J Inf Process Syst, Vol.10, No.3, pp.395~411, September 2014
http://dx.doi.org/10.3745/JIPS.04.0007

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

Imputation of Medical Data Using Subspace Condition Order Degree Polynomials

Klaokanlaya Silachan* and Panjai Tantatsanawong**

Abstract—Temporal medical data is often collected during patient treatments that require personal analysis. Each observation recorded in the temporal medical data is associated with measurements and time treatments. A major problem in the analysis of temporal medical data are the missing values that are caused, for example, by patients dropping out of a study before completion. Therefore, the imputation of missing data is an important step during pre-processing and can provide useful information before the data is mined. For each patient and each variable, this imputation replaces the missing data with a value drawn from an estimated distribution of that variable. In this paper, we propose a new method, called Newton's finite divided difference polynomial interpolation with condition order degree, for dealing with missing values in temporal medical data related to obesity. We compared the new imputation method with three existing subspace estimation techniques, including the k-nearest neighbor, local least squares, and natural cubic spline approaches. The performance of each approach was then evaluated by using the normalized root mean square error and the statistically significant test results. The experimental results have demonstrated that the proposed method provides the best fit with the smallest error and is more accurate than the other methods.

Keywords—Imputation, Personal Temporal Data, Polynomial Interpolation

1. INTRODUCTION

Temporal medical data are sequences of event values occurring over a period of time. Depending on the measurements of various indicators related to the medical condition being studied, the medical data in a patient case series is collected from the patient's health records and treatments (temporal data). Each event that occurs at each time point has a value that is recorded. The aggregate of all these values represents a single variable (such as a body mass index [BMI] over a time period). However, the dimension of each case series might not be the same as there were the difference in the number of treatments or it was about the time for patients treated. Problems involving temporal data can be analyzed from patterns over time and so on. Therefore, processing information within the time domain in medical research often requires appropriate technical knowledge and an understanding of processing. At the same time, there are numerous pitfalls that come with these benefits; one being that the medical data in a

※ This research was supported by the Higher Education Commission scholarship. The data was from the Cardiovascular and Metabolism Center, Ramathibodi Hospital and thank the reviewers for comments to improve the quality of this paper.

Manuscript received May 14, 2013; first revision October 04, 2013; accepted July 07, 2014.

Corresponding Author: Panjai Tantatsanwong (panjai@gmail.com)

* Department of Computing, Silpakorn University, Nakhon Pathom, Thailand (klao_99@yahoo.com)

** Department of Computing, Silpakorn University, Nakhon Pathom, Thailand (panjai@gmail.com)

time series often contains missing or irregular data. The missing temporal data are significant and may cause major problems, although this has received little attention in the field of medical analysis. Missing data may result from patients not completing treatments as indicated, by patients dropping out of a study before completion, and so on [1]. In addition, if we want a set of temporal medical data for analyzed or mining the data. This missing data can be a problem when analyzed by information processing models. This is due to the fact that many learning algorithms require complete data sets to enable the analysis or mining of data. Therefore, missing attribute values are a very important issue in data analysis. In situations where missing data is likely to occur, the researcher is often advised to plan to use mining data methods.

The imputation of missing data becomes an important method in temporal medical data, where it is crucial to use all of the available data and to not discard records with missing values. Imputing missing values is one of the biggest tasks in pre-processing when performing data mining or data analysis, and it can help produce good quality medical data to provide a complete dataset [2]. The occurrence of missing values can be captured by three types of missing data, which are as follows: missing-completely-at-random, where the gaps in y are independent of both x and y; missing-at-random (MAR), where the gaps in y depend on x but not on y; and non-ignorable, where the gaps in y depend on y and possibly also on x [3-6]. There are many approaches for improving the quality of temporal medical data when information is missing, which are as follows: 1) ignoring objects containing missing values is the easiest and most commonly approach applied; 2) filling in missing values manually; or 3) the imputation of missing values [5-7].

The concept of this paper was to using specific measured data to estimate the missing values in each case series should result in acceptable values [8,9]. Patterns in temporal data can be represented in samples at discrete time points based on various sampling intervals. Some general guidelines for estimating data values with interpolation are as follows: interpolation schemes should model values between and possibly beyond known data points, the principal polynomial should fit with the selected data points, and the polynomial function should be able to reasonably estimate the value of the unknown function [10]. In this work, a combination of polynomial functions and the values associated with the missing polynomial are used to estimate the missing values for each patient.

Our main purpose is to discuss a method for imputing missing data in temporal medical data using Newton's finite divided difference polynomial interpolation with a condition order degree (NFDC). We have proposed this method as a means to estimate values by polynomial interpolation, to determine base time points, and to compensate for missing values in clinical and two-dimensional medical data. To measure the performance of the imputation, we evaluated the accuracy of the approach on a real data set related to obesity and compared it with existing imputation methods.

This paper contains six sections. Section 2 presents a brief overview of the research on the estimation of temporal medical data, while new and traditional polynomial interpolation methods are detailed in Section 3. Section 4 provides the theoretical framework for NFDC, which we compared with the k-nearest neighbor (k-NN), local least squares (LLS), and natural cubic spline methods. Section 5 analyzes the performance of the four imputation methods and we present our conclusions in Section 6.

2. RELATED WORK

Interpolation algorithms aim to predict the value of a variable by using other values of the same variable. The structure of an interpolation model is a function between known points. Therefore, applying the interpolation technique to estimate and impute missing data involves finding an approximation function to estimate values. The known points (x_1, x_2, …, x_n) are called interpolation points. For interpolation with equally spaced data, as the distance between each pair of interpolation points is constant, we write $x_i = x(t_i)$ for i = 1, 2, …, n, that is, ($x(t_1)$, $x(t_2)$, ..., $x(t_n)$). Interpolation will determine the function f(x) at unknown points, while the estimation function will allow the calculation of the function f(x) at desired points for developing the closest estimate. The function f(x) is estimated to determine the values of x, which estimate a smooth curve on the entire domain of the function f(x). Therefore, x is a value between x_1 and x_n in the data set [10-12]. This is because random data requires finding an approximate function to use in place of a more complicated function. This method is distinguished by a degree of continuity [13-15]. Estimating the function used for data collection requires estimating the unknown parameters at each point of interest using values obtained from measurements. If data from a number of different positions (discrete data points) are used, the proposed data will change the results of the continuous function. The function is affected not only by indicating the location of the interpolation, but also by including values at a point between two different positions. This is, therefore, a complete data representation. Recently, several papers have focused on modeling and analyzing temporal medical data. The following sentence is a summary of related work on the principles of estimating missing values. We found no previous research on our proposed finite divided difference method for estimating incomplete temporal medical (clinical) data with different time points in each patient case series.

Noraziana et al. [16] studied effective steps for extracting data that is missing by replacing each value with the mean value of the two points that occur before and after (mean-before-after). Hourly annual inspection records for PM10 in Seberang Perai, Penang, Malaysia, were selected for modeling the missing data (PM107). This analysis was used to estimate six techniques. The techniques used to estimate the missing value were linear, quadratic, and cubic splines and nearest neighbor.

In a study conducted by Bose et al. [17], it was shown that microarray experiments could generate data sets with multiple missing expression values normally, due to various experimental problems. Unfortunately, many algorithms for gene expression analysis require a complete matrix of gene array values as the input. Therefore, the effective estimation of missing values is essential for minimizing the effect of incomplete data sets and to increase the range of data sets to which these algorithms can be applied. In this regard, a new interpolation-based imputation method was proposed to predict missing values in microarray gene expression data. The proposed method selects a subset of similar genes and a subset of similar samples with respect to each missing position, and then applies interpolation in a novel manner to predict a missing value.

Viana et al. [12] focused on the data preparation process used by air-mobile satellites to predict the degradation of the solar array. In particular, the problem of the loss of information due to missing values was addressed.

Jerez et al. [18] presents methods that are based on machine learning techniques for imputation in medical databases. The authors concluded that methods based on machine learning techniques appear to be suited for the imputation of missing values (i.e., for the multi-layer perceptron, self organizing maps, and k-NN). It has been asserted that this has led to a significant enhancement in prognosis accuracy as compared to imputation methods based on statistical estimation, that is, mean values, hot-deck and multiple imputation. The methods were used to impute absent values in the 'El Alamo-I' breast cancer data set, which contains 3,679 records from different hospitals and the Spanish Breast Cancer Research Group (GEICAM) [18].

The analysis by Eisemann et al. [19] is based on the malignant melanoma data set and the female breast cancer data set from the Schleswig-Holstein Cancer Registry in Germany. The cases with complete tumor stage information were extracted and their stage information was partly removed according to an MAR pattern, resulting in five simulated data sets for each cancer entity. The missing tumor stage values were then treated with multiple imputation using chained equations, polynomial regression, predictive mean matching, random forests, and proportional sampling as imputation models. The estimated tumor stages, stage-specific numbers of cases, and survival curves after multiple imputations were compared to the observations. The observed tumor stages on the individual level, the stage-specific numbers of cases, and the observed survival curves were most accurately estimated by polynomial regression and predictive mean matching, while the random forest and proportional sampling models were less accurate.

Tsumoto [20] introduced a combination of the extended moving average method and the rule induction method for continues attribute, called CEARI, to discover new information in temporal databases. CEARI was then applied to a medical dataset related to motor neuron diseases. The data set have many missing values,. In this way, medical physician selected a specific test. They may not take the same test. Missing values will be observed very often in clinical situations.

3. METHODOLOGY

In this section, the proposed imputation model from temporal medical dataset is explained in detail. The temporal data structure is transformed from a low dimension to a subspace with missing data to reduce the dimension of the feature vectors. A subset of the rows and columns of the matrix form the subspace. One must first locate a subspace for a patient case series and fit it to a low-rank matrix. A subspace of the data may be used to compute the missing elements. The subspace matrix has size m × n. The subspaces are different for each patient case series, and even the dimensions of patient case might not be the same. The four imputation techniques are applied to the subspace matrices.

3.1 Data Sets

The data set that we used contains 400 cases with approximately 1,200 records from patients at the Cardiovascular and Metabolism Center, Ramathibodi Hospital with obesity and cardiovascular or heart diseases. The structure of medical temporal data was shown in Table 1. The data structure consists of re-code numbers, patient ID numbers, and treatment dates, as well as the patient's age, sex, height, weight, BMI, basal metabolic rate (BMR), skeletal muscle mass (SMM), percentage of body fat, waist/hip ratio, edema examination, target control, weight control, fat control, muscle control, fitness score, body mass, and protein (g).

Table 1. Structure of medical temporal data

Measurement variable					
X	T_1	T_2	T_3	…	T_n
Body mass index	x	-	x	…	x
Basal metabolic rate	-	-	x	…	x
…	…	…	…	…	…
N	x	x	x	…	x

The Table 2 shows a set of patient visits is defined, where t_i is the i-th visit of a patient j, $T_j = \{t_1, t_2, t_3, \ldots, t_n\}$, where n is the total number of visits. Personal patient data is collected over n visits in an m × n dimension matrix. In the time series $X = \{x_1, x_2, \ldots, x_n\}$, X is the measurement variable, x_i is the recorded value of the measurement variable X at time i, and n is the number of observations. Each event occurring at each time point has a value that was recorded.

Table 2. Structure of obesity medical temporal data

Case-PID	Time	Weight	BMI	BMR	SMM	Protein (g)	X_n	Class label
1	1	x	x	-	x	x	.	1
1	2	x	-	-	x	-	.	1
1	3	-	x	x	-	x	.	1
….	…	…	…	…	…	…	.	..
2	1	x	-	x	x	x	.	0

BMI=body mass index, BMR=basal metabolic rate, SMM=skeletal muscle mass.

Formally, a patient case series in temporal medical data can be represented by PIDxx-X = $\{x_it_1, x_it_2, \ldots, x_it_n\}$.

3.2 Comparison with Other Imputation Methods

In the imputation method for medical temporal data, the data are considered to be in the matrix X with n rows and n columns. The rows are assumed to correspond to entities (observations) and for the columns to correspond to variables (features). The elements of X are denoted by x_{ik} (i = 1, ..., n; k = 1, …, n). The situation in which some entries (i, k) in X are missing is modeled with an additional matrix. The Table 3 shows the summaries of some related method by imputation technique .

Table 3. Summaries of some related method

Method	Imputation technique
Natural cubic spline	Polynomial coefficients
k-NN	Euclidean distance function
LLS	Regression and correlation
NFDC (proposed)	Polynomial function with a condition order degree polynomial

k-NN=k-nearest neighbor, LLS=local least squares, NFDC=Newton's finite divided difference polynomial interpolation with a condition order degree

3.2.1 k-NN

The k-NN imputation method, as developed by Troyanskaya et al. [7], imputes the missing values in an instance of interest by considering a given number of variables. In the k-NN method, variables are imputed using variables that are the most similar. Similarity is measured to calculate the distance. The k-NN method imputes missing values using a selection based on the similarity of the expression. The missing values for an instance are imputed by considering a given number of instances that are the most similar to the instance of interest. This step of k-NN imputation for temporal medical data is used to determine the k closest-neighbors. That is, to find a target similar to the missing values to impute. Euclidean distance metrics are used for this purpose. The problem with temporal medical data for a selected value k is that the data needs to be reliably transformed to a lower-dimensional subspace. In this research, the conditions for selecting k are determined by monitoring the number of treatment visits for each patient's subspace using the Euclidean distance from the number of rows in each of patient subspace [21,22].

3.2.2 LLS

The LLS imputation is comprised of the following two processes: selecting the vector or k variables based on Pearson correlation coefficients, and estimating the missing values by setting up linear regression equations. To estimate the missing values using the most appropriate linear equation, the least square is used, which, when applied to estimating the missing values in temporal medical data, results in the minimum sum of squared distances between the points (x, y) and a line [2,23].

3.2.3 Natural cubic spline

Natural cubic spline interpolation was also selected as an appropriate method for estimating intermediate values between the measured values. Assuming that a collection of known points is: (x_0, y_0), (x_1, y_1), ..., (x_n, y_n), the natural cubic spline estimates the missing values by observing the value of the spline. To interpolate between data points using constrained cubic splines. In general, constrained is well known that are condition function of the natural cubic spline to give only the data point (x_i, y_i), we must determine the polynomial coefficients for each partition so that the resulting polynomials pass through the data points and are continuous in their first and second derivatives. Data points can be represented using a third-order polynomial, $S_i(x) = a_ix^3 + b_ix^2 + c_ix + d_i$, which is constructed on each closed interval $[x_{i-1}, x_i]$. Cubic splines have the following properties: 1) they interpolate the given data; 2) there is continuity of the zeroth, first, and second derivatives at interior points; and 3) they satisfy certain boundary conditions. The natural free boundary condition is the most common. Alternatively, one may use clamped or fixed boundary conditions.

Given n+1 distinct knots x_i such that: $x_0 < x_1 < \ldots < x_n$, and the corresponding value is y_i, a cubic spline function is constructed using a cubic polynomial in each interval, such that the spline and its first and second derivatives are all continuous [24,25]. If there are missing values in the first or last position, extrapolation is used to impute that value.

4. PROPOSED MODEL: NFDC

In this paper, we propose a new method called NFDC. NFDC imputation is based on the idea that using a value of a measured variable for each patient to replace the missing values should result in an acceptable estimate. The imputation method is specified separately for each patient case series. The dimension of each patient case series may differ because of the different numbers of time-treatments or time points for each patient. Therefore, the method for estimating the missing values can be represented as sampling at discrete time events based on various sampling intervals. The function that interpolates the data is an interpolating polynomial. The NFDC method is used to find a function f(x) from data in every given position, such that f(x) interpolates the data. The interpolation degree of f(x) depends on the n data points in the low-rank matrix subspace for each patient.

We next describe the procedure for applying the polynomial to estimate the missing data. Given a temporal set of n data nodes (x_1, y_1), …, (x_n, y_n), the interpolation polynomial in Newton form is chosen to interpolate f(x) at x for estimates using the divided difference table and formula for the set of low-rank matrix subspaces in the patient subspace. A detailed process model developed by the estimation method is used to determine the conditions of the estimators used in the condition order degree. The condition order degree is used to find a polynomial from the observed values in the low-rank matrix for each variable and each patient. This is done by using the highest order degree obtained from the time points for each patient. The process uses value estimation in the degree format of the continuous value of the estimation function at time x(t) during $[xt_1,\dots,xt_{n+1}]$ in a random data pattern. A polynomial of degree n may be written as $f_n(x) = a_0 + a_1x + a_2x^2 +\dots+ a_nx^n$, $a_n = 0$. For any n+1 data points, there is an interpolating degree n polynomial. When the only function consistent with the n+1 data points is the divided difference, the number of points used in the interpolation structure is called the order of the interpolation. Thus, linear interpolation, $y = a_0 + a_1x$, uses two points and is second; quadratic interpolation, $y = a_0 + a_1x + a_2x^2$, uses three point and is third; and cubic interpolation, $y = a_0 + a_1x + a_2x^2 + a_3x^3$, uses four points and is fourth. If the number of time points for a patient receiving treatment is two, then the observed values in the low-rank matrix are 1 and 2. According to the condition of the order degree polynomial, the higher degree is used, in this case 2, to find a polynomial $f_n(x)$, which is later used to estimate the missing value. This method enables for all missing values to be estimated. NFDC uses each occurring data value $X(t_i)_{obs}$ to calculate a polynomial per the temporal estimation principle with the finite difference table. When a polynomial value is achieved, the coefficient was used to calculate for the condition order degree to impute the missing temporal data $X(t_i)_{miss}$, regarding to the position of the value. The value may be before or after the resulting value, which can be in the first position, $f_n(x(t_1))$, or in the last position, $f_n(xt_n)$. There are no additional conditions or methods, and therefore this method will always provide an estimate.

The main concept of NFDC for estimation and imputation is comprised of

1) transforming temporal data to a low dimensional subspace;
2) interpolating the function $f(x_i)$ to find data points for observed values and missing values;
3) imputation in the subspace for each patient case series wherever there is a missing value.

4.1 Proposed Procedure for NFDC Imputation of Temporal Medical Data

The procedure for NFDC imputation of temporal medical data shown in Fig.1. There are seven steps in the proposed method, which are as listed below.

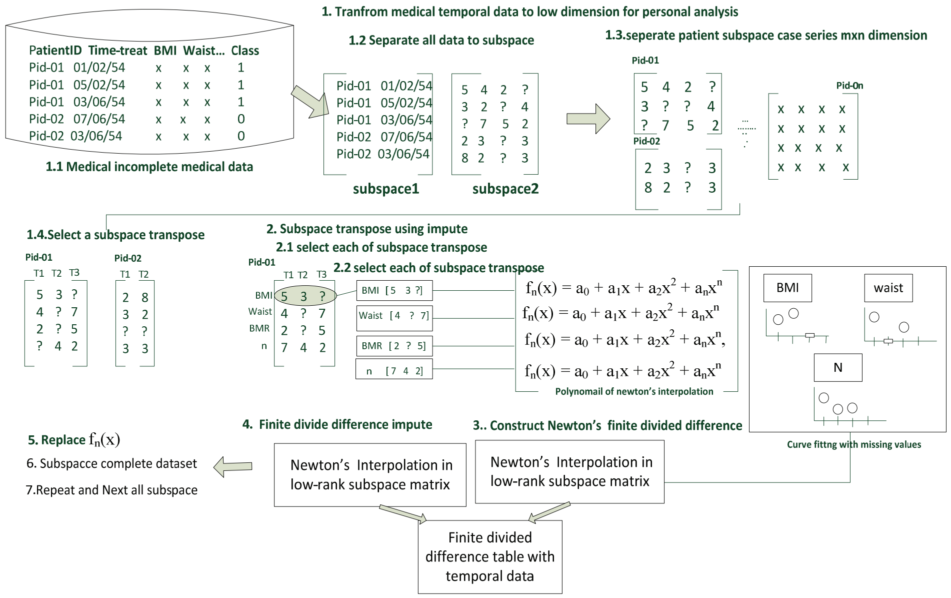

1. Tranfrom medical temporal data to low dimension for personal analysis
PatientID Time-treat BMI Waist... Class
Pid-01 01/02/54 x x x 1
Pid-01 05/02/54 x x x 1
Pid-01 03/06/54 x x x 1
Pid-02 07/06/54 x x x 0
Pid-02 03/06/54 x x x 0
1.1 Medical incomplete medical data
1.2 Separate all data to subspace
Pid-01 01/02/54
Pid-01 05/02/54
Pid-01 03/06/54
Pid-02 07/06/54
Pid-02 03/06/54
5 4 2 ?
3 2 ? 4
? 7 5 2
2 3 ? 3
8 2 ? 3
subspace1
subspace2
1.3.seperate patient subspace case series mxn dimension
Pid-01
5 4 2 ?
3 ? ? 4
? 7 5 2
Pid-02
2 3 ? 3
8 2 ? 3
Pid-0n
x x x x
x x x x
x x x x
x x x x
1.4.Select a subspace transpose
Pid-01
T1 T2 T3
5 3 ?
4 ? 7
2 ? 5
? 4 2
Pid-02
T1 T2
2 8
3 2
? ?
3 3
2. Subspace transpose using impute
2.1 select each of subspace transpose
Pid-01
T1 T2 T3
BMI 5 3 ?
Waist 4 ? 7
BMR 2 ? 5
n 7 4 2
2.2 select each of subspace transpose
BMI [5 3 ?]
Waist [4 ? 7]
BMR [2 ? 5]
n [7 4 2]
$f_n(x) = a_0 + a_1x + a_2x^2 + a_nx^n$
$f_n(x) = a_0 + a_1x + a_2x^2 + a_nx^n$
$f_n(x) = a_0 + a_1x + a_2x^2 + a_nx^n$,
$f_n(x) = a_0 + a_1x + a_2x^2 + a_nx^n$
Polynomail of newton's interpolation
BMI
waist
N
Curve fittng with missing values
3.. Construct Newton's finite divided difference
Newton's Interpolation in low-rank subspace matrix
4. Finite divide difference impute
Newton's Interpolation in low-rank subspace matrix
Finite divided difference table with temporal data
5. Replace $f_n(x)$
6. Subspacce complete dataset
7.Repeat and Next all subspace

Fig. 1. Procedure with steps: Newton's finite divided difference polynomial interpolation with condition order degree (NFDC).

Step 1. Transform temporal medical data to a lower dimension for personal analysis.

1.1. Determine the system's input variables for a temporal data matrix by creating the matrix: A = [x_{ij}], i = 1, …, n, j = 1, …, n. The dataset A = {PIDxx, time, x_1, x_2, …., x_n} is stored as the m × n matrix below, where PID is the patient ID number and x_{ij} are values of the measurement variables.

$$\begin{pmatrix} \text{PID01 time} & x_{i1} & x_{i2} & x_{i3} & \dots x_{in} \\ \text{PID01 time} & x_{i1} & ? & x_{i3} & \dots x_{in} \\ \text{PID01 time} & x_{i1} & x_{i2} & x_{i3} & \dots x_{in} \\ \text{PID02 time} & x_{i1} & x_{i2} & ? & \dots x_{in} \\ \text{PID02 time} & x_{i1} & x_{i2} & ? & \dots x_{in} \end{pmatrix} \quad m \times n$$

1.2. Separate the matrix into patient case series. Each of the subspaces has a different dimension. For example,

$$\underset{2\,\times\,n}{\overset{\textbf{X-PID}_{01}}{\begin{pmatrix} ? & x_{12} & x_{13}\dots x_{1n} \\ x_{21} & x_{22} & x_{23}\dots x_{2n} \end{pmatrix}}} \quad \underset{4\,\times\,n}{\overset{\textbf{X-PID}_{02}}{\begin{pmatrix} x_{11} & x_{12} & ? & \dots & x_{1n} \\ x_{21} & ? & x_{23} & \dots & x_{2n} \\ ? & x_{32} & ? & \dots & x_{3n} \\ x_{41} & x_{42} & ? & \dots & x_{4n} \end{pmatrix}}} \quad \underset{m\,\times\,n}{\overset{\textbf{X-PID}_{0n}}{\begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ ? & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ xnn & xnn & ? & \dots & xnn \end{pmatrix}}}$$

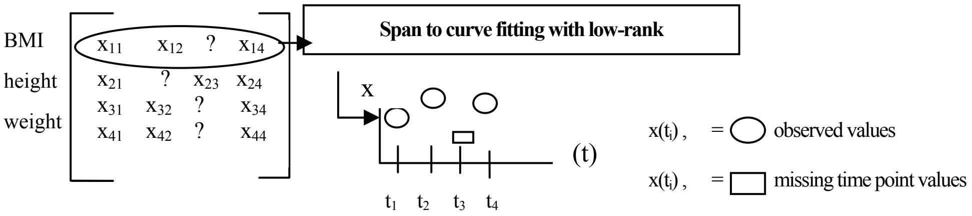
The column vectors in each subspace are sequences of measurement variables (height, BMI, and so on). The rows in each subspace are time points in time treatment.

1.3. Transpose (T) subspaces with missing values.

$$\underset{2\times n}{\textbf{X-PID}_{01}{}^{T}\begin{pmatrix} ? & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \\ \dots & \dots \\ x_{1n} & x_{2n} \end{pmatrix}} \quad \underset{4\times n}{\textbf{X-PID}_{02}{}^{T}\begin{pmatrix} x_{11} & x_{21} & ? & \dots & x_{41} \\ x_{12} & ? & x_{32} & \dots & x_{42} \\ x_{13} & x_{23} & ? & \dots & x_{43} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & ? & \dots & x_{4n} \end{pmatrix}} \quad \underset{m\times n}{\textbf{X-PID}_{0n}{}^{T}\begin{pmatrix} x_{11} & ? & \dots & x_{in} \\ x_{12} & x_{22} & \dots & x_{in} \\ \dots & \dots & \dots & \dots \\ x_{nn} & x_{nn} & ? & x_{nn} \\ x_{in} & x_{in} & \dots & x_{in} \end{pmatrix}}$$

In the transposed patient subspace case series, the column vectors are time points in the treatment, the row vectors are measurement variables (height, BMI, and so on) from patient information.

1.4. To find the missing point values and fit curves in a low-rank matrix, transpose the subspace matrix, which is separated from low-rank matrix. Using the low-rank matrix, locate non-missing and missing values in the subspace.


BMI
height
weight
x11 x12 ? x14
x21 ? x23 x24
x31 x32 ? x34
x41 x42 ? x44
Span to curve fitting with low-rank
X
(t)
t1 t2 t3 t4
x(ti) , = observed values
x(ti) , = missing time point values

Step 2. Subspace transposed using imputation of missing values (incomplete data).

To estimate the missing values using the interpolation polynomial, index the data points starting with $y_i = f(x_i)$, $i = 1, 2, \ldots, n$.

2.1. Generate a subspace data set for a patient case series with partial missing data.

$$\{x(t_1), x(t_2), \ldots.., x(t_n)\}$$

Note that t_i is distinct in the low-rank matrix for subspaces with missing values.

2.2. Select the transposed subspace in each patient case for imputation using missing values.

Step 3. Personal subspaces and constructing Newton's divided difference.

We estimated the missing data value $x(t_i)$ in the low-rank patient subspace matrix by computing Newton's interpolation polynomial $X_{t\text{-}obs}$ with the condition order degree in the recursive divided difference.

3.1. Compute in transposed patient subspace .

3.2. Select rows from the matrix $[x(t_1)\ x(t_2)\ x(t_3) \ldots x(t_n)]$. The row gives measurement values at discrete points in time, $t_1, t_2, t_3, \ldots, t_n$. To find the measurement value at any point in time, use the continuous function y = f(x). This function interpolates the n+1 data points, so that the value of y may be estimated at any other time. For example,

$$\begin{aligned}
&y \text{ at } x(t) = [\ ?\ \ x_{i1}\ \ x_{i2}\ \ x_{i3}\ \ x_{in}]^T\\
&t_i = [1\ \ 2\ \ 3\ \ 4]\\
&x(t_i) = [1\ \ 2\ \ 3\ \ 4]\\
&y(t_i) = [1\ \ 2\ \ 3\ \ 4],
\end{aligned}$$

where x is BMI.T

3.3. Construct Newton's divided difference table to generate divided differences for this set of row vectors in the subspace, and construct f(x) and Newton's divided difference interpolation polynomial, using all rows in the transposed subspace with point values for imputation. Conditions for the order degree for imputing in low-rank subspaces are shown in the divided difference table (Table 4).

Table 4. Finite divided difference table

x_i	y_i	$f(x_i, x_{i+1})$	$f(x_i, x_{i+1}, x_{i+2})$	$f(x_i, x_{i+1}, x_{i+2}, x_{i+3})$	$f(x_i, x_{i+1}, x_{i+2}, x_{i+3}, \ldots x_{i+n})$
x_1	f_1	$F(x_1, x_2)$	$f(x_1, x_2, x_3)$	$F(x_1, x_2, x_3, x_4)$	$F(x_1, x_2, x_3, x_{4,\ldots} x_n)$
x_2	f_2	$F(x_2, x_3)$	$F(x_2, x_3, x_4)$	$F(x_2, x_3, x_4, x_5)$	
x_3	f_3	$F(x_3, x_4)$	$F(x_3, x_4, x_5)$		
x_4	f_4	$f(x_4, x_5)$			

Conditions for order degree imputation in low-rank subspace matrix using the divided difference table with Eqs. (5)–(8).

Definition: The finite divided difference form of equation.

$$b_1 = f(x_1) \tag{1}$$

$$b_2 = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \tag{2}$$

$$b_3 = \frac{\dfrac{f(x_3) - f(x_2)}{x_3 - x_2} - \dfrac{f(x_2) - f(x_1)}{x_2 - x_1}}{x_3 - x_1} \tag{3}$$

$$b_4 = \frac{\dfrac{f(x_4) - f(x_3)}{x_4 - x_3} - \dfrac{f(x_3) - f(x_2)}{x_3 - x_2} - \dfrac{f(x_2) - f(x_1)}{x_2 - x_1}}{x_4 - x_1} \tag{4}$$

Condition with order degree

Observe treatment values at the second time point in rows in the subspace using linear interpolation (second degree):

$$f_1(x)=b_1+b_2(x-x_1). \tag{5}$$

Observe treatment values at the third time point in rows in the subspace using quadratic interpolation (third degree):

$$f_2(x)=b_1+b_2(x-x_1)+b_3(x-x_1)(x-x_2). \tag{6}$$

Observe treatment values at the fourth time point in rows in the subspace using cubic interpolation (fourth degree):

$$f_3(x)=b_1+b_2(x-x_1)+b_3(x-x_1)(x-x_2)+b_4(x-x_1)(x-x_2)(x-x_3). \tag{7}$$

Observe treatment values at other time points in the rows in the subspace using n-th degree interpolation:

$$f_n(x)=b_1+b_2(x-x_1)+b_3(x-x_1)(x-x_2)+...+b_n(x-x_1)..(x-x_{n-1}). \tag{8}$$

3.4. Find a set of observed data points with a direct point (X_{obs}) in the patient subspace to find the missing values, where $y(t_i)_{obs}$, $X(t_i)_{obs}$ is an indexed time point, $y(t_i)_{obs}$ is an observed value in the matrix ($y(t_i)_{obs}$ is f(x) = variable), and $X(t_i)_{obs}$ is an observed time value. Given that the data points are ($t_{i\text{-}obs}$, $y_{i\text{-}obs}$), i = 1, ..., n in the subspace, count the positions in the row with $y(t_i)_{obs}$ at $x(t_i)$. This should contain a set with a feature and no missing values.

$$\begin{array}{ll} t(i) & = [\,1 \;\; 2 \;\; 3\,] \\ X(t_i)_{obs} & = [\,2 \;\; 3 \;\; 4\,] \\ y(t_i)_{obs} & = [\,2 \;\; 3 \;\; 4\,] \end{array}$$

3.5. Find a set with a feature and no missing values ($y(t_i)_{miss}$). Find the missing data point, where t(i) is the time point, y_{miss} is a missing value in the matrix ($y_{(ti)miss}$ is f(x) = estimate value in variable), and $X(t_i)_{miss}$ is a time point with a missing value. If values are found to be missing, the time point t denotes that $y(t_i)_{miss}$ = [?] are missing values.

$$\begin{array}{ll} t(i) & = [\,1\,] \\ X(t_i)_{miss} & = [\,1\,] \\ y(t_i)_{miss} & = [\,?\,] \end{array}$$

Step 4. Calculate Y_{obs}, Y_{miss} and X_{obs}, X_{miss} with finite divided difference imputation.

4.1. Calculate $X(t_i)_{obs}$ for the polynomial from the set of observed value features. The divided differences table, denoted by f[x] = $y(t_i)_{obs}$, is defined recursively by:

$$\begin{aligned} t(i) &= [1\ \ 2\ \ 3] \\ X(t_i)_{obs} &= [2\ \ 3\ \ 4] \\ y(t_i)_{obs} &= [70.7\quad 70.8\quad 75.6] \end{aligned}$$

The finite divided difference in Eqs. (5)–(8).

4.2. Calculate $X(t)_{miss}$ from the set of missing value features:

$$\begin{aligned} t(i) &= [1] \\ X(t_i)_{miss} &= [1] \\ y(t_i)_{miss} &= [?] \end{aligned}$$

Calculate this using the values shown in the divided difference table in the Newton divided difference formula and the Newton form of the interpolating polynomial for interpolating f(x) at $x(t_i)$:

$$f_n(x) = \sum_{i=1}^{n} \left\{ F[x_1, x_2, ..., x_i] \prod_{j=1}^{i-1} \left(x - x_j\right) \right\}, \tag{9}$$

$$\begin{aligned} f_n(x) &= f(x_1) + (x - x_1) f[x_2, x_1] + (x - x_1)(x - x_2) f[x_3, x_2, x_1] \\ &+ \cdots + (x - x_1)(x - x_2) \cdots (x - x_{n-1}) f[x_n, x_{n-1}, \cdots, x_1]. \end{aligned} \tag{10}$$

Consider the condition degree by observing the value of three points using quadratic interpolation, that is, $f_n(x)$, $n = 3$, $x(t_i)_{miss} = 1$:

$$(f_3(x(t_i)) = f_3(1). \tag{11}$$

Step 5. Use the estimate $f_n(x(t_i))$ in the missing position $[X((t_i))_{miss}]$.

Step 6. Repeat steps 3 to 5 for next row in the patient subspace matrix.

Step 7. Repeat for patient subspace_{n+1}, so that results are calculated for each patient subspace that has no missing values.

5. EXPERIMENTS AND RESULTS

The performance of each imputation method was evaluated by comparing the estimated values with the corresponding real values in the complete subspace of a patient case series. The accuracy of each imputation method was evaluated by calculating the error between the observed values and the imputed values after the missing values had been estimated. The efficiency of the missing data estimation system was evaluated using the normal root mean standard error (NRMSE). NRMSE measures the error between real values and estimated values and quantifies the accuracy of the estimated values. The Wilcoxon rank sign significance test was also applied [3,26].

5.1 Rank Metric Normalization in a Complete Set

Min-max normalization performs a linear transformation on the original data to normalize the values of each input data feature vector. Min-max normalization maps a value v from A to v' in the range [new_min$_A$, new_max$_A$], where the range was selected as 1 and 0, that V is the current value of variable x, *v'* is the new value of variable x, and min$_A$ and max$_A$ are the minimum and maximum values of an attribute.

$$v' = \frac{v\text{-}min_A}{max_A\text{-}min_A}\ (new\_max_{A,}\ new\_min_A) + (new\_min_A) \tag{12}$$

The method for ranking metric normalization is as follows:

1) Temporal data rank normalization in a complete set.
2) Random missing values in one incomplete set.
3) Temporal incomplete data rank normalization using four methods for imputation.
4) Evaluation.

5.2 Evaluation of a Patient in a Subspace

5.2.1 NRMSE

NRMSE was used to compare the accuracy of the different imputation methods. NRMSE measures the error between the real values and the estimated values and quantifies the accuracy. This means that a smaller value for the criterion indicates greater accuracy [5]. The NRMSE is the RMSE/root mean square deviation (RMSD) divided by the range of observed values of a variable, in which y$_t$ is the real value, ye$_t$ is the estimated value, and n is the number of missing values. The formula is given by Eqs. (13) and (14).

$$RMSE = \frac{\sqrt{\sum_{t=1}^{n}(y_t - ye_t)^2}}{n} \tag{13}$$

$$NRMSE = \frac{RMSE}{X_{\max} - X_{\min}} \tag{14}$$

In analyzing the missing data, the cases are first ordered according to data size. The missing dataset is then divided into four parts with different percentages of the missing data. The percentage of missing data is a proportion of the mean data size, that is, p-missNo = (Row*Col*pct_mis)/100, where p-missNo% is the total missing data percentage. For each missing data pattern and each missing data mechanism, five different percentages of missing values (10%, 15%, 20%, 25%, and 30%) were used randomly. The results of the experiment are shown in Table 5 and Fig. 2.

5.2.2 Statistical significance

We tested the significance of all missing probabilities. Each result was compared using the Wilcoxon signed-rank statistical significance test (W). The Wilcoxon significance test was performed to test the validity of the different imputation methods for temporal obesity data at the statistical significance level of 0.05. Under the null hypothesis, the HL statistic suggests evidence of a lack of model fitting (a p-value much greater than 0.05 would indicate very good model, while $p < 0.05$ reveals a poor model). The results were significant for a higher percentage of missing values. [26,27].

5.3 Experimental Results and Discussion

Table 5 and Fig. 2 show the experimental results. The performance was measured with respect to the percentage of missing values in the data set. The accuracy of data was computed using the NRMSE. There were four methods used to estimate missing values, which were then compared with the real data set for the temporal obesity data. These pieces of data had 10%, 15%, 20%, 25%, or 30% of the values missing. NFDC was compared with three existing subspace estimation techniques: k-NN, LLS, and natural cubic splines. The results are shown in Fig. 2. The results confirm the good performance of NFDC for estimating missing values. Of the four methods, NFDC method gives the smallest error. The average accuracy values for the methods are 14.4, 19.0, 39.4, and 358.2 for NFDC, cubic splines, k-NN, and LLS, respectively.

Table 5. Results of the proposed model and original methods as shown in percentages of missing value

Method	Percentage of missing values					Average
	10%	15%	20%	25%	30%	
NFDC (proposed method)	10.4931	15.0097	13.1893	16.2048	20.5221	14.4830
Natural cubic spline	12.2079	19.0056	17.4129	22.5382	24.1026	19.0535
k-NN	31.623	39.017	40.6742	42.3140	43.6030	39.4463
LLS	349.591	350.237	354.4324	367.5914	369.3242	358.2352

NFDC=Newton's finite divided difference polynomial interpolation with a condition order degree, k-NN=k-nearest neighbor, LLS=local least squares.

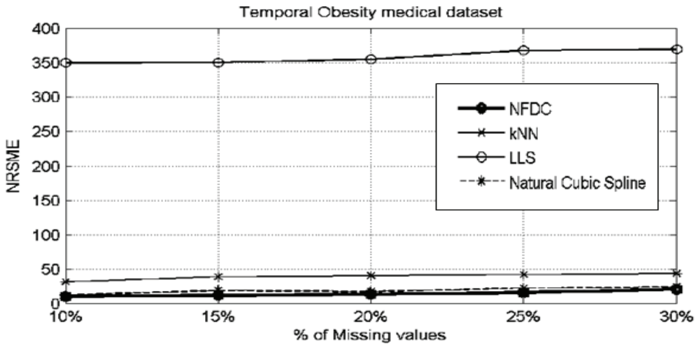


Fig. 2. Comparison of normal root mean standard error (NRMSE) with missing values for the four methods. NFDC=Newton's finite divided difference interpolation with a condition order degree, k-NN=k-nearest neighbor, LLS=local least squares.

Table 6. Average ranking of the *p*-values of temporal obesity data

Method	Percentage of average ranking values				
	10%	15%	20%	25%	30%
NFDC (proposed method)	0.6285	0.5900	0.4829	0.5129	0.4595
Natural cubic spline	0.5064	0.4801	0.4653	0.5111	0.4598
k-NN	0.0940	0.0164	0.0509	0.0516	0.5233
LLS	0.3427	0.3646	0.3520	0.3536	0.5508

NFDC=Newton's finite divided difference polynomial interpolation with a condition order degree, k-NN=k-nearest neighbor, LLS=local least squares.

Table 6 provides an average ranking of the *p*-values obtained using W under the hypothesis that the estimated value of the proposed method has acceptable accuracy. The significance of this finding is that NFDC from the data size (10% and 30%) when compared with the existing methods. The performance of NFDC is significantly better than that of the other methods.

6. Conclusion

This paper presents a new method known as, NFDC, which is based on the concept of estimations in each patient case series from temporal obesity data. NFDC, which uses polynomial interpolation, creates a new data point between given data points in each patient's medical indicators related to the number of real time points. The estimation of the missing values is obtained by using the observed data to impute values in positions where data are missing with a condition order degree. Under the condition order degree, which is considered according to the number of observed treatment values, it can still be processed based on interpolation. Missing values at any time point in the temporal obesity data can be estimated in every subspace without other conditions or methods. The experimental results were evaluated using the NRMSE and Wilcoxon statistical significance test. Our proposed method has a small estimation error as compared with existing subspace estimation techniques, including k-NN, LLS, and natural cubic splines. The NRMSE results show that NFDC has the smallest amounts of errors and the highest significance rate.

References

[1] C. M. Antunes and A. L. Oliveira, "Temporal data mining: an overview," in *KDD 2001 Workshop on Temporal Data Mining*, San Francisco, CA, August 26, 2001.

[2] X. Zhang, X. Song, H. Wang, and H. Zhang, "Sequential local least squares imputation estimating missing value of microarray data," *Computers in Biology and Medicine*, vol. 38, no. 10, pp. 1112-1120, Oct. 2008.

[3] A. R. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons, "Review: a gentle introduction to imputation of missing values," *J Clinical Epidemiology*, vol. 59, no. 10, pp. 1087-1091, Oct. 2006.

[4] A. Sahu, T. Swarnkar, and K. Das, "Estimation methods for microarray data with missing values:a review," *International Journal of Computer Science & Information Technologies*, vol. 2, no. 2, pp. 614-620, Mar. 2011.

[5] B. Mehala, P. Ranjit Jeba Thangaiah, and K. Vivekanandan, "Selecting scalable algorithms to deal with missing values," *International Journal of Recent Trends in Engineering*, vol. 1, no. 2, pp. 80-83, May 2009.

[6] K. Raja, G. Tholkappia Arasu, and C. S. Nair, "Imputation framework for missing values," *International Journal of Computer Trends and Technology*, vol. 3, no. 2, pp. 215-219, 2012.

[7] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B.

Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520-525, Jun. 2001.

[8] J. F. Roddick and M. Spiliopoulou, "A survey of temporal knowledge discovery paradigms and methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 4, pp. 750-767, Jul. 2002.

[9] M. H. Dunham, *Data Mining Introductory and Advanced Topics*. Upper Saddle River, NJ: Prentice Hall/Pearson Education, 2003.

[10] M. Dvornikov, "Spectral properties of numerical differentiation," *Journal of Concrete and Applicable Mathematics*, vol. 6, no. 1, pp. 81-89, Jan. 2008.

[11] J. M. Jerez, I. Molina, J. L. Subirats, and L. Franco, "Missing data imputation in breast cancer prognosis," in *Proceedings of the 24th IASTED International Conference on Biomedical Engineering*, Innsbruck, Austria, 2006, pp. 323-328.

[12] N. Viana, A. Pereira, R. Ribeiro, and A. Donati, "Handling missing values in solar array performance degradation forecasting," in *Proceedings of the 15th Mini-EURO Conference on Managing Uncertainty in Decision Support Models*, Coimbra, Portugal, September 22-24, 2004.

[13] D. N. Varsamis and N. P. Karampetakis, "On a special case of the two-variable Newton interpolation polynomial," in *2nd International Conference on Communications, Computing and Control Applications*, Marseilles, France, December 6-8, 2012, pp. 1-6.

[14] J. B. Scarborough, *Numerical Mathematical Analysis*, 6th ed. Baltimore, MD: Johns Hopkins Press, 1966.

[15] K. E. Atkinson, *An Introduction to Numerical Analysis*, 2nd ed. New York, NY: Wiley, 1989.

[16] M. N. Noraziana, Y. A. Shukric, R. N. Azamc, and A. M. M. Al Bakrib, "Estimation of missing values in air pollution data using single imputation techniques," *ScienceAsia*, vol. 34, no. 3, pp. 341-345, 2008.

[17] S. Bose, C. Das, S. Dutta, and S. Chattopadhyay, "A novel interpolation based missing value estimation method to predict missing values in microarray gene expression data," in *International Conference on Communications, Devices and Intelligent Systems*, Kolkata, India, December 28-29, 2012, pp. 318-321.

[18] J. M. Jerez, I. Molina, P. J. Garcia-Laencina, E. Alba, N. Ribelles, M. Martin, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105-115, Oct. 2010.

[19] N. Eisemann, A. Waldmann, and A. Katalinic, "Imputation of missing values of tumour stage in population-based cancer registration," *BMC Medical Research Methodology*, vol. 11, p. 129, Sep. 2011.

[20] S. Tsumoto, "Rule discovery in large time-series medical databases," in *Principles of Data Mining and Knowledge Discovery, Lecture Notes in Computer Science Volume 1704*, J. Żytkow and J. Rauch, Eds., Heidelberg: Springer Berlin, pp. 23-31, 1999.

[21] E. Acuña and C. Rodriguez, "The Treatment of Missing Values and its Effect on Classifier Accuracy," in *Classification, Clustering, and Data Mining Applications*, D. Banks, F. McMorris, P. Arabie, and W. Gaul, Eds.ed: Springer Berlin Heidelberg, 2004, pp. 639-647.

[22] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, Jan. 1967.

[23] T. H. Bo, B. Dysvik, and I. Jonassen, "LSimpute: accurate estimation of missing values in microarray data with least squares methods," *Nucleic Acids Research*, vol. 32, no. 3, p. e34, Feb. 2004.

[24] C. De Boor, A Practical Guide to Splines, Applied Mathematical Sciences Volume 27. New York. NY: Springer-Verlag, 1978.

[25] G. Walberg, "Cubic spline interpolation: a review," Department of Computer Science, Columbia University, New York, NY, Technical Report CUCS-389-88, 1988.

[26] B. Rosner, R. J. Glynn, and M. L. Lee, "The Wilcoxon signed rank test for paired comparisons of clustered data," *Biometrics*, vol. 62, no. 1, pp. 185-192, Mar. 2006.

[27] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*. New York, NY: Wiley, 1973.

Klaokanlaya Silachan

She received a M.S. of science degree in technology of information system management from Mahidol University. She started her teaching career in 1994 and has been working as a full-time lecturer in the Computer Department at Nakorn Pathom Rajaphat University, Thailand. She is currently pursuing her Ph.D. in Computer Science and information at the Department of Computing, Silpakorn University, Thailand. Her research interests are in the areas of data mining technique, neural network, and computational information processing.

Panjai Tantasanawong, Ph.D.

He obtained his doctoral degree in computer science from Asia Institute of Technology (AIT), in Thailand, in the year 2000. His major area is in networking and software engineering. He received his M.S. degree in Computer Science from Chulalongkorn University, Bangkok, Thailand in 1992 and B.S. degree in Public Health from Mahidol University, Bangkok, Thailand in 1984, respectively. Currently, he is the Associate Professor at the Department of Computing, Faculty of Science, Silpakorn University, Bangkok, Thailand. He has publications in national and international conference proceedings and academic journals.