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