Clinical Time-series Data Interpolation Based on Regression Diagnosis

Masakazu Nakase
Graduate School of Engineering
Doshisha University
1-3 Tataramiyakodani, Kyotanabe-shi,
Kyoto 610-0321, JAPAN
Email: dti0742@mail4.doshisha.ac.jp

Miho Ohsaki1, Shigeru Katagiri1, Yukari Hatada
 Faculty of Science and Engineering
Doshisha University
1-3 Tataramiyakodani, Kyotanabe-shi,
Kyoto 610-0321, JAPAN
Email: {mohsaki1, skatagir1}@mail.doshisha.ac.jp

Abstract—With the development of the clinical database infrastructure, medical knowledge discovery in clinical databases has been actively studied recently. To obtain useful knowledge for medical treatment, it is necessary to apply time-series analysis methods to clinical time-series data. However, such data is nonstationary because of changes in symptom, and the clinical examination intervals of the data are extremely irregular. That makes the application of time-series analysis methods difficult because conventional time-series analysis methods assume stationarity and regularity. Therefore, we herein propose an interpolation method to make the time intervals of clinical time-series data equal based on recursive regression diagnosis and segmentation. We conducted an evaluation experiment using real clinical examination records of hepatitis and compared the performance of the proposed method to that of a conventional interpolation method based on averaging points in each constant period. The results indicated the potential of the proposed method to segment periods of different symptoms and interpolate the points during each period. We mainly confirmed the segmentation performance in this experiment, and the future work will be the estimation of the whole interpolation performance.

I. INTRODUCTION

The development of information technology has made it possible to accumulate clinical examination results as a database. The importance of Evidence-Based Medicine (EBM) has led to an increased need for utilizing clinical data. The further trials were started recently to discover useful knowledge in clinical data to understand symptoms and obtain the hints for medical treatment. Especially for chronic diseases, it is expected that the knowledge discovery in clinical time-series data will play a potent role. The reason is that it is humanly impossible even for experienced medical doctors to directly extract the tendencies of a great deal of multivariate timeseries. In order to obtain knowledge on changes in symptom, time-series analysis should be applied to clinical time-series data.

However, the unfavorable characteristics of clinical timeseries data makes it impossible to apply conventional time-series analysis methods that assume regular sampling. Patient’s clinical examination records basically have irregular intervals, which are caused by irregular visits of the patient for clinical examination and the intensive execution of clinical examination when the patient’s illness becomes worse. It is needed to make the intervals of clinical time-series data equal by interpolating them to for the application of the time-series analysis methods. We therefore propose an interpolation method for this purpose.

The organization of the present paper is as follows. Section II introduces related research and the motivation for the present study. Section III describes the concept and the process of the proposed method. Section IV describes an experiment to evaluate the performance of the proposed method compared to that of a conventional method. Section V presents the conclusion and future research.

II. CONVENTIONAL INTERPOLATION METHODS

Currently, not many studies have examined the interpolation specific to clinical time-series data. One reason is that medical informatics and clinical database infrastructure have been developed over the past two decades, and clinical time-series data have recently become available because they require observations of up to a dozen years. We therefore introduce general interpolation methods of time-series data.

The simplest interpolation methods are averaging or linear regression with a constant period [2], [3]. These methods slide a time window of constant width and calculate the average or the estimate by linear regression on the points in each window. They regard it as an interpolated point, namely a set of an equally spaced time point and the estimated value corresponding to the time point. Consequently, the time intervals are made equal via this process. However, the symptoms are different among patients and nonstationary for each patient, these methods thus have some problems, how the width and overlap of a time window are determined and whether it is proper to fix the width and overlap.

Other methods interpolate missing points using state space model and Kalman filter to more strictly model the generation mechanism [4]. A statistical model such as AR and ARMA is represented in state space, the optimal coefficients of the model are obtained by maximum likelihood, and the missing points are recursively estimated by Kalman filter. There is the following problem to apply such type of methods to clinical time-series data. The selection and fitting of the statistical model, which strongly affect the interpolation performance of
III. PROPOSED METHOD

We first present an outline of regression diagnosis. We then explain the proposed method, which is based on regression diagnosis, in detail.

Regression diagnosis[1] is a method by which to estimate regression model fitness and identify influential observations as outliers. Regression diagnosis consists of three procedures: linear regression modeling, coefficient estimation, and outlier detection. Suppose that we have a set of \( n \) points \((x_i, y_i), i = 1, 2, \ldots, n\), and adopt a simple linear regression. A linear model between the explanatory variable \( x_i \) and the objective variable \( y_i \) is first assumed, as shown in Equation (1). Second, the coefficients \( \beta_0 \) and \( \beta_1 \) are solved as \( \beta_0 \) and \( \beta_1 \) through the calculation of the residual sum of squares \( S_e \) and its minimization by least square. Equation (2) summarizes this process. Note that maximum likelihood estimation derives the same solution in the case of linear regression. Finally, the fitness of the regressed linear function is diagnosed, and outliers, which strongly influence the regression results, are detected.

\[
y_i = \beta_0 + \beta_1 x_i + \varepsilon_i
\]

\[
S_e = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2, \quad \frac{\partial S_e}{\partial \beta_0} = 0, \quad \frac{\partial S_e}{\partial \beta_1} = 0 \tag{2}
\]

Leverage, the most basic criterion of regression diagnosis, estimates the influence of a point \((x_k, y_k)\) on the regression result as an outlier. Leverage is derived through the following process. The deformation of the predicted value \( \hat{y}_i = \beta_0 + \beta_1 x_i \) leads to the summation of the original value \( y_i \) multiplied the weight \( h_{kj} \), as shown in Equation (3). Letting \( j = k \), the Leverage \( h_{kk} \) is obtained as in Equation (4). Equation (4) means that Leverage is defined in terms of the degree of influence per data point \( \frac{1}{n} \) (where \( n \) is the number of data) and the distance of a focused point in the \( X \) direction \((x_k - \bar{x})^2\) (where \( \bar{x} \) is the average of \( x \)). Points with large Leverage values are regarded as outliers.

\[
y_k = \frac{\sum y_i + (x_k - \bar{x}) \sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum h_{kj} y_i \tag{3}
\]

\[
h_{kj} = \frac{1}{n} + \frac{(x_k - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \tag{4}
\]

Cook’s distance[1] is the extension of Leverage and is a popular criterion of regression diagnosis. Cook’s distance additionally reflects the distance of a focused point in the \( Y \) direction, the residual \( y_k - \hat{y}_k \). Equation (5) indicates that Cook’s distance \( D_k \) is defined as the normalized difference with the estimate of variance \( s^2 \), between the estimates \( \hat{y} \) and \( \hat{y}_{mib} \) which are estimated including and excluding \( y_k \), respectively. Here, \( e_k(s) \) is the standardized residual. Cook’s distance can be deformed to consist of three diagnosis criteria, which are the degree of influence per data point \( \frac{1}{n} \), the distance of a focused point in the \( X \) direction \((x_k - \bar{x})^2\), and that in the \( Y \) direction \( y_k - \hat{y}_k \). The larger the value of Cook’s distance (i.e., the smaller the number of data and the further the distances in the \( X \) direction and \( Y \) direction), the more outlying the focused point \((x_k, y_k)\).

\[
D_k = \frac{(\hat{y}_{sub} - \hat{y}) (\hat{y}_{sub} - \hat{y})}{2s^2} \frac{1}{2 - h_{kk}} (e_k(s))^2 \tag{5}
\]

The proposed method makes the intervals of clinical time-series data equal by interpolating missing points with recursive regression diagnosis and segmentation.

The process of the proposed method is described as a conceptual diagram in Figure 1. In the proposed method, a linear function is regressed on the points in a certain period, which is initialized with the entire period of input time-series data. (a) in Figure 1 corresponds to this procedure. On each focused point, if the values of three diagnosis criteria exceed preliminarily determined thresholds, the point is found to be an outlier. After all of the outliers are found, the most outlying point among the outliers is regarded as the boundary of the

---

![Fig. 1. Process of the proposed method (conceptual diagram).](image-url)
Interpolation ( DataX[ ], DataY[ ], Start, End ) {
    Boundary = Regression_and_Segmentation ( DataX, DataY, Start, End );
    ..... (c)
    if ( Boundary == -1 || Boundary == Start || Boundary == End )
      return; ..... (e)
    Interpolation ( DataX, DataY, Start, Boundary ); ..... (d)
    Interpolation ( DataX, DataY, Boundary, End ); ..... (d)
}

Regression_and_Segmentation ( DataX[ ], DataY[ ], Start, End ) {
    if ( The period, End - Start + 1 < Tn )
      return -1; ..... #1
    Regress DataX and DataY in the period; ..... (a)
    for ( i=Start; i<End; i++ ) {
      if ( |DataX[i] - DataX average| > Tx && |DataY[i] - DataY estimate| > Ty ) ..... #2, #3
        Regard the point (DataX[i], DataY[i]) as an outlier;
    }
    Find the most outlying point (DataX[k], DataY[k]) among outliers and return k as a boundary; ..... (b)
}

Fig. 2. Process of the proposed method (pseudo code).

symptom period, as in (b). The clinical time-series data is segmented into two periods at the boundary, as in (c). The same procedures (a), (b), and (c) are recursively executed on the points in the former period, before the boundary, and those in the latter period, after the boundary as in (d). When no outlier is found, the recursive regression and segmentation are stopped, as in (e). Finally, all segmented and regressed periods are obtained. We can easily make the time intervals equal by interpolating missing points in each period using the regressed linear function.

The process of the proposed method is shown as a pseudo code in Figure 2. The i-th data point is expressed as (DataX[i], DataY[i]), where DataX[i] is the number of elapsed days from the first clinical examination and DataY[i] is the value of clinical examination result. The variables Start, End, and Boundary are the indices of DataX and DataY, which represent the start point (DataX[Start], DataY[Start]), the end point (DataX[End], DataY[End]), and the boundary point (DataX[Boundary], DataY[Boundary]). The constants $T_x$, $T_y$, and $T_x$ are the thresholds of the three diagnosis criteria. The method Interpolation is the main body of interpolation and recursively calls the sub-method Regression_and_Segmentation. The procedures explained above are coded as the commands (a) through (e), and the diagnosis based on three criteria are coded as the if statements #1, #2, and #3 in Figure 2.

IV. Evaluation Experiment

A. Experimental Conditions

We conducted an evaluation experiment to examine the performance of the proposed method, as compared to the conventional method, which is based on averaging in each constant period. We used the records of clinical examinations of hepatitis as input clinical time-series data, which was accumulated over approximately 20 years at a university medical school hospital[5]. This dataset consists of the sequential results of 459 types of clinical examinations for 607 patients. We focused on the G-GTP, which is a common clinical examinations used to observe the symptoms of chronic hepatitis. As the first step of the present study, we selected the G-GTP examination histories of 14 patients who had relatively clear symptom changes.

The proposed method requires the preset thresholds of three diagnosis criteria to be determined. In order to mark the optimal thresholds, we investigated various conditions: The threshold of the degree of influence per data point $T_{n} = 1/3$, that of the distance of a focused point in the X direction $T_{x}, 0.01R_{x} \leq T_{x} \leq 0.25R_{x}$ (where $R_{x}$ is the range of X), and that of the distance of a focused point in the Y direction $T_{y}, 0.01s \leq |T_{y}| \leq 0.8s$ (where s is the estimate of residual standard deviation).

On the other hand, the conventional method requires the preset values of the width and overlap of a constant period time window to be determined. Statistical observation of time intervals of the input data revealed that the intervals from 28 days or more to less than 56 days appear most frequently [3]. We then set the width as 56 days and the overlap as 14 days. In the time window, the average of all points is regarded as an interpolated point in the period with no overlap. That is, each period of $56 - 14 \times 2 = 28$ days, where $14 \times 2$ are the former and latter overlaps, is given an interpolated point and becomes the equal time interval.

The measure of performance evaluation is discussed and determined as follows. The frequency distribution of the time interval of the input data has two peaks, namely, a large peak at rank greater than or equal to 28 days and less than 56 days, and a small peak at rank less than 7 days. It is reasonable to assume that routine clinical examinations are conducted once every one or two months during remission and that
intensive clinical examinations are conducted several times a month when the patients condition is worsening. Therefore, if time intervals of less than 7 days occurred at least three times, i.e., approximately one month, then the beginning of the sequence of these intervals can be defined as a switch point of symptoms. We evaluated the performance of the proposed and conventional methods with this measure based on whether each method correctly detected the switch points and segmented periods at them.

B. Results and Discussion

Figure 3 and 4 show examples of experimental results. In both of the figures, the left-side graphs draw the original time-series of G-GTP examination histories and the interpolation result on them by the conventional method. The right-side graphs draw the same by the proposed method. We magnified a part of the upper graphs and plotted on the lower graphs to give in-depth consideration to how the segmentation works. The boundaries between non-colored and gray-colored areas, in other words, the closest points in the gray-colored area to its start or end, are the switch points of symptoms.

Comparing the right upper and left upper graphs of Figure 3, the proposed method seems to properly describe the whole original time-series with a smaller number of functions than that of the conventional method. As in the right lower graph of Figure 3 the proposed method definitely detected the switch points of symptoms if the thresholds of the three diagnosis criteria were properly set. Meanwhile, as in the left lower graph of Figure 3 the conventional method also detected those points, but a certain amount of discrepancies could be accompanied.

The similar tendency as a whole appears in the right upper and left upper graphs of Figure 4 to that of Figure 3. The proposed method seems to outperform the conventional method from the viewpoint of overall description. However, as in the right lower and left lower graphs of Figure 4, the detection performance of the switch points of symptoms, namely the segmentation performance, does not differ between the proposed method and the conventional one. Although the both methods detected the close points to the switch points of symptoms, they could not definitely detect the switch points of symptoms.

To quantitatively estimate the performance, we calculated Recall (the rate of correctly detected switch points per all true switch points) and Precision (the rate of those per all detected switch points). The number of all switch points was 25 in the 14 original time-series used in this experiment. The proposed method detected 284 switch points, and 22 switch points were correct among them. The conventional method detected 1109 switch points, and 20 were correct. Therefore, the value of Recall was $22/25 = 0.88$, and that of Precision was $22/284 = 0.08$ for the proposed method. The value of Recall was $20/25 = 0.80$, and that of Precision was $20/1109 = 0.02$ for the conventional method.

The value of Recall of the proposed method was a little higher than that of the conventional method. It can be implied that the proposed method detects correct switch points with less miss detections. The proposed method showed consid-
erably higher Precision, about four times of that of the conventional method. It is indicated that the proposed method detects correct switch points with much less false detections.

The experimental results indicated the potential of the proposed method to segment different symptom periods. In the experiment, the preset parameters of the conventional method was determined based on the domain knowledge which the most frequent interval was about 28 days. In contrast, those of the proposed method was determined ad-hoc. The proposed method outperformed the conventional method in spite of an adverse condition for the proposed method. Thus, it is expected that the proposed method will achieve a higher performance if the domain knowledge is reflected to the preset parameters.

There remains some issues on the proposed method. The difference in performance was not statistically tested between the proposed method and the conventional methods because of the limited number of clinical examination records used in the present experiment. It was also found that the performance of the proposed method depends strongly on the threshold setting of diagnosis criteria. In the future, we will conduct a statistical performance evaluation by increasing the number of clinical examination records. We will also discuss how to automatically set the thresholds and/or discuss an alternative diagnosis criterion based on the value of Cook’s distance.

V. CONCLUSION

Clinical time-series data has unfavorable characteristics, the nonstationary and irregularity, to apply conventional time-series analysis methods. To solve this problem, we proposed an interpolation method for making the time intervals equal based on recursive regression diagnosis and segmentation. We conducted an evaluation experiment using a dataset of clinical examination records of hepatitis and compared the performance of the proposed method and that of a conventional constant period averaging method. The results indicated the potential of the proposed method to segment different symptom periods and interpolate the points during each period when the thresholds of regression diagnosis criteria are properly set. As the future work, we will statistically estimate and compare the performance of the proposed and conventional methods and consider an automatic setting of the thresholds.

ACKNOWLEDGEMENTS

This research was supported in part by the Academic Frontier Research Project on "New Frontier of Biomedical Engineering Research” and by a Grant-in-Aid for Scientific Research (C), No.19500132 from MEXT, Japan.

REFERENCES