrange interpolation work better.

We define **k** as:

$$k = \begin{cases} \max\limits_{k \in U}(P_k) & P_k < 0 \\ \min\limits_{k \in U}(P_k) & P_k > 0 \\ k & P_k = 0 \end{cases} \tag{6}$$

$$P_k = \frac{L(x_i)_k - y_i - 1}{x_i - x_{i-1}} \tag{7}$$

x_i is the position of the missing value, $L(x_i)_k$ is the approximation of the missing value y_i using Lagrange interpolation (when the window size is k), and (x_{i-1}, y_{i-1}) is previous value of the missing value. $L(x_i)_k$ is calculated by the equation (5) (n = k).

U is a set of values of k. It is proved by experiments that when $k_{\max} > 20$, the interpolation precision using the proposed algorithm tends to be constant. Therefore, the upper limit of k in this paper is set to 20, that is, U= [1, 20]. But one thing has to be aware of is that U may be different for different datasets.

It can be seen from the equation(6) that when $P_k < 0(k \in U)$ is true, k is the corresponding value that maximize the slope; When $P_k > 0(k \in U)$ is true, k is the corresponding value that Minimize the slope; When neither of the above two is true, k has two or more alternatives at this time. Taking k with two alternatives, note them as k_1, k_2 (the calculation method is the same when multiple alternatives appear), and the last k is further determined by the following formula:

$$k = \begin{cases} \max\limits_{k \in (k_1, k_2)} (P_k) & P_k < 0 \\ \max\limits_{k \in (k_1, k_2)} (P_k) & P_k > 0 \end{cases} \tag{8}$$

$$P_k = \frac{y_{i+2} - L(x_i)_k}{x_{i+2} - x_i} \tag{9}$$

x_i is the position of the missing value, $L(x_i)_k$ calculated by the equation (5) is the approximation of the missing value y_i (when n=k), and (x_{i+2}, y_{i+2}) is the second data behind the missing value, which is chosen as a new reference value used for calculation.

If $P_k(k \in U)$ is still not all positive/negative, then other reference values adjacent will be selected to calculate the slope helping determine k until it has only one alternative. In general, k can be determined by comparing the slopes within choosing three references. The neighboring data used to calculate the slope can theoretically be freely selected. In this paper, we first choose the previous value and the second is the value second behind the missing data. The reason is that the data distribution on both sides of missing data may be different if it is at the peak / valley, so choosing references from different directions can help to determine k faster.

The algorithm is drawn as follows:

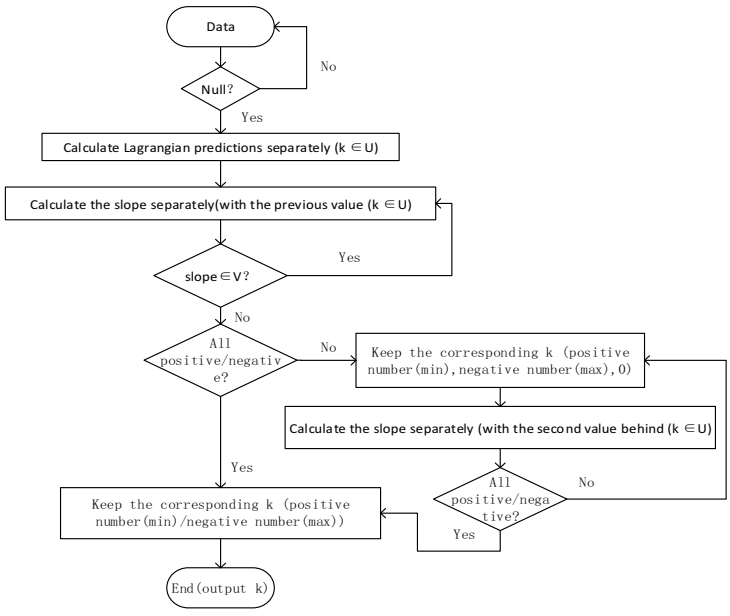

Fig. 1. Algorithm flowchart

In Fig. 1, V is used to avoid the useless calculation caused by the "Runge" phenomenon. Due to the influence of the "Runge" phenomenon, as interpolation window (k) increases, there will be obvious oscillation at both ends of the dataset to be interpolated, which will make the slope calculated here appears to be abnormally large/small. Therefore, choosing a suitable V, when n = k ∈ U, if the corresponding slope exceeds this range V, the predicted value (when k > n) is no longer calculated. This step can reduce the amount of calculation. In this paper, V is set to [-2, 2], which can be adjusted according to different datasets.

3 Evaluation Rule

In this paper, we use Mean Absolute Error (MAE) to evaluate the accuracy of each interpolation method, which is defined as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | E_i | = \frac{1}{n}\sum_{i=1}^{n} | Y_i - \hat{Y}_l | \quad \textbf{(10)}$$

MAE: Mean Absolute Error

E_i : The absolute error between the i_{th} actual value and the predicted value

Y_i : the i_{th} actual value

$\hat{Y}_i$: the i_{th} predicted value

The smaller the MAE, the better the performance.

4 Experiments

In this paper, the performance of the proposed algorithm is verified by vehicle positioning data. We randomly select the "latitude" and "longitude" positioning data of a single vehicle from August 18, 2018 18:36 to 21:36. For various reasons, the collected vehicle positioning data is not always clean, and there may be some data missing and abnormal. In this paper, we first compare and analyze the impact of different window sizes on the performance of Lagrange interpolation method. Then we compare the performance of the proposed algorithm and four kinds of commonly used interpolation methods: fixed window Lagrange interpolation, linear interpolation and forward interpolation. Last, the improvement of "Runge" phenomenon will be further shown and analyzed.

4.1 Lagrange Interpolation with Different Sizes of Window

In order to show that the impact of different window sizes on the performance of Lagrange interpolation method, we first delete some of the data in the sample data and repairing it using Lagrange interpolation with different window sizes, then compare them with the raw data to see the effect of window sizes on Lagrange interpolation. The results have been show in Table 1. In order to save space, we only list MAE values when the window size $k \in [1, 6]$. (We give 5 MAE values between raw data numbered 12, 23, 59, 79, 92 and its corresponding filled data here.) In this section, the sample data we use is randomly chosen from the vehicle positioning data we have mentioned above and the evaluation rule is the MAE values.

Table 1. MAE values after Lagrange interpolation with different window sizes ($k \in [1, 6]$)

	12	23	59	79	92	Average MAE
k=1	1.5×10^{-6}	2×10^{-6}	3.65×10^{-5}	1.3×10^{-4}	6.5×10^{-5}	4.85×10^{-5}
k=2	1.16×10^{-6}	1.67×10^{-7}	4.43×10^{-5}	1.36×10^{-4}	4.37×10^{-5}	4.51×10^{-5}
k=3	9.5×10^{-7}	2×10^{-7}	5.06×10^{-5}	1.38×10^{-4}	3.2×10^{-5}	4.44×10^{-5}
k=4	8.16×10^{-7}	3.57×10^{-7}	3.5×10^{-4}	1.8×10^{-3}	2×10^{-2}	4.62×10^{-3}
k=5	6.27×10^{-7}	3.54×10^{-5}	1.8×10^{-2}	2.05	23.35	5.08
k=6	4.11×10^{-7}	6.58×10^{-4}	142.53	888.43	21654.76	--

The bolded in the 2~5 columns in Table 1 are optimal interpolation results. Obviously, for the missing data at different locations, the optimal interpolation results correspond to different window sizes (k). For example, the data numbered 23 achieves its best interpolation value when the window size **k**=2 while for No.92 data, the optimal window size **k**=3. Apparently, if a fixed window Lagrange interpolation is used, the effect of interpolation will definitely be affected. The last column shows the average error under the fixed window leaved for the further use. In addition, we can see from the data at Bottom right in Table 1 that some data will be Seriously affected if we use window fixed Lagrange interpolation method, which is result from "Runge" phenomenon. We will study it in the following experiments.

4.2 Comparison of Interpolation Accuracy Using Different Methods

In order to prove the improvement on the performance of the existing interpolation methods, we further do some experiments in this section. We choose part of data of the "Latitude" and "Longitude" in the vehicle positioning data and take MAE values as evaluation rule. Then four kinds of methods has been used to interpolate the missing data respectively. The results have been shown as in Table 2.

Table 2. Comparison of the performance of different interpolation methods

	longitude	latitude
Forward interpolation	2.48×10^{-4}	1.08×10^{-4}
Linear interpolation	1.96×10^{-5}	3.82×10^{-4}
Common Lagrange interpolation (k=5)	1.12×10^{-2}	2.23
Common Lagrange interpolation (k=5)*	4.89×10^{-4}	0.54
Algorithm proposed	1.13×10^{-5}	1.35×10^{-5}

In Table 2, Forward interpolation, Linear interpolation, Common Lagrange interpolation and Algorithm proposed show the performance using Forward interpolation, Linear interpolation, Common Lagrange interpolation and the Algorithm proposed in this paper respectively. What's more, Common Lagrange interpolation (k=5)* means that it has removed the interpolation value affected by the "Runge" phenomenon when calculating the Mean Absolute Error (MAE) in the Common Lagrange interpolation (k=5).

It can be seen from Table 2 that MAE values of the commonly used Lagrange algorithm (k=5) is the largest, because the Lagrange interpolation has "Runge" phenomenon after multiple interpolations, making the filled data oscillate, so its MAE increase by several orders of magnitude. In addition, it can be seen that even if the interpolation value severely affected by "Runge" phenomenon is removed (see common Lagrange interpolation (k=5)*), the MAE value is still relatively large. Moreover, although the accuracy of linear interpolation is very close to the algorithm we have proposed, the linear interpolation method is too singular, that is, not each missing data can be repaired by this method perfectly. And more importantly, the algorithm proposed in this paper has taken linearly interpolation (i.e. equivalent to Lagrange Interpolation when n=k=1) into consideration, which is why the accuracy of the proposed algorithm is higher than the accuracy of linear interpolation. In summary, the performance of the proposed algorithm is optimal compared to the other commonly used interpolation methods. Although the improvement is not as large as several orders of magnitude, it still means a lot to data that requires very high precision, such as vehicle positioning data [10].

4.3 "Runge" Phenomenon

According to the "Runge" phenomenon, when using high-order interpolation polynomial based on equidistant nodes to approximate the "Runge" function, the interpolation polynomial will produce obvious data oscillation, that is, traditional Lagrange interpolation will cause "Runge" phenomenon, which result in data oscillation appearing at both ends of the data interval to be interpolated making the average error of the interpolation much larger. But experiments in this paper have proved that our method is beneficial to mitigate the impact of the "Runge" phenomenon.

The dataset used in this part is sampled from the data "latitude" in the vehicle positioning dataset we have mentioned above. We first show readers the "Runge" phenomenon appearing in the fixed window lagrange interpolation in Fig. 2. and then show the improvement of our proposed algorithm on it in Fig. 3.

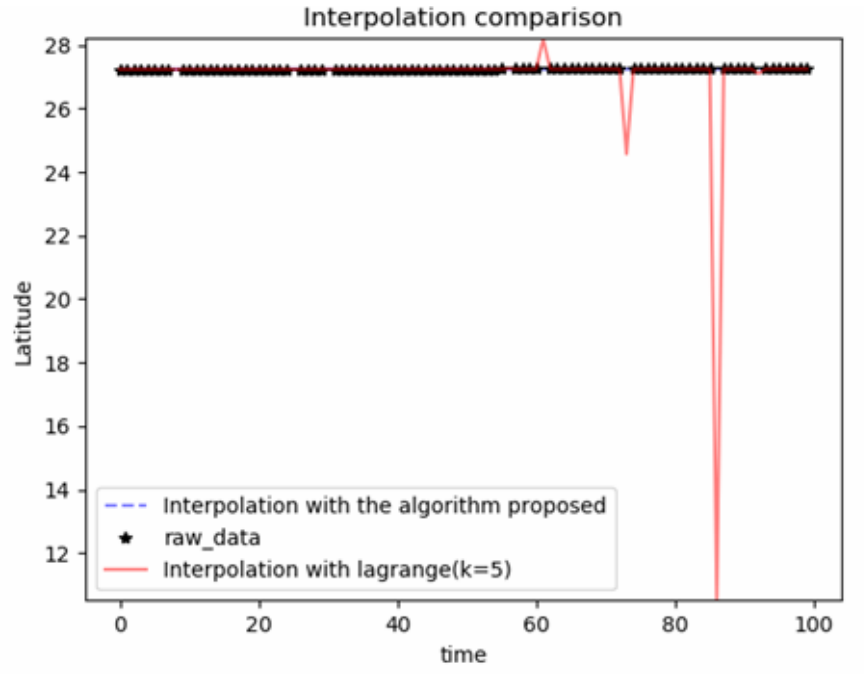

Fig. 2. The "Runge" phenomenon appearing in the Lagrange interpolation with fixed window size

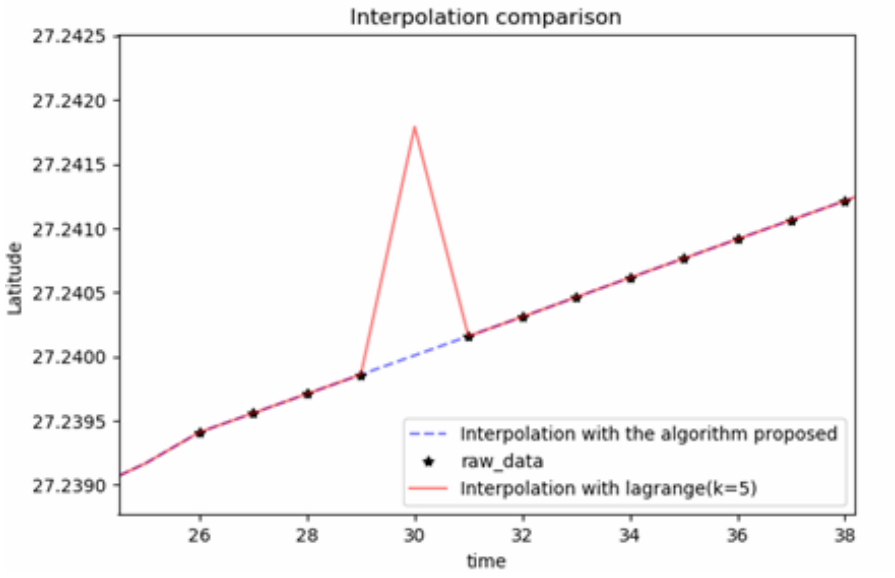

Fig. 3. Comparison of the "Runge" phenomenon before and after improvement (Fig. 3. is a partial enlarged view of Fig. 2.)

"*" identifies the trajectory of the raw data, and the solid line identifies the trajectory of cleaned data after using Lagrange interpolation with fixed window (k=5). As Fig. 1. shows, when we use the window fixed Lagrange interpolation method to repair missing data, data oscillation appears in the end of the data interval, which is caused by "Runge" phenomenon.

"*" identifies the trajectory of the raw data, the solid line identifies the trajectory after using Lagrange interpolation with fixed window (k=5), and the dotted line marks our proposed algorithm. As can be seen from Fig. 3. No. 30 data is obviously missing(*) and the fixed-window Lagrange interpolation value (solid line)obviously does not match on the track of the raw data, which is due to the "Runge" phenomenon. But the algorithm proposed in this paper does not have this problem, which we can see from the dotted line in Fig. 3. Obviously, the filled data using the algorithm proposed in this paper (dotted line) is consistent with the trajectory of the raw data. Therefore, it can be concluded that the algorithm proposed in this paper avoids data oscillation caused by "Runge" phenomenon.

5 Conclusion

Based on the problem that the precision of existing window fixed Lagrange interpolation is not high enough [13], we propose a slope based selection algorithm of optimal window in Lagrange interpolation. By considering the slope calculated by connecting the predicted value with some adjacent values, the algorithm can help Lagrange interpolation to determine an optimal window size for missing values at different positions to find the optimal interpolation value. Experiments have shown that the interpolation accuracy of the algorithm is better than the existing commonly used interpolation methods. In addition, our algorithm proposed also avoid data oscillation caused by the "Runge" phenomenon. Therefore, the algorithm proposed in this paper means a lot in high-precision data restoration such as vehicle positioning data.

Acknowledgements

This research was funded by the National Key Research and Development Program of China, grant number 2017YFC0840200, the Fundamental Research Funds for the Central Universities, grant number 2017JBZ107 and the National Natural Science Foundation of China under grant 61271308.

References

[1] M. Hu, S. Salvucci, A study of imputation algorithms, Working Paper Series 41(3)(2001) 343-371.

[2] R. Jiang, Method to verify and repair road traffic flow data. <http://en.cnki.com.cn/Journal_en/I-I138-JTJS-2006-06.htm>, 2006.

[3] W.-G. Han, J.-F. Wang, J.-J. Hu, Imputation methods for missing values in traffic flow data, Computer & Communications, 2005.

[4] Z.-M. Guo, A.-Y. Zhou, Research on data quality and data cleaning: a survey, Journal of Software 13(11)2002 2076-2082.

[5] F.-J. Feng, J.-P. Yao, X.-S. Li, Research on the data cleaning framework in big data, in: Proc. 2018 2nd International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2018), 2018.

[6] X. Jiang, X.-W. Liu, Knowledge Service-oriented data cleaning in big data, Library & Information 157(5)(2013) 16-21.

[7] S.-E. Madnick, R.-Y. Wang, W.-L. Yang, Overview and framework for data and information quality research, Journal of Data & Information Quality 1(1)(2009) 1-22.

[8] S. García, J. Luengo, F. Herrera, Tutorial on practical tips of the most influential data preprocessing algorithms in data mining, Knowledge-Based Systems 98(2016) 1-29.

[9] C.-K. Enders, Analyzing longitudinal data with missing values, Rehabil Psychol 56(4)(2011) 267-288.

[10] B. Smith, W. Scherer, J. Conklin, Exploring imputation techniques for missing data in transportation management systems, Transportation Research Record Journal of the Transportation Research Board 1836(1)(2003) 132-142.

[11] A. Kaw, Lagrangian Interpolation, University of South Florida, 2009.

[12] L.-J. Zhang, L. Wang, Python Practice of Data Analysis and Mining, China Machine Press, 2016.

[13] Y. Chen, high-order polynomial interpolation based on the interpolation center's neighborhood, in: Proc. the 2009 WRI World Congress on Software Engineering, 2009.