Computational Statistics and Data Analysis 112 (2017) 63–79

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Imputing missing values in unevenly spaced clinical time series data to build an effective temporal classification framework

Jane Y. Nancy[a], Nehemiah H. Khanna[a,*], Kannan Arputharaj[b]

[a] Ramanujan Computing Centre, Anna University, Chennai-600025, India

[b] Department of Information Science and Technology, Anna University, Chennai-600025, India

HIGHLIGHTS

- Imputing missing values in unevenly spaced clinical time series data.
- Tolerance rough set induced bio-statistical (TRiBS) framework is used in imputation.
- TRiBS adopts and improve inverse distance weight (IDW) interpolation technique.
- TRiBS uses the tolerance rough set and particle swarm optimization to improve IDW.
- Performance of the imputed results proves the effectiveness of TRiBS.

ARTICLE INFO

Article history:
Received 2 November 2015
Received in revised form 10 November 2016
Accepted 16 February 2017
Available online 24 February 2017

Keywords:
Time series
Missing value
Tolerance rough set
Particle swarm optimization
Inverse distance weight

ABSTRACT

BACKGROUND: In healthcare domain, clinical trials generate time-stamped data that record set of observations on patient health status. These data are liable to missing values since there are situations, where the patient observations are neither done regularly nor updated correctly.

OBJECTIVE: This paper aims to impute missing values in an unevenly spaced clinical time-series data by proposing a tolerance rough set induced bio-statistical (TRiBS) framework. The proposed framework adopts an inverse distance weight (IDW) interpolation technique and improves it using the concept of tolerance rough set (TR) and particle swarm optimization (PSO).

METHOD: To interpolate an unknown data point, the classical IDW interpolation suffers from two major drawbacks: first, in selecting the known data points and second, choosing an optimal influence factor. TRiBS framework overcomes the first limitation using TR and the second using PSO. TR derives the dependent attributes for each attribute using non-missing records. The nearest significant set is then generated for each missing value based on its attribute dependencies. The PSO technique fixes the weights for the data in a nearest significant set by finding an optimized influence factor. The obtained significant set and its influence factor are used in IDW computations to impute missing value.

RESULT: The proposed work is experimented using clinical time series dataset of hepatitis and thrombosis patients. However, the proposed system can support other clinical time series dataset with minor domain specific changes.

CONCLUSION: The performance of the imputed results proves the effectiveness of TRiBS. Experimental evaluation with the classifiers such as neural networks, support vector

[*] Corresponding author.
E-mail addresses: nancy@annauniv.edu (J.Y. Nancy), nehemiah@annauniv.edu (N.H. Khanna), kannan@annauniv.edu (K. Arputharaj).

http://dx.doi.org/10.1016/j.csda.2017.02.012

machine (SVM) and decision tree have shown an improvement in the classification accuracy when a missing data is pre-processed with the proposed framework.

1. Introduction

The impact of missing data and its management has been studied in several research studies (Little and Rubin, 2014; Enders, 2010; Van der Heijden et al., 2006; Scheuren, 2005; Schafer, 1997; Dempster et al., 1977). The occurrence of missing data is obvious in many real-life applications where there is periodic record maintenance. Treating these missing values is considered as a vital task, since it improves the effectiveness of knowledge discovery process (Enders, 2010; Ford, 1983). In healthcare domain, clinical data are liable to have missing values since the observations are done for each patient at irregular intervals and the number of observations done varies for every patient. Missing data can be classified into two categories based on its pattern and relationship between observed variable with missing data (Little and Rubin, 2014; Enders, 2010). First category corresponds to six patterns, namely univariate pattern, unit non-response pattern, monotone pattern, general pattern, planned missing pattern and latent variable pattern. Second category classifies missing data as missing at random (MAR), missing not at random (MNAR) and missing completely at random (MCAR) (Little and Rubin, 2014). The two common strategies for handling missing values are ignorance (deletion) and imputation (Enders, 2010). There are several missing value imputation techniques, namely mean, median, nearest neighbour, hot-deck, maximum likelihood, regression (Little and Rubin, 2014; Enders, 2010; Dempster et al., 1977; Ford, 1983).

The applicability of these missing imputation techniques in non-time series data differs from time series data, due to the presence of temporal patterns like trend, seasonal, cyclic and irregular variations in time series data. Several research works have been carried out to illustrate the importance of imputing the missing values in time series (Enders, 2010). Clinical time series data are characterized by the temporal patterns, which identify the change in the temporal sequence of observed patient's lab examinations for a particular disease. Thus, missing value imputation in clinical time series data becomes challenging when the observations are done irregularly.

1.1. Outline of the paper

This paper proposes a TRiBS framework for imputing missing values in an unevenly spaced clinical time series data. TRiBS adopts and improves the classical IDW presented by Shepard (1968) using two key concepts, namely the tolerance rough set (TR) analysis and particle swarm optimization (PSO). TR analysis identifies similar records, which forms the significant set. The PSO technique finds the influence factor value for fixing the weights of known data included in the significant set. The significant set and the influence factor are then used in the IDW process to derive the interpolated values. These interpolated values impute the missing values in clinical time-series dataset. The performance measures such as mean absolute deviation (MAD), mean absolute percentage error (MAPE), root mean squared error (RMSE), fractional bias error (FB) and index of agreement (IA) derived from experimental analysis prove the efficiency of the proposed framework. Classification on TRiBS imputed data using classifiers such as neural network, support vector machine and decision tree shows an improved accuracy.

The rest of the paper is organized as follows. The Section 2 discusses the related works. In Section 3 Materials and methods used in the proposed framework are discussed. Experimental results and discussions are presented in Section 4. Conclusion and scope for future work are presented in Section 5.

2. Related work

This section reviews the work carried out by the researchers in missing value imputation using statistical and machine learning techniques.

2.1. Statistical techniques

Ford (1983) has presented a hot-deck imputation method in which the available complete records act as donors for the records that contain missing values. This method attempts to impute missing values from the observed values with similar pattern, hence it is also termed as similar response pattern imputation. Andridge and Little (2010) have provided a comprehensive review about the various versions of hot deck imputations and its applications. Perez et al. (2002) in their work have illustrated the usage of various imputation techniques like mean, hot deck and multiple imputation for predicting the outcome in the Intensive care units (ICU).

Mean imputation method (Van der Heijden et al., 2006) can be conditional or unconditional. In unconditional imputation, the overall mean of the attribute corresponding to the missing value from the observed dataset is used to impute the missing values. Conditional mean replaces the missing values with the mean of the specific subgroup to which it belongs. If the attributes are categorical, then the missing data are replaced by its modal value.

Srebotnjak et al. (2012) employed the Hot-deck method to impute water quality index. The results show that the hot deck method is more applicable as they preserve the distribution of the data than mean imputation. Sullivan and Andridge (2015) have developed a non-ignorable proxy pattern mixture hot deck multiple imputation method which suits all types of missingness. The authors have combined the ideas of hot deck imputation using distance-based donor selection and a parametric non-ignorable imputation procedure which are based on the assumption of MNAR. A donor quality metric named minimum mean distance has been proposed and a sensitivity parameter was used to identify the missingness mechanism. Inverse distance weight (IDW) interpolation presented by Shepard (1968) finds the missing attribute value using the weighted average of known values from the considered neighbourhood. IDW suffers from many drawbacks which include the distance calculation, selection of neighbourhood and its weight factor. To overcome these limitations several enhancements related to the distance computation and influence factor optimization for the selected known points were incorporated to the traditional IDW (Lu and Wong, 2008; Cressman, 1959; Gandin, 1970; Barnes, 1964; Sen and Sahin, 2001).

Little and Rubin (2014) have presented multiple imputation method for handling missing data. The process involves three steps, first the missing values are imputed multiple '*n*' times in order to represent the uncertainty in identifying which value to impute, second step involves the analyses of the '*n*' complete datasets and as the third step the results are combined to yield a single estimate, e.g., *p*-values, standard errors, regression coefficients that incorporate missing data uncertainty. Van der Heijden et al. (2006) have presented a work that illustrates the effectiveness and impact of handling missing data in clinical diagnostic studies. For handling the missing data the authors have considered the following methods, namely missing-indicator method, complete case analysis, multiple imputation, single imputation of unconditional and conditional mean. For experimentation the authors have considered diagnostic patient data for pulmonary embolism. The experimental results show that imputing missing data improves the prediction results compared to the complete case analyses and missing-indicator method. It has been observed that multiple imputations attain effective results, but when there is a low number of missingness the single imputation method works effectively.

In regression imputation (Enders, 2010), a regression model is constructed using the complete observations. This model is used to predict the value of the missing data. Expectation maximization (EM) method is a two-step iterative process proposed by Dempster et al. (1977) in which a complete dataset is obtained by imputing the missing values by performing the E-steps (expectation) and the M-steps (maximization) repeatedly until convergence occurs. Initially the E-step computes the expected value of the sum of missing data variables with an assumption that the value for the population mean and variance–covariance matrix are known. The expected value of the sum of a variable is used in the M-step to estimate the population mean and covariance. The process iterates until the values of the estimates do not change. The major advantage of the EM method over mean and hot deck imputation is that it preserves the relationship between the variables, whereas mean and hot-deck imputations reduce the variance and the absolute value of the covariance.

Full information maximum likelihood (FIML) is a model based approach very much related to EM method where the parameters estimated in both the methods are almost identical. In this approach missing data and parameter estimation are handled in a single step (Enders and Bandalos, 2001). It is also referred to as raw-data maximum likelihood, which reads in the raw data one case at a time, and maximizes the maximum likelihood function one case at a time; finally the results are combined to produce an overall estimate of the maximum likelihood function. The FIML produces the estimated parameters, but does not impute missing data. Olinsky et al. (2003) presented a comparative study that illustrates the efficacy of mean imputation, regression imputation, multiple imputation, EM, maximum likelihood and FIML imputation for missing data in structural equation modelling. The results indicated FIML to be far superior to other methods and could be used for the determination of parameters in structural equation modelling, however, if a complete set of imputed data is required then maximum likelihood is found to be superior.

Strike et al. (2001) have evaluated the performance of mean imputation, hot deck imputation and listwise deletion with respect to bias and precision for software cost modelling. The results show that listwise deletion is suitable when there is less percentage of missingness. However the precision shows variation when the missingness increases or is non-ignorable. In the case of mean imputation, the precision tends to be higher and bias is nearly zero for MAR and MCAR but accuracy decreases for non-ignorable missing data. Hot deck imputation with z-score standardization and Euclidean distance outperforms listwise deletion and mean imputation as it shows less bias and higher precision even for non-ignorable missingness.

2.2. Machine learning techniques

In machine learning techniques, a data model is built from the complete data available for each of the attribute and later the constructed model is used to predict the missing values. The methodology remains the same for several supervised machine learning algorithms such as decision trees, probabilistic, and association rules (Farhangfar et al., 2007). K-nearest neighbour method imputes missing values from the computed nearest neighbour of suitable distance.

Artificial neural networks (ANN) extract patterns hidden in the data and hence find its application in imputing missing values (Farhangfar et al., 2007; Nordbotten, 1996; Silva-Ramírez, 2011; Junninen et al., 2004; Gheyas and Smith, 2010). Nordbotten (1996) has presented a work that uses an ANN to impute missing values. The training model is built using non-missing records and the trained model is used for imputing missing records. Silva-Ramírez et al. (2011) developed a multilayer perceptron method of imputation for handling data which are missing completely at random. The performance of this method was almost similar to other methods like hot deck, KNN, mean/mode and regression for quantitative variables but it outperformed the others for categorical variables. Junninen et al. (2004) have presented an evaluation study on missing data imputation using air quality datasets. Missing patterns were simulated and the dataset was evaluated with regression-based imputation, self-organizing map (SOM), linear, spline and nearest neighbour interpolation, multivariate nearest neighbour, multi-layer perceptron (MLP) and hybrid methods. The comparison results illustrate that SOM and MLP outperform the multivariate nearest neighbour and other techniques. Gheyas and Smith (2010) have proposed two missing value imputation algorithms based on an ensemble model of generalized regression neural networks (GRNN) namely GRNN-ensemble for multiple-imputation and GRNN-ensemble for single imputation. The key advantages of these algorithms lie in the fact that they are local approximators and are non-parametric algorithms which avoid assumptions made on distributions.

In decision tree method the missing variable is treated as the target and the remaining variables are used as predictors. Decision tree imputation employs learning algorithm such as ID3 to build a decision-tree classifier using the rows which are complete. The decision tree rule is then applied on the row with missing value to predict the missing value. Rahman and Islam (2013) have presented two techniques for imputing missing values using decision trees and decision forests. These techniques form segments in the dataset which contains records with high similarities and attribute correlations. Missing values are then imputed based on the similarity between the segments and missing record. The precision of the imputed results is increased because the authors have considered all the attributes based on their correlation strength.

Bayesian Principal Component Analysis (BPCA) is a probabilistic model based on Bayesian principal component analysis for estimating the missing values (Oba et al., 2003). This method consists of three processes: principal component regression, Bayesian estimation and EM algorithm. The method divides the dataset into complete and incomplete and represents the dataset in the form of matrix. Bayes theorem is used to calculate the PCA and the Bayesian estimation calculates posterior distribution of model parameter and input matrix containing samples. The iterative EM method is employed to compute the unknown parameter.

Cismondi et al. (2013) have employed a fuzzy logic based classification model which identifies whether the missing values are recoverable (dependent) or non-recoverable (independent) based on the dependency of the missing values with other variables. According to the classification result, dependent values are imputed using mean imputation and independent values are deleted. Ding and Ross (2012) have proposed a gaussian mixture model based K-nearest neighbour method which outperformed methods based on likelihood, regression, Bayesian and multiple imputation for handling missing scores in biometric fusion. Qu et al. (2009) have presented a probabilistic principal component analysis for imputing missing values in time series data that integrates principal component analysis and maximum likelihood estimation. Principal component analysis extracts the dominant parts and the maximum likelihood estimation imputes the missing values using the likelihood function derived from the sampled data.

Jerez et al. (2010) have presented an analysis of various missing data imputation methods that incorporates statistical and machine learning methods. Statistical methods included multiple imputation method, hot deck methods and mean methods. Machine learning methods included Multilayer perceptron, K-nearest neighbour and SOM. The authors have concluded that the machine learning methods outperformed statistical methods with significant reduction in the error rates.

Comparing to the works discussed in the literature the proposed work is different in the following ways:

Most of the imputation techniques discussed in the literature can work effectively with the time series data that are regular. Since clinical time series data are considered to be unevenly spaced, the direct application of these techniques may degrade their performance. Hence, this work proposes an enhancement to the IDW interpolation by using the concept of TR analysis and PSO for missing value imputation. The traditional IDW has two major limitations in choosing the number of known data points (nearest neighbourhood) and finding the optimal value for influence factor parameter. The proposed work overcomes these limitations by using the TR analysis concept for selecting the relevant known points and PSO techniques to find the optimal value for influence factor parameter. From the experimental results it has been inferred that the proposed framework has effectively reduced the error rate and improved the accuracy of the imputed results.

3. Materials and methods

This section describes the datasets used for experimentation and the methods used in the proposed framework.

3.1. Dataset description

For experimentation, this work uses two time series clinical data of hepatitis and thrombosis patients. The datasets were released in Principles and Practice of Knowledge Discovery in Databases (PKDD) discovery challenge for a data mining contest (Hepatitis, 2005; Thrombosis, 1999). These datasets were collected from Chiba hospital between the years 1982 to 2002 and 1980 to 1992 respectively. Hepatitis dataset (Hepatitis, 2005) contains laboratory examination reports of 771 hepatitis B and C patients. In hepatitis dataset, each patient has undergone 983 laboratory tests. It has to be noted that not all the laboratory tests taken relate to hepatitis. Hence by considering the expert guidance, the dataset descriptions given with the dataset and by Ohsaki et al. (2002), 29 suggested tests have been selected for experimentation. Out of 29 attributes 11 attributes are identified to have missing values. Thrombosis dataset consists of 1000 patient's laboratory examination

Table 1
Dataset summary.

Dataset	Total records	Lab tests	Suggested lab test (by expert)	Total patients	Average missing value (%)
Hepatitis	1 565 876	983	29	771	11
Thrombosis	57 543	564	33	1000	8

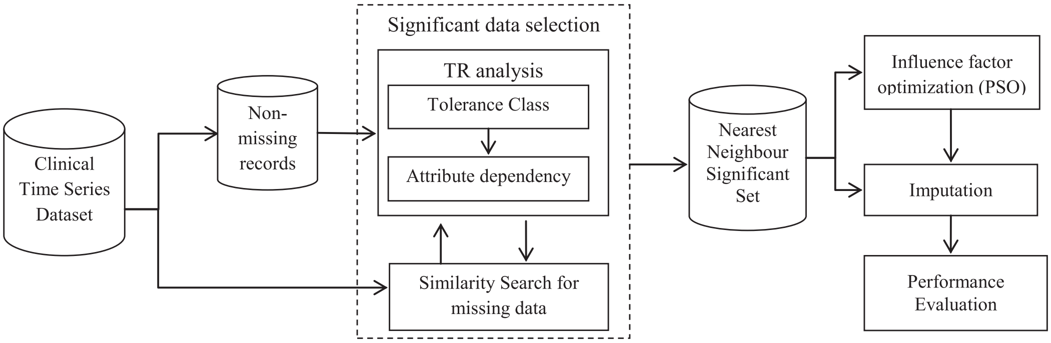

Clinical Time Series Dataset
Non-missing records
Significant data selection
TR analysis
Tolerance Class
Attribute dependency
Similarity Search for missing data
Nearest Neighbour Significant Set
Influence factor optimization (PSO)
Imputation
Performance Evaluation

Fig. 1. Proposed framework.

reports (Thrombosis, 1999). Each patient has undergone 564 laboratory tests. On combining the expert's knowledge and dataset descriptions given with the dataset (Jensen, 2001; Tsumoto, 1999; Thrombosis, 1999), 33 suggested tests have been identified for our experimentation. Out of 33 attributes 8 attributes are identified to have missing values. Table 1 shows the general dataset summary for the hepatitis and thrombosis datasets.

All the data included in the hepatitis and thrombosis dataset are associated with an unevenly spaced (irregular) time stamp. From these datasets, patients who have more than thirty percentages of incomplete data were not included for experimentation. Thus for hepatitis datasets 694 records and for thrombosis datasets 770 records were only considered in further experimentation.

A detailed description of the suggested lab tests for hepatitis and thrombosis patients is provided in Appendix A. These lab tests along with their date of examination are considered as temporal input attribute set for demonstrating the proposed temporal mining framework.

3.2. Methods

The proposed framework is shown in Fig. 1. The components of the proposed framework are significant data selection, influence factor optimization and imputation.

Imputing missing value in the clinical data becomes challenging due to the uneven spacings in the observed data period.

3.2.1. Significant data selection (SDS)

The classical IDW (Shepard, 1968) method finds the value for any unknown point using the measured known data significant to it, which forms the neighbourhood set. The size of the neighbourhood set chosen has a direct influence in the accuracy of the determined value. The IDW has limitation in choosing the number of known values for interpolation. The proposed work overcomes this limitation using significant data selection process, which aims at identifying the nearest significant known data points. This process is done in two steps: First, attribute dependent set generation and second, similarity search. The concept of tolerance rough set (TR) analysis is used in significant data selection process.

The following notations are used in the discussion and equations,

$I = (\mathbb{U}, A)$ represents an Information system
$\mathbb{U} = \{x_1 \ldots x_i \ldots x_j, \ldots, x_n\}$ refers to a universe, which is a nonempty set of finite objects.
x_i and x_j are the i^{th} and j^{th} objects in the universe '$\mathbb{U}$'.
A refers to knowledge in the universe which are the non-empty finite set of attributes, 'a' refers to an attribute where $a \in A$.
$a\,(x_i)$, $a\left(x_j\right)$ are the values of x_i and x_j for the attribute a.
τ is the tolerance value (0–1).
$a\left(t_j\right)$ and $a\,(t_i)$ are the observation periods of x_i, x_j for the attribute a.
$a_{\max}$ and $a_{\min}$ are the maximum and minimum values of a.
$a_{t_{\max}}$, $a_{t_{\min}}$ are the maximum and minimum durations.
$a\langle\left(x_i(t_i), x_j(t_j)\right)\ \tau\rangle$ value of x_i, x_j for the attribute at time t_i and t_j.
B and Q are two attributes such that $B \epsilon A$, $Q \epsilon B$; $B \notin Q$ $\bar{a}$ represents the mean.

Generally, an information system in rough sets is represented as $I = (\mathbb{U}, A)$, where $\mathbb{U} = \{x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n\}$ is called as a universe, which is a nonempty set of finite objects and A is the knowledge in the universe which is the non-empty finite set of attributes (Pawlak, 1982). The equivalence classes of the B-indiscernibility relation (IND(B)) is also denoted as $[x]_B$ and $X \subseteq \mathbb{U}$ represents an elementary portion of knowledge that can be extracted. This concept of equivalence class becomes complex when the attributes hold real value data. So to address the real value data problem, the concept of tolerance rough set was chosen (Komorowski et al., 1999).

The tolerance rough set forms tolerance classes based on the similarity in attributes. Lower and upper approximations are the two main operations that are used in characterizing the knowledge. Lower approximation of set contains all elements that surely belong to the set. Upper approximation of set contains all elements that possibly belong to set. Based on these approximations three regions are characterized namely positive, negative and boundary region. Positive region contains all the objects of '$\mathbb{U}$' that can be classified to equivalence classes of '$\mathbb{U}/B$' where $B \subseteq A$. Negative region contains all the objects, which are certainly non-member of X. Boundary region contains all objects which are possibly a member of X. The degree of dependency of the attributes is an important factor to be considered in the attribute selection process (Pawlak, 1982; Komorowski et al., 1999). The degree of dependency between two attribute sets is defined using the positive region.

In the first step of significant data selection process, dependent attributes are identified for each clinical attribute. A temporal similarity measure is used to compute similarity for each attribute based on its observation time. The temporal tolerance similarity measure, $\mathrm{SIM}_{a\langle(x_i(t_i),xj(t_j)),\tau\rangle}$ and temporal tolerance class relation, $\mathrm{tempTOL}_\tau(B)$ are defined in Eqs. (1) and (2). This temporal based similarity measures are then used to define lower approximations $(\underline{B_\tau}X)$ to construct positive regions $(\mathrm{POS}_{B,\tau}(Q))$ as defined in Eqs. (3) and (4). Finally, the significance of the attribute is computed using temporal tolerance based degree of dependency defined in Eq. (5).

$$\mathrm{SIM}_{a\langle(x_i(t_i),x_j(t_j)),\tau\rangle} = 1 - \bar{a}\left\{\left(\frac{a(x_i) - a(x_j)}{a_{\max} - a_{\min}}\right), \left(\frac{a(t_j) - a(t_i)}{a_{t_{\max}} - a_{t_{\min}}}\right)\right\}. \tag{1}$$

$$\mathrm{tempTOL}_\tau(B) = \{(x_i, x_j) \epsilon U \mid \forall a \epsilon B, (x_i, x_j) \in \mathrm{SIM}_{a\langle(x_i(t_i),x_j(t_j)),\tau\rangle}\}. \tag{2}$$

$$\underline{B_\tau}X = \{x \mid \mathrm{SIM}_{B,\tau}(x) \subseteq X\}. \tag{3}$$

$$\mathrm{POS}_{B,\tau}(Q) = \cup_{x\in U/Q}\, \underline{B_\tau}X. \tag{4}$$

$$K = \gamma_{B,\tau}(Q) = |\mathrm{POS}_{B,\tau}(Q)|/|\mathbb{U}|. \tag{5}$$

An attribute a in A where $a \in A$ is said to be dependent on other attributes a_i in A, $a_i \in A$ when its tolerance based degree of dependency value meets the dependency threshold $(D\tau)$. The value of $D\tau$ is selected by the domain based expert guidance. The complete data is grouped into two dataset categories namely missing records and non-missing records. TR analysis generates tolerance classes with the non-missing records. Lower approximation, positive region and degree of dependency are computed for each clinical attribute. Based on the degree of dependency the attribute dependent set is generated for each clinical attribute. The identified set represents the dependencies among the clinical attributes. In the second step, a similarity search process starts for each missing record. The similarity search identifies the number of significant known points that should be considered for interpolation. For each missing record, its attribute dependent set is extracted. The similar records are grouped based on the tolerance classes for the non-missing records of the identified dependent attributes.

The Significant_data_selection algorithm (Algorithm 1) describes similarity search process of SDS which selects the significant data points that form neighbourhood set for interpolation process. Fig. 2 illustrates the similarity search process in SDS.

The observation for the attribute $A1$ with OID (Object ID) 3, taken on 23/4/1981 is found missing. The attributes dependent for this missing attribute $A1$ are obtained from the attribute dependency set. Here, in this example $\{A2, A3\}$ are dependent attributes for $A1$. The tolerance class from the non-missing records for the attributes $\{A2, A3\}$ is generated. The maximum set which includes the missing OID 3 is identified as significant set for imputing its attribute $A1$ using interpolation.

3.2.2. *Influence factor optimization and imputation*

The traditional IDW (Shepard, 1968) works with an assumption that the distance between the measured values and its location has direct influence on the predicted value. To estimate the value at unknown point x, the general mathematical formulation for IDW (Shepard, 1968) is given in Eq. (6),

$$\mathrm{M_E}(x) = \frac{\sum_{i=1}^{n} \mathrm{M}_A(i)/d(x,i)^k}{\sum_{i=1}^{n} 1/d(x,i)^k} \tag{6}$$

where $\mathrm{M_E}(x)$ is the estimated value at point x, $\mathrm{M}_A(i)$ is the actual value at point i, n is the total number of known neighbour points for point i, $d(x, i)$ is the distance between the point x and the point i, and k is the influence factor parameter. This parameter value controls the weightage for each considered known point in interpolation. However, choosing this value

Algorithm 1: Significant_data_selection (X, Y, Z, τ, A, M, Dτ)

Input: *X- Complete dataset, Y- non-missing records, Z- missing records, τ-tolerance factor, A – complete attribute set, M- missing attribute set, Dτ -dependency threshold*

Output: *Attribute dependent set AD, nearest significant set (NS).*

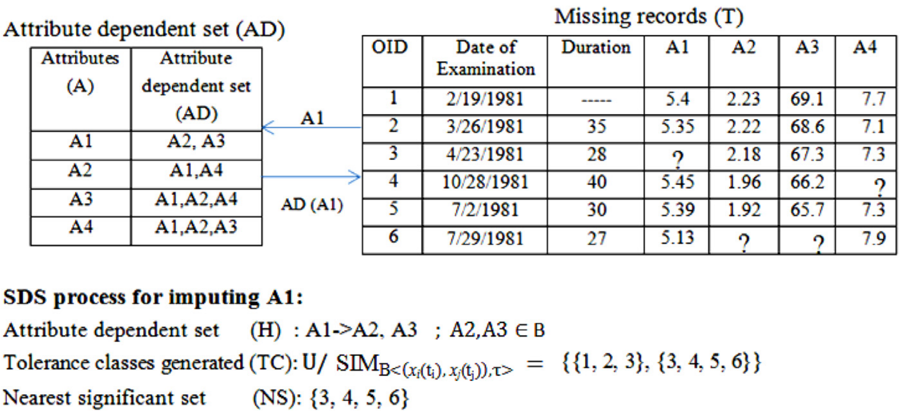
```
Attribute_dependency (Y, τ, A, Dτ)
Similarity_ search (X, Z, τ, AD ,A, M)
```

Algorithm 1.a: Attribute_dependency (Y, τ, **A,** Dτ)

```
m=size (A)
For i=1 to m do
D= A_i
R={ }
 For j=i+1 to m do
     $K = \gamma_{A_j,\tau}(A_i)$ // compute temporal degree of dependency using equation (5)
     If k> Dτ then
       $R = R \cup A_i$ // generate dependent attribute set of $A_i$
     end if
  End for
   AD_i = R
  End for
 return AD
```

Algorithm 1.b: Similarity_ search (X, Z, τ, AD, A, M)

```
n= size(M)
For i= 1 to n
  H = dependent attribute set (AD ) for $A_i$
  T= all non-missing records of H in  X
  TC=computed tolerance classes (tempTOL_τ(H)) for T using equation (2)
  $NS_i$ = Maximum set  in TC
End for
Return  NS
```



Attribute dependent set (AD)

Attributes (A)	Attribute dependent set (AD)
A1	A2, A3
A2	A1,A4
A3	A1,A2,A4
A4	A1,A2,A3

Missing records (T)

OID	Date of Examination	Duration	A1	A2	A3	A4
1	2/19/1981	-----	5.4	2.23	69.1	7.7
2	3/26/1981	35	5.35	2.22	68.6	7.1
3	4/23/1981	28	?	2.18	67.3	7.3
4	10/28/1981	40	5.45	1.96	66.2	?
5	7/2/1981	30	5.39	1.92	65.7	7.3
6	7/29/1981	27	5.13	?	?	7.9

SDS process for imputing A1:

Attribute dependent set (H) : A1->A2, A3 ; A2,A3 ∈ B

Tolerance classes generated (TC): $U/\ SIM_{B<(x_i(t_i),\,x_j(t_j)),\tau>} = \{\{1, 2, 3\}, \{3, 4, 5, 6\}\}$

Nearest significant set (NS): {3, 4, 5, 6}

Fig. 2. Significant data selection process.

needs to be done carefully since the wrong choice of this value affects the accuracy of the interpolation results. The proposed Influence factor optimization (IFO) process uses PSO technique to overcome the limitation of choosing optimized value for the influence factor (k). The chosen value for 'k' fixes the weights for each considered known data points in the significant

Algorithm 2: Influence_Factor_Optimization (n, l)

```
Input  : n-number of particles, l-iterations
Output: 'k'- Influence factor value
For each particle i =1 to n do
    Initialize particle position ($P_i$) using the equation (7 ) // represents the values for influence factor (k).
End for
$\text{pbest}_i = P_i$        // Initialize the pbest for each particle as its initial value.
do
    For each particle i =1 to n do
        Compute the fitness value $f_{RMSE_i}$ using the equation (10)
        If the fitness value $f_{RMSE_i} > \forall\ \text{pbest}_i$ then
            $\text{pbest}_i = P_i$ // set current particles position as local best
        end if
        If the fitness value $f_{RMSE_i} > \text{gbest}$
            gbest= $P_i$ ; gbestval= $f_{RMSE_i}$ // select gbest( the particle with the least fitness value )
        end if
        update the particle velocity using the equation (8)
        update the particle position using the equation (9)
    End for
while (maximum iteration or minimum error criteria)
```

set formed from the SDS process. The optimal value minimizes the error rate and improves the accuracy of the imputed results.

The concept of PSO was introduced by Kennedy and Eberhart (1995), which is a robust evolutionary computation technique based on the swarm intelligence. In PSO, particles represent solutions in a search space. A fitness function is used to evaluate the particles, which is represented by the objective function that has to be optimized. Each particle is guided by a measure called velocity to follow the best particles to search optimum point in problem space. The initial particle's position, velocity and its updated position are defined using Eqs. (7)–(9).

$$P_i = \text{range}_{\min} + (\text{range}_{\max} - \text{range}_{\min}) \cdot *\text{rand}\,(\text{numInd}, n_{\text{var}})\,. \tag{7}$$

$$V_{i+1} = V_i + c1 * r1 * (\text{pbest}_i - P_i) + c2 * r2 * (\text{gbest}_i - P_i)\,. \tag{8}$$

$$\text{P}_{i+1} = P_i + V_{i+1} \tag{9}$$

where rand() refers to the function that generates random numbers, $\text{range}_{\min}$ and $\text{range}_{\max}$ are the minimum and maximum ranges of the search space, pbest_i is the local best position of particle i, P_i is the current position of particle i and gbest_i is the global best position of particle i, V_i is the velocity of the i^{th} particle and V_{i+1} is the velocity of the $i+1^{\text{th}}$ particle, P_{i+1} is the current position of particle $i+1$, $c1$ is the cognitive coefficient and $c2$ is the social component. $r1$ and $r2$ are the stochastic influence component whose values are randomly taken in the range 0–1.

The fitness function of the proposed TRiBS is shown in Eq. (10).

$$\text{Minimize}\, f_{RMSE_i} = \sqrt{\frac{(\text{M}_A(i) - \text{M}_\text{E}\,(x))^2}{n}}. \tag{10}$$

The algorithm Influence_Factor_Optimization (Algorithm 2) given below describes the overall steps involved in IFO process.

4. Experimental results and discussions

The proposed TRiBS framework has been experimented with two real clinical time series and a simulated data. A detailed explanation about the experimental settings and the results obtained during the imputation process using the proposed framework is discussed in this section.

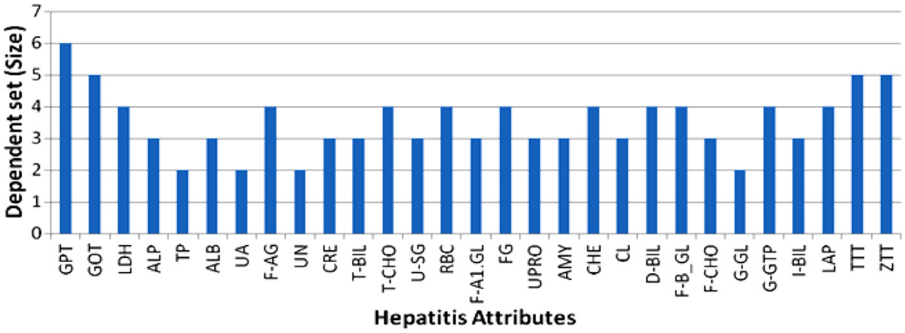


Fig. 3. Dependency set for hepatitis attributes.

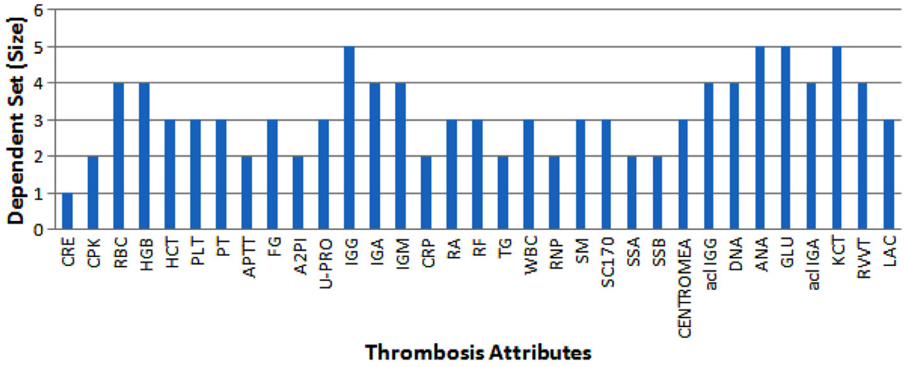


Fig. 4. Dependency set for thrombosis attributes.

4.1. An application on real data: hepatitis and thrombosis

TRiBS framework performs two major functionalities namely significant data selection and influence factor optimization. Two real clinical time series datasets of hepatitis and thrombosis patients are used for experimentation. The concept of tolerance rough set (TR) is used in significant data selection process to select the relevant known data points which form the nearest significant set. The initial step of TR starts with finding the dependent attribute set for every attribute using the non-missing records of hepatitis and thrombosis datasets. Since the data considered for experimentation is a time series data, the TR procedure computes the tolerance class, lower approximation, positive region and degree of dependency with respect to the temporal duration among the clinical observations. The computational procedures are illustrated in Section 3.2.

Figs. 3 and 4 summarize the number of dependent attributes identified for each clinical attribute in hepatitis and thrombosis datasets. The X-axis in Figs. 3 and 4 denotes the clinical attribute that will be considered for hepatitis and thrombosis diagnosis. The Y-axis in Figs. 3 and 4 denotes its corresponding number of dependent attributes. To identify and group the dependent attributes a tolerance threshold level of 0.48 is considered with expert guidance. An attribute is said to be dependent on the other only when its tolerance degree of dependency is less than or equal to 0.48. This dependency attribute set acts like a dictionary in the imputation process. To impute a missing data first its dependent attributes are identified and a similarity search is performed for the non-missing records in the data for those dependent attributes. This is done by generating interval-based tolerance class. The maximum set in the tolerance class, which includes the missing record, is considered as nearest significant set.

The characteristics of the missing attributes namely its temporal interval and its degree of dependency with other attributes have higher impact in choosing the nearest significant set boundaries. The missing attribute that is highly dependent on the other attributes and has a lower temporal interval compared to that of average interval spacing among the clinical observations has high nearest significant set. The PSO technique now fixes weights for each element in nearest significant set by finding the optimal value for the influence factor parameter. The cognitive and social coefficient is taken to be 1. The stochastic constants $r1$ and $r2$ are any random numbers in the range 0–1. The search space to form the solution set is fixed to be in the range of 0–15. Table 2 summarizes the best fitness position on an average obtained for each particle at the iterations 100, 150, 200, 300, 500 and 600 of a when imputing a missing data in the attribute LAP for hepatitis patients.

The selection of the number of particles and iterations is based on minimal RMSE value. For hepatitis and thrombosis data the number of particles was chosen to be 20 and the iterations to be 300. Fig. 5 shows the plot to illustrate the best fitness position taken on average at different iterations when imputing a missing data in the attribute LAP for hepatitis patients. The X-axis represents the 300 iterations; Y-axis represents the best fitness value of nearest significant neighbour set formed for the corresponding missing attribute. It was observed that on an average best fitness value is found to be at the position 1.5;

Table 2
Fitness position (missing attribute-LAP).

Number of particles	Fitness position					
	100 iterations	150 iterations	200 iterations	300 iterations	500 iterations	600 iterations
5	4.22	4.25	4.11	3.98	4.01	3.87
10	3.145	3.241	3.334	3.51	3.11	3.07
15	3.411	3.417	3.356	2.882	2.545	2.541
20	2.449	2.401	2.417	1.503	1.503	1.502

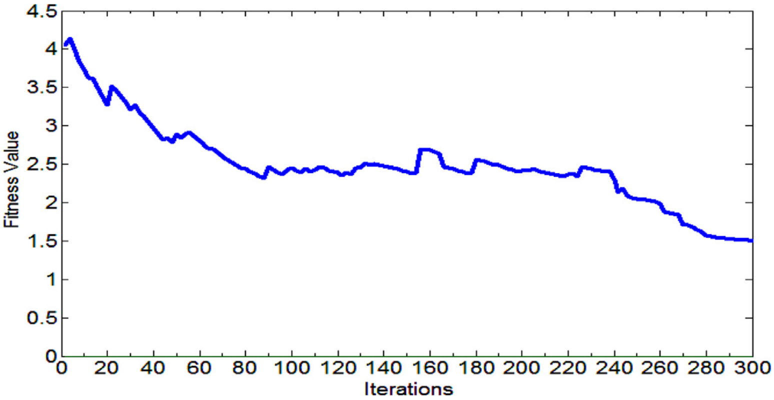


Fig. 5. Plot on fitness value vs. iterations (missing attribute-LAP).

this value is taken as optimal value for the influence factor. This process is repeated for each generated nearest significant set in significant data selection process.

A 10-fold cross validation with two independent runs evaluation scheme is adopted for generating the training and test subsets. For evaluating the experimental results the performance metrics considered are mean absolute deviation (MAD), root mean squared error (RMSE), mean absolute percentage error (MAPE), fractional bias error (FB) and index of agreement (IA). Let D_i be the i^{th} observed value, F_i be the i^{th} estimated value and $\bar{D}$ be the mean of observed value. Eqs. (11)–(15) define the MAD, RMSE, MAPE, FB and IA measures.

Mean Absolute Deviation (MAD):

Mean absolute deviation or MAD, is a measure that represents the positive and negative deviations between the forecast and the actual demand. Mathematically, it is represented as

$$\mathrm{MAD}_n = \frac{1}{n}\sum_{i=1}^{n} |D_i - F_i|. \tag{11}$$

Root Mean Squared Error (RMSE):

RMSE measures the average squares of the errors. Mathematically, it is represented as

$$\mathrm{RMSE}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i - F_i)^2}\,. \tag{12}$$

Mean Absolute Percentage Error (MAPE):

MAPE is a measure that represents the magnitude of the error relative to the magnitude of the demand. The average ratio is multiplied by 100 to represent this relative measure as a percent. Mathematically, it is represented as

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \frac{|\,(D_i - F_i)\,|}{D_i} * 100. \tag{13}$$

Index of Agreement (IA):

It is a measure of correlation between the observed and estimated values. It varies between 0 and 1. A value of 1 indicates strong correlation and a value of 0 indicates no agreement.

$$\mathrm{IA} = 1 - \left(\left(\sum_{i=1}^{n}(D_i - F_i)^2\right) \Big/ \sum_{i=1}^{n}\left(\left|\left(D_i - \bar{D}\right)\right| + \left|\left(F_i - \bar{D}\right)\right|\right)^2\right). \tag{14}$$

Table 3
Comparison of imputation results TRiBS, KNN, EM, IDW for hepatitis patients.

Performance measures	Imputation techniques	Hepatitis attribute								
		F-A/G	LAP	F-A1_GL	F-B_GL	AMY	CHE	CL	F-CHO	G-GTP
MAD	TRiBS	0.22	14.2	0.32	0.73	4.3	0.41	6.9	2.8	2
	KNN	0.72	36.8	0.88	1.65	6.9	1.17	22.2	7	5.1
	EM	1.04	43.2	0.75	2.33	22.9	1.61	23.7	7.7	8.2
	IDW	1.25	61.9	1.29	2.81	27.1	2.37	31.5	10.2	9.3
RMSE	TRiBS	0.26	16.04	0.33	0.82	4.83	0.5	7.4	2.93	2.1
	KNN	0.77	44.59	0.96	1.77	17.93	1.24	25.1	7.55	5.92
	EM	1.05	50.63	0.9	2.37	24.77	1.76	26.17	8.61	9.1
	IDW	1.26	69.43	1.84	2.87	28.74	2.46	31.68	11.33	10.44
MAPE	TRiBS	10.37	3.94	9.56	9.62	5.41	4.39	6.61	5.03	7.76
	KNN	33.29	10.38	26.25	21.55	21.07	12.62	21.26	12.52	18.58
	EM	48.36	12.12	22.01	30.39	28.15	17.16	22.71	13.69	30.68
	IDW	58.74	17.41	40.77	36.7	33.37	25.19	30.16	18	34.33
IA	TRiBS	0.63	0.93	0.86	0.73	0.84	0.77	0.35	0.63	0.98
	KNN	0.24	0.7	0.39	0.44	0.37	0.46	0.08	0.08	0.88
	EM	0.2	0.63	0.37	0.34	0.26	0.38	0.09	0.17	0.76
	IDW	0.18	0.53	0.24	0.29	0.23	0.21	0.09	0.15	0.7
FB	TRiBS	0.11	0.02	0.09	0.1	0.02	0.01	0.07	0.03	0.02
	KNN	0.41	0.11	0.31	0.25	0.18	0.03	0.13	0.04	0.09
	EM	0.64	0.13	0.26	0.36	0.27	-0.04	0.15	0.06	0.14
	IDW	0.84	0.2	-0.18	0.45	0.31	0.05	0.35	0.09	0.17

Fractional Bias (FB):

This measure is used to identify the underestimate and overestimate of the predicted value. This value lies in the range -2 to $+2$ and desired value is 0.

$$\mathrm{FB} = \frac{1}{n}\sum_{i=1}^{n}\left(D_i - F_i\right) / \left(\left(D_i + F_i\right)/2\right). \tag{15}$$

The proposed TRiBS method is compared with the other imputation techniques such as KNN, EM and IDW. Tables 3 and 4 summarize the imputation results over the considered performance measures for hepatitis and thrombosis datasets. In the hepatitis dataset, the results of nine attributes (F-A/G, LAP, F-A1_GL, F-B_GL, AMY, CHE, CL, F-CHO and G-GTP) and in the thrombosis dataset the results of seven attributes (TG, CPK, WBC, RBC, PLT CRE and HGB) are shown in Tables 3 and 4. The comparative results of other suggested attributes in hepatitis and thrombosis dataset are provided in Appendix B.

The error rates, MAD, RMSE and MAPE values, for the proposed TRiBS have been reduced compared to the KNN, EM and IDW. The IA value of TRiBS is closer to 1 which proves the effectiveness of the imputed value. Similarly the FB value of TRiBS for the imputed attributes in hepatitis and thrombosis datasets is closer to 0 which shows the reduction in the bias error of the imputed results.

From the obtained imputation results it can be observed that on an average the IA for the proposed improved IDW method is closer to 1 thereby ensuring that there is strong correlation between the actual observed and imputed values. The FB values prove that the proposed methodology highly limits underestimation or overestimation. A statistical paired t-test (Zimmerman and Donald, 1997) was carried out with significance value of 0.05 to assess the error rates of TRiBS over KNN, EM and IDW. The p-value obtained (p-value < 0.05) shows that there is a significant difference in the error rates for TRiBS over KNN, EM and IDW.

The impact of the proposed missing data imputation is also evaluated by extracting temporal features from the imputed data and applying classification techniques SVM, Neural networks, Decision Trees using R-tool (Team, 2013; Zhao, 2011). Fig. 6 shows the comparison of classification accuracies of the classifiers NN, DT and SVM in combination with the proposed TRiBS, KNN, EM and IDW imputation techniques.

The obtained classification results show that the combination of NN with TRiBS attains the classification accuracies of 83.57% and 80.15% for the missing rates of 10% and 15% respectively. This was found to be higher when compared to NN with KNN, EM and IDW imputations. The combination of DT with TRiBS attains the classification accuracy 81.14% and 77.91% for the missing rate of 10% and 15% respectively which was found to be higher when compared to DT with KNN, EM and IDW imputations. The combination of SVM with TRiBS attains the classification accuracy 78.89% and 76.19% for the missing rate of 10% and 15% respectively which was found to be higher when compared to DT with KNN, EM and IDW imputations. It can be inferred from Fig. 6 that there is a significant improvement in the classification accuracy for the pre-processed data using TRiBS imputation technique even when there is an increase in the percentage of missingness.

Table 4
Comparison of imputation results TRiBS, KNN, EM, IDW for thrombosis patients.

Performance measures	Imputation techniques	Thrombosis attribute						
		TG	CPK	WBC	RBC	PLT	CRE	HGB
MAD	TRiBS	8.7	1.9	0.58	0.39	13.7	0.08	0.59
	KNN	19.7	6.2	1.038	0.79	51.3	0.13	1.78
	EM	16.8	6	1.34	0.75	56.1	0.12	2.12
	IDW	27.4	8	1.76	1.267	62.4	0.22	2.84
RMSE	TRiBS	10.38	1.97	0.65	0.49	16.54	0.1	0.68
	KNN	23.83	7.5	1.15	0.94	56.81	0.15	1.88
	EM	21.61	6.84	1.41	0.91	62.04	0.13	2.22
	IDW	29.25	8.99	1.85	1.52	64.55	0.26	2.9
MAPE	TRiBS	8.86	8.2	8.41	8.38	4.78	8.97	5.06
	KNN	20.04	24.34	14.74	17.16	16.93	14.53	15.67
	EM	16.4	24.69	19.21	16.2	18.21	13.02	18.61
	IDW	29.44	31.85	25.11	27.17	21.11	26.39	24.85
IA	TRiBS	0.95	0.98	0.75	0.39	0.97	0.92	0.97
	KNN	0.73	0.74	0.44	0.15	0.67	0.81	0.82
	EM	0.74	0.62	0.38	0.17	0.63	0.87	0.76
	IDW	0.66	0.67	0.3	0.11	0.62	0.69	0.68
FB	TRiBS	0.15	0.08	0.1	0.11	0.05	0.05	0.08
	KNN	0.34	0.1	0.29	0.17	0.26	0.07	0.24
	EM	0.27	0.16	0.38	0.19	0.32	0.07	0.28
	IDW	0.47	0.25	0.5	0.31	0.32	0.14	0.38

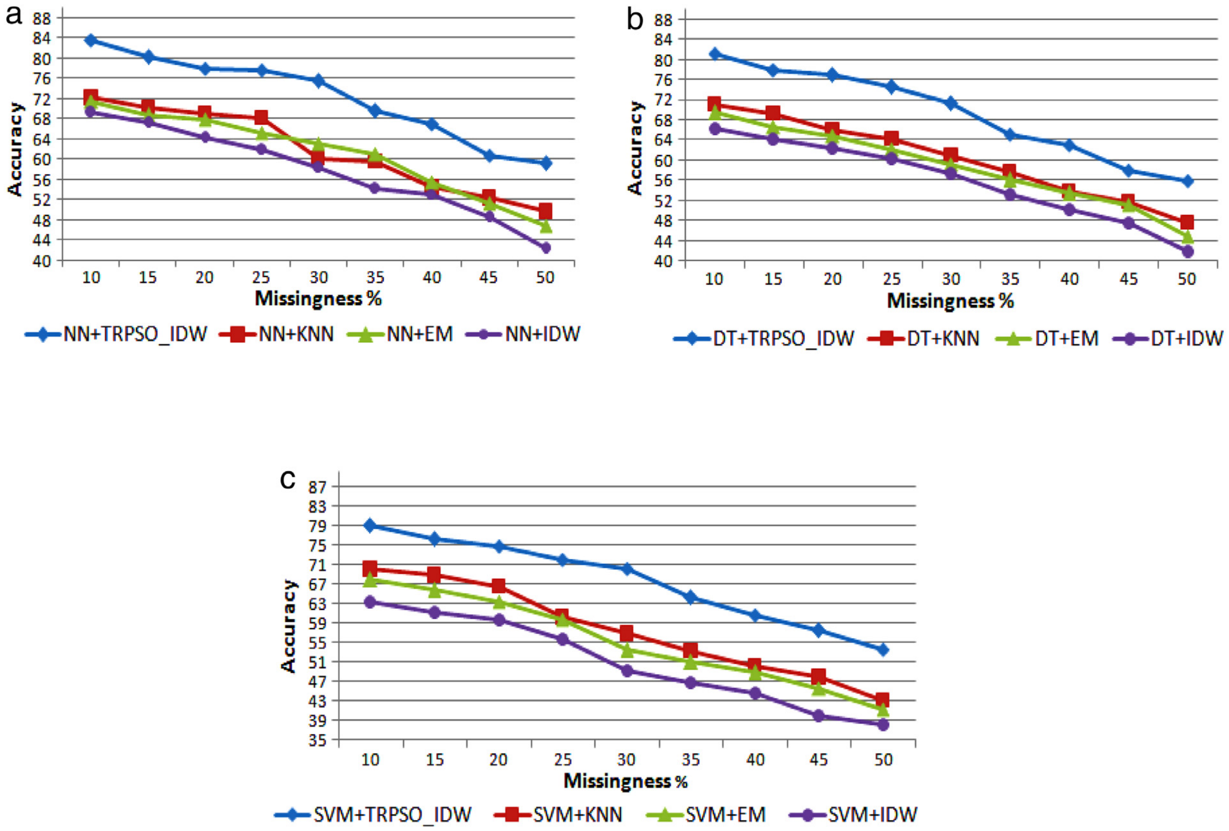


Fig. 6. Comparison of classification accuracies for the real data: (a) NN with TRiBS, KNN, EM and IDW. (b) DT with TRiBS, KNN, EM and IDW. (c) SVM with TRiBS, KNN, EM and IDW.

4.2. A simulation study

This section presents the design and implementation of a simulation study that is used to validate the performance of the proposed TRiBS framework. Data simulation is the process of generating data that imitates the actual scenarios using

Table 5
Simulation parameters.

Setting	Correlation coefficient (ρ)	Missing rate
1	0.3	10 20
2	0.5	10 20
3	0.7	10 20

a mathematical model (Liu, 2015; Boker et al., 2011). This work performs a simulation to generate a data that imitates a chronic disease model. Generally, in the clinical domain, for diagnosing a disease, a patient undergoes several laboratory examinations. The measurement of these laboratory examinations is done repeatedly for a patient. Due to these repeated measurements, clinical data are often considered to be a longitudinal data. A data is said to be longitudinal, if it records, multiple observations on the same subject over a period of time (Liu, 2015). Longitudinal data represent unevenly spaced time series data, if its observations are done at irregular intervals. Clinical data are often considered to be longitudinal and unevenly spaced, since the observation of health conditions for each patient are done at irregular intervals of time. Moreover, the number of observations done also differs for each patient.

The process of data simulation consists of the following three steps: first, identification of mathematical structure that fits the data; second, sample data generation; third, formatting the simulated data (Boker et al., 2011). In a chronic disease diagnosis, the patients undergo several laboratory examinations, which make the data to be multivariate. This work initiates the simulation process by identifying the mathematical structure of hepatitis and thrombosis data. To understand the mathematical structure of the hepatitis and thrombosis data, a data fitting model in MATLAB is used (MATLAB and Statistics Toolbox, 2013). A simulated data for a chronic disease are generated based on the underlying structure of hepatitis and thrombosis data. Clinical observations are found to exhibit linear relationship and follow a normal distribution. Thus, a multivariate normal distribution model with covariance structure is adopted for simulating an unevenly spaced clinical time series data.

Let $Y_j = \left\{Y_{jt_n}(i), t_n \in T, j \in m, i \in A; t_{n+1} > t_n\right\}$ be a clinical time series data where t_n is n^{th} time-point of observation, T is the set of observed time points ($T = \{t_1, t_2, \ldots, t_n\}$), m is the number of subjects and A is the attribute set. In clinical time series data, the observations are done consecutively and observations at two different time points say t_n and t_{n+1} are correlated. The observations that are done consecutively with shorter intervals are said to be highly correlated and as the interval of observations gets farther they are less correlated. These time variations can be described using a first order auto-regression AR(1) covariance structure. However, the AR(1) covariance structure is specific to evenly spaced time series.

The spatial power, SP(POW) covariance structure is an adaptation of first order auto regressive, AR(1) covariance structure for unevenly spaced time series (Liu, 2015). Since, clinical time series data are unevenly spaced this work adopts a spatial power, SP(POW) covariance structure. The difference between AR(1) and SP(POW) covariance structure lies in choosing the power factor of the correlation coefficient. In SP(POW), the correlation coefficient is raised to the power of difference between two observed times (d_{ij}), whereas for AR(1) it is raised to the power factor such as $1, 2, 3 \ldots n$. Let σ^2 represent observations taken at any given time, which has the same variance, $\sum$ represents the covariance structure, n is the point of observation and $\rho^{d_{12}}$ is the correlation between two observations (Liu, 2015). Eq. (16) defines the SP(POW) covariance structure.

$$\sum = \sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \cdots & \rho^{d_{1n}} \\ \rho^{d_{12}} & 1 & \rho^{d_{23}} & \cdots & \rho^{d_{2n}} \\ \rho^{d_{13}} & \rho^{d_{23}} & 1 & \cdots & \rho^{d_{3n}} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho^{d_{1n}} & \rho^{d_{2n}} & \rho^{d_{3n}} & \cdots & 1 \end{bmatrix}. \tag{16}$$

A multivariate normal distribution with linear response function is represented as $\upsilon = \propto_0 + \propto_1 t$, where $\propto_0$ is the intercept and $\propto_1$ is the slope. The samples are generated from the chosen multivariate normal distribution with SP(POW) covariance structure (MATLAB and Statistics Toolbox, 2013). The parameters used in designing this simulation study are shown in Table 5. In this work, the simulation process generates 10 attributes measured repeatedly at 100 different time points for 100 subjects. This process is repeated 100 times and hence the number of simulated samples generated is 10 000. For experimentation, this work considers the intercept ($\propto_0$) to be 5 and the slope ($\propto_1$) to be 0.1. The correlation between two observations is assumed to be positive. The correlation coefficient (ρ) takes values such as 0.3, 0.5 and 0.7. The missing rates of 10% and 20% are considered in the evaluation study. Two missingness types, namely missing completely at random (MCAR) and missing at random (MAR) are considered in this simulation study. MCAR test presented by Little (1988) was performed to ensure the incorporation of MCAR type missingness. For incorporating MAR type of missingness, the missing probabilities were chosen based on the observed data.

Table 6
Comparison of imputation results for TRiBS, KNN, EM and IDW using simulation study (MCAR missing type).

Imputation techniques	ρ	MAD		RMSE		MAPE		IA		FB	
		Missing 10%	Missing 20%	Missing 10%	Missing 20%	Missing 10%	Missing 20%	Missing 10%	Missing 20%	Missing 10%	Missing 20%
TRiBS	0.3	0.340	0.400	0.454	0.539	6.279	7.426	0.769	0.724	0.064	0.087
	0.5	0.275	0.335	0.355	0.454	5.077	6.189	0.831	0.764	0.050	0.065
	0.7	0.157	0.206	0.197	0.252	2.88	3.795	0.932	0.9	0.004	0.016
KNN	0.3	0.715	0.795	0.843	0.902	13.234	14.651	0.569	0.552	0.135	0.128
	0.5	0.685	0.705	0.793	0.816	12.578	12.956	0.588	0.580	0.087	0.095
	0.7	0.525	0.615	0.644	0.768	9.784	11.448	0.676	0.608	0.137	0.170
EM	0.3	0.705	0.745	0.821	0.843	12.991	13.722	0.586	0.577	0.110	0.124
	0.5	0.595	0.605	0.731	0.697	10.913	11.104	0.613	0.639	0.068	0.058
	0.7	0.495	0.565	0.671	0.708	9.158	10.377	0.635	0.626	0.103	0.079
IDW	0.3	0.811	0.881	0.953	1.048	14.984	16.297	0.532	0.499	0.166	0.193
	0.5	0.702	0.781	0.881	0.958	12.962	14.374	0.540	0.510	0.160	0.154
	0.7	0.601	0.682	0.785	0.875	11.123	12.577	0.571	0.536	0.138	0.153

Table 7
Comparison of imputation results for TRiBS, KNN, EM and IDW using simulation study (MAR missing type).

Imputation techniques	ρ	MAD		RMSE		MAPE		IA		FB	
		Missing 10%	Missing 20%	Missing 10%	Missing 20%	Missing 10%	Missing 20%	Missing 10%	Missing 20%	Missing 10%	Missing 20%
TRiBS	0.3	0.295	0.305	0.355	0.385	5.352	5.684	0.823	0.806	0.005	0.044
	0.5	0.196	0.215	0.242	0.265	3.603	3.930	0.905	0.891	0.012	0.006
	0.7	0.160	0.197	0.173	0.254	2.915	3.634	0.936	0.890	0.013	0.003
KNN	0.3	0.611	0.631	0.716	0.741	11.207	11.578	0.590	0.577	0.106	0.114
	0.5	0.537	0.597	0.644	0.723	9.863	10.954	0.625	0.574	0.101	0.113
	0.7	0.411	0.474	0.559	0.623	7.449	8.543	0.583	0.527	0.084	0.095
EM	0.3	0.684	0.731	0.791	0.835	12.572	13.477	0.567	0.563	0.115	0.121
	0.5	0.501	0.557	0.632	0.668	9.118	10.15	0.602	0.556	0.082	0.110
	0.7	0.447	0.507	0.551	0.603	8.255	9.308	0.690	0.650	0.092	0.113
IDW	0.3	0.754	0.812	0.869	0.923	13.843	14.868	0.535	0.505	0.141	0.129
	0.5	0.657	0.697	0.756	0.795	12.086	12.813	0.598	0.577	0.110	0.118
	0.7	0.581	0.604	0.714	0.728	10.644	11.132	0.576	0.611	0.101	0.104

The results of simulation are summarized in Tables 6 and 7. The performance measures such as mean absolute deviation (MAD), root mean squared error (RMSE), mean absolute percentage error (MAPE), fractional bias error (FB) and index of agreement (IA) are considered for evaluating the performance of the proposed TRiBS over other imputation methods such as KNN, EM and IDW. The experimental results show that the MAD, RMSE and MAPE errors decrease when there is a positive correlation $\rho = 0.7$ and a decrease in the slope. It can be observed that when there is an increase in the slope and missing rate the error measures increase in all the methods.

Table 6, shows the comparative results of MCAR type missingness obtained for TRiBS, KNN, EM and IDW imputation using various simulation settings mentioned in Table 5. It can be inferred from the results that for MCAR type missingness, the proposed TRiBS imputation shows reduced error rates compared to KNN, EM and IDW. Similarly, Table 7 shows the comparative results of MAR type missingness obtained for TRiBS, KNN, EM and IDW imputations using various simulation settings mentioned in Table 5. The obtained results prove that there is a reduction in the error rates for the TRiBS imputation compared to KNN, EM and IDW.

The significance of error rates of TRiBS over KNN, EM and IDW is assessed using a statistical paired t-test (Zimmerman and Donald, 1997) analysis with a significance level of 0.05. The observed p-value is less than the expected value for the level of significance; this indicates that there is a significant decrease in the error rates for TRiBS over KNN, EM and IDW. Asymptotic properties are used for analysing the performance of any statistical estimators (Liu, 2015). They are generally used to determine the behaviour of a given estimator as the number of observations (n) grows to infinity ($n \to \infty$). The following are the basic asymptotic properties: consistency, asymptotic normality and asymptotic efficiency (Liu, 2015).

TRiBS exhibits the asymptotic property of mean squared error consistency. A sequence of estimators (δ_n) of the parameter (δ) is said to be mean square error consistent, if $\lim_{n\to\infty} \mathrm{E}\left[(\delta_n - \delta)^2\right] = 0$, where '$n$' is the sample size (Liu, 2015). To analyse the mean squared error consistency of the TRiBS framework, samples of various sizes are generated with the simulation parameters considered in the simulation study. The sample sizes ($n = 100, 150, 200, 300, 400, \ldots 1600$) are generated using the simulation setting such as positive correlation coefficients of 0.3, 0.5 and 0.7. The mean square value is computed for the MCAR and MAR missing types for 10% and 20% of missingness. Fig. 7(a) shows the average value of MSE obtained for

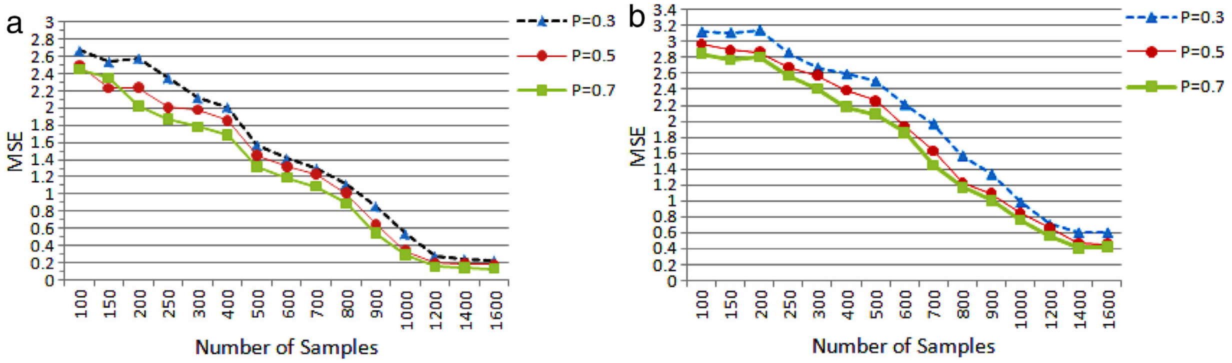


Fig. 7. (a) Number of samples vs. MSE for MAR missing type. (b) Number of samples vs. MSE for MCAR missing type.

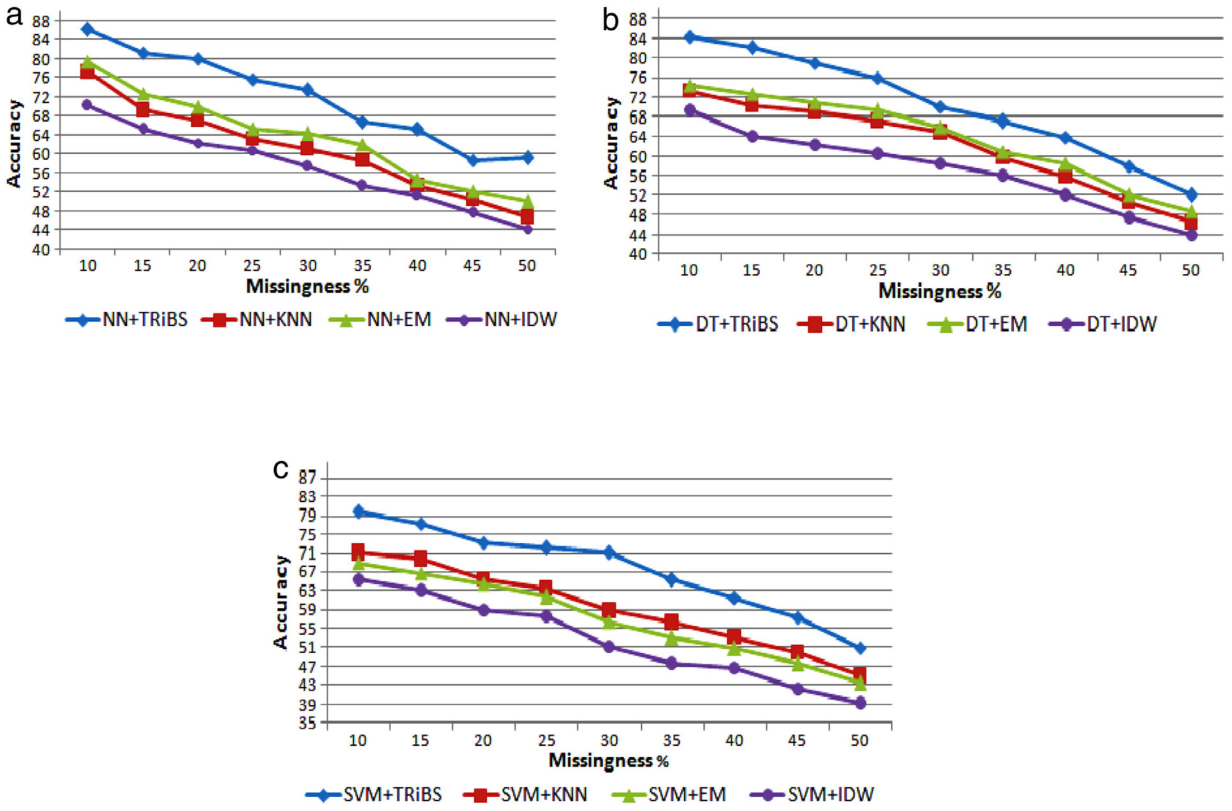


Fig. 8. Comparison of classification accuracies for the simulated data: (a) NN with TRiBS, KNN, EM and IDW. (b) DT with TRiBS, KNN, EM and IDW. (c) SVM with TRiBS, KNN, EM and IDW.

various numbers of sample size for 10% and 20% MAR type missingness. Fig. 7(b) shows the average value of MSE obtained for various numbers of sample size for 10% and 20% MCAR type missingness.

It can be observed that for both MAR and MCAR type missingness, as the sample size grows the bias value tends to get zero and the MSE gets reduced and equals to its variance or near zero. Thus, it can be inferred that TRiBS exhibit mean squared error consistency.

The impact of the proposed missing data imputation has been evaluated by applying classification techniques such as SVM, K-NN, Neural networks and Decision Trees using R-tool (Team, 2013; Zhao, 2011). Fig. 8 shows the comparison of classification accuracies taken on average for the classifiers NN, DT and SVM in combination with the TRiBS, KNN, EM and IDW imputation techniques.

From Fig. 8, it can be observed that the classification model built after the TRiBS imputation shows an increase in the classification accuracies compared to KNN, EM and IDW imputation techniques.

5. Conclusion

Clinical time series data are often observed at irregular intervals and hence they are referred to as unevenly spaced data. The presence of missing data makes it difficult to be used in the knowledge discovery process. The proposed tolerance rough set induced bio-statistical framework handles the missing value by improving the traditional IDW techniques using the concept of TR and PSO. A rough set concept is used in TR to select the neighbourhood set for each unknown data point. The PSO technique is then used to find the optimal value for the influence factor to fix the weightage of each known data point in the neighbourhood set. IDW interpolation process is carried out using the obtained neighbourhood set and influence factor for the corresponding unknown data locations. To demonstrate the proposed work two clinical time series data of hepatitis and thrombosis patients were considered. The experimental results show that the proposed system has reduced the error rate of imputed results compared to the other imputation techniques like KNN, EM and traditional IDW.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2017.02.012.

References

Andridge, R.R., Little, R.J., 2010. A review of hot deck imputation for survey non-response. Int. Statist. Rev. 78 (1), 40–64. http://dx.doi.org/10.1111/j.1751-5823.2010.00103.x.

Barnes, S.L., 1964. A technique for maximizing details in numerical weather map analysis. J. Appl. Meteorol. 3 (4), 396–409. http://dx.doi.org/10.1175/1520-0450(1964)003<0396:ATFMDI>2.0.CO;2.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., 2011. OpenMx: an open source extended structural equation modeling framework. Psychometrika 76 (2), 306–317. http://dx.doi.org/10.1007/s11336-010-9200-6.

Cismondi, F., Fialho, A.S., Vieira, S.M., Reti, S.R., Sousa, J.M., Finkelstein, S.N., 2013. Missing data in medical databases: Impute, delete or classify? Artif. Intell. Med. 58 (1), 63–72. http://dx.doi.org/10.1016/j.artmed.2013.01.003.

Cressman, G.P., 1959. An operational objective analysis system. Mon. Weather Rev. 87 (10), 367–374.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 1–38. http://www.jstor.org/stable/2984875.

Ding, Y., Ross, A., 2012. A comparison of imputation methods for handling missing scores in biometric fusion. Pattern Recognit. 45 (3), 919–933. http://dx.doi.org/10.1016/j.patcog.2011.08.002.

Enders, C.K., 2010. Applied Missing Data Analysis. Guilford Press.

Enders, C.K., Bandalos, D.L., 2001. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. Struct. Equ. Model. 8 (3), 430–457.

Farhangfar, A., Kurgan, L.A., Pedrycz, W., 2007. A novel framework for imputation of missing values in databases. IEEE Trans. Syst. Man. Cybern.-Part A: Syst. Hum. 37 (5), 692–709. http://dx.doi.org/10.1109/TSMCA.2007.902631.

Ford, B.L., 1983. An Overview of Hot-Deck Procedures: Incomplete Data in Sample Surveys. 2.

Gandin, L.S.L.S., 1970. The planning of meteorological station networks (No. 04; QC875. 5, G3.).

Gheyas, I.A., Smith, L.S., 2010. A neural network-based framework for the reconstruction of incomplete data sets. Neurocomputing 73 (16), 3039–3065. http://dx.doi.org/10.1016/j.neucom.2010.06.021.

2005. Hepatitis dataset for discovery challenge. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 16th ECML + 9th PKDD Conference, October 3–10, Porto, Portugal, http://lisp.vse.cz/challenge/ecmlpkdd2005/ (Accessed: 01.01.06).

Jensen, S., 2001. Mining medical data for predictive and sequential patterns: PKDD 2001. In: Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases.

Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M., Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 50 (2), 105–115. http://dx.doi.org/10.1016/j.artmed.2010.05.002.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmos. Environ. 38 (18), 2895–2907. http://dx.doi.org/10.1016/j.atmosenv.2004.02.026.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks IV, pp. 1942–1948.

Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, , 1999. Rough sets perspective on data and knowledge. In: Klosgen, W., Zytkow, J.M. (Eds.), The Handbook of Data Mining and Knowledge Discovery. Oxford University Press, New York.

Little, R.J., 1988. A test of missing completely at random for multivariate data with missing values. J. Amer. Statist. Assoc. 83 (404), 1198–1202. http://www.jstor.org/stable/2290157.

Little, R.J., Rubin, D.B., 2014. Statistical Analysis with Missing Data. John Wiley & Sons.

Liu, X., 2015. Methods and Applications of Longitudinal Data Analysis. Elsevier.

Lu, G.Y., Wong, D.W., 2008. An adaptive inverse-distance weighting spatial interpolation technique. Comput. Geosci. 34 (9), 1044–1055. http://dx.doi.org/10.1016/j.cageo.2007.07.010.

MATLAB and Statistics Toolbox Release 2013. The MathWorks, Inc., Natick, Massachusetts, United States.

Nordbotten, S., 1996. Neural network imputation applied to the Norwegian 1990 Population Census Data. J. Off. Stat. 12 (4), 385–401. http://www.jos.nu/Articles/abstract.asp?article=124385.

Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I., Ishii, S., 2003. A Bayesian missing value estimation method for gene expression profile data. Bioinformatics 19 (16), 2088–2096. http://dx.doi.org/10.1093/bioinformatics/btg287.

Ohsaki, M., Sato, Y., Yokoi, H., Yamaguchi, T., 2002. A rule discovery support system for sequential medical data, in the case study of a chronic hepatitis dataset. In: Workshop Notes of the International Workshop on Active Mining, at IEEE International Conference on Data Mining, December.

Olinsky, A., Chen, S., Harlow, L., 2003. The comparative efficacy of imputation methods for missing data in structural equation modeling. European J. Oper. Res. 151 (1), 53–79. http://dx.doi.org/10.1016/S0377-2217(02)00578-7.

Pawlak, Z., 1982. Rough sets. Int. J. Comput. Inf. Sci. 11, 341–356. http://dx.doi.org/10.1007/BF01001956.

Perez, A., Dennis, R.J., Gil, J.F., Rondón, M.A., López, A., 2002. Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia. Stat. Med. 21 (24), 3885–3896. http://dx.doi.org/10.1002/sim.1391.

Qu, L., Li, L., Zhang, Y., Hu, J., 2009. PPCA-based missing data imputation for traffic flow volume: a systematical approach. IEEE Trans. Intell. Transp. Syst. 10 (3), 512–522. http://dx.doi.org/10.1109/TITS.2009.2026312.

Rahman, M.G., Islam, M.Z., 2013. Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques. Knowl.-Based Syst. 53, 51–65. http://dx.doi.org/10.1016/j.knosys.2013.08.023.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. CRC Press.

Scheuren, F., 2005. Multiple imputation: How it began and continues. Amer. Statist. 59 (4), 315–319. http://dx.doi.org/10.1198/000313005X74016.

Sen, Z., Sahin, A.D., 2001. Spatial interpolation and estimation of solar irradiation by cumulative semivariograms. Sol. Energy 71 (1), 11–21. http://dx.doi.org/10.1016/S0038-092X(01)00009-3.

Shepard, D., 1968. A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM National Conference, pp. 517–524. http://dx.doi.org/10.1145/800186.810616.

Silva-Ramírez, E.L., Pino-Mejías, R., López-Coello, M., Cubiles-de-la-Vega, M.D., 2011. Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Netw. 24 (1), 121–129. http://dx.doi.org/10.1016/j.neunet.2010.09.008.

Srebotnjak, T., Carr, G., de Sherbinin, A., Rickwood, C., 2012. A global Water Quality Index and hot-deck imputation of missing data. Ecol. Indic. 17, 108–119. http://dx.doi.org/10.1016/j.ecolind.2011.04.023.

Strike, K., El Emam, K., Madhavji, N., 2001. Software cost estimation with incomplete data. IEEE Trans. Softw. Eng. 27 (10), 890–908. http://dx.doi.org/10.1109/32.962560.

Sullivan, D., Andridge, R., 2015. A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. Comput. Statist. Data Anal. 82, 173–185. http://dx.doi.org/10.1016/j.csda.2014.09.008.

Team R.C., 2013. R: A language and environment for statistical computing.

1999. Thrombosis dataset for discovery challenge. In: 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, September 15–18, Prague, Czech Republic. http://lisp.vse.cz/pkdd99/ (Accessed: 01.01.06).

Tsumoto, S., 1999. Guide to the medical data set. In: (Berka P. ed.) PKDD.99 Workshop Notes on Discovery Challenge, Prague, pp. 45–47.

Van der Heijden, G.J., Donders, A.R.T., Stijnen, T., Moons, K.G., 2006. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J. Clin. Epidemiol. 59 (10), 1102–1109. http://dx.doi.org/10.1016/j.jclinepi.2006.01.015.

Zhao, Y., 2011. R and Data Mining: Examples and Case Studies.

Zimmerman, , Donald, W., 1997. A note on interpretation of the Paired-Samples t Test. J. Educ. Behav. Stat. 22 (3), 349–360.