Knowledge-Based Systems 37 (2013) 37–47

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

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

Optimum estimation of missing values in randomized complete block design by genetic algorithm

A. Azadeh [a,b,*], S.M. Asadzadeh [a,b], R. Jafari-Marandi [a,b], S. Nazari-Shirkouhi [a,b], G. Baharian Khoshkhou [c], S. Talebi [d], A. Naghavi [a,b]

[a] Department of Industrial Engineering, Center of Excellence for Intelligent Experimental Mechanics, University of Tehran, P.O. Box 11365-4563, Iran
[b] Department of Engineering Optimization Research, College of Engineering, University of Tehran, P.O. Box 11365-4563, Iran
[c] Department of Mechanical and Industrial Engineering, University of Illinois, Urbana–Champaign, USA
[d] Department of Industrial Engineering, North Carolina State University, Raleigh, USA

ARTICLE INFO

Article history:
Received 18 March 2011
Received in revised form 20 May 2012
Accepted 25 June 2012
Available online 17 July 2012

Keywords:
Missing values
Genetic algorithm (GA)
Artificial Neural Network (ANN)
Particle swarm optimization (PSO)
Regression methods
Complete randomized block design

ABSTRACT

Missing data are a part of almost all research, and we all have to decide how to deal with it from time to time. There are a number of alternative ways of dealing with missing data. The problem of handling missing data has been treated adequately in various real world data sets. Several statistical methods have been developed since the early 1970s, when the manipulation of complicated numerical calculations became feasible with the advancement of computers. The purpose of this research is to estimate missing values by using genetic algorithm (GA) approach in a randomized complete block design (RCBD) table and to compare the computational results with three other methods, namely, particle swarm optimization (PSO), Artificial Neural Network (ANN), approximate analysis and exact regression method. Furthermore, 30 independent experiments were conducted to estimate missing values in 30 RCBD tables by GA, PSO, ANN, exact regression and approximate analysis methods. Computational results indicated that the best answer (in the last 10-chromosome population) obtained by GA is frequently the same as the missing value, with the mean value being close to the missing observation. It is concluded that GA provides much better estimation than the other methods. The superiority of GA is shown through lower error estimations and also Pearson correlation experiment. Therefore, it is suggested to utilize GA approach of this study for estimating missing values for RCBD.

1. Introduction

Missing data are a part of almost all research, and we all have to decide how to deal with it from time to time. Data values may be absent from a dataset for various reasons, for example, the inability to measure certain attributes. In such cases, the most popular and simple method of handling missing data is to ignore either the projects or the attributes with missing observations. This technique causes the loss of valuable information and therefore may lead to inaccurate cost estimation models. Gad and Ahmed [24] claimed that ignoring the missing values in this case leads to biased inferences. Moreover, when an attribute contains a missing value in a test case, it may or may not be worthwhile to take the extra effort in order to obtain a value for that attribute(s). There are a number of alternative ways of dealing with missing data.

In this paper, we employ GA to estimate missing values in RCBD table and compare the results of GA with two other soft computing techniques namely PSO and ANN. The RCBD table has n (with index $i = 1,2,\ldots,n$) treatments and m (with index $j = 1,2,\ldots,m$) randomized blocks. The main assumptions of RCBD table are: (i) observations are adequately described by the model; (ii) errors are normally and independently distributed with mean zero and constant but unknown variance δ^2. Once we perform a method to estimate missing values in such table, for each cell of this table say (i,j) we assume the cell value to be missed. We use the methods to estimate all values in cells and compare the methods based on the errors in estimating all missed values in a table. Moreover, we use the conventional methods of approximate analysis and exact regression to estimate the missing value and compare the results to show the superiority of the proposed GA method. The main contribution of the paper is to use specific type of GA coding which improves the quality of missing value estimation in comparison with other conventional (i.e. approximate and regression) and soft computing techniques (i.e. PSO and ANN). Since the algorithm assumes that a cell in RCBD table is missing, the estimated missing value can be compared with the actual one. Once this estimation

* Corresponding author at: Department of Industrial Engineering, Center of Excellence for Intelligent Experimental Mechanics, University of Tehran, P.O. Box 11365-4563, Iran.
E-mail addresses: ali@azadeh.com, aazadeh@ut.ac.ir (A. Azadeh).

0950-7051/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2012.06.014

procedure is repeated for all the cells of a RCBD table, the correlation of the estimated values and actual values is the criteria to choose between alternative methods.

The structure of the paper is as follows: In Section 2, the related work and the latest literature related to the problem of missing value estimation is outlined. In Section 3, the different mechanisms used to create missing data and the most common techniques for handling them are described. Section 4 explains the general outline of soft computing techniques namely, GA, PSO and ANN. In Section 5, the specific approaches used for GA-coding, PSO-coding and ANN structure are expressed. Section 6 is designated to the computational results. Finally, Section 7 is a conclusion to this research.

2. Literature review

The problem of handling missing data has been treated adequately in various real world data sets. Several statistical methods have been developed since the early 1970s. Some of the most important review papers on the subject are Afifi and Elashoff [2], Hartley and Hocking [27], Dempster et al. [15], Little and Rubin [35], Little and Rubin [36]. Zhang [66] proposed a new learning algorithm which induces decision trees from training datasets with missing data. Missing value techniques may be used to make classification and learning algorithms robust to the missing values (see [14]).

In the field of software engineering there are rather few published works concerning missing data. In Song and Shepperd [55], two imputation methods, class mean imputation (CMI) and k-nearest neighbors (k-NN), were considered with respect to two mechanisms of creating missing data: missing completely at random (MCAR) and missing at random (MAR). Bashir et al. [6] introduced a novel partial matching concept in association rules mining to improve the accuracy of missing values imputation. Their imputation technique combined the partial matching concept in association rules with k-nearest neighbor approach. Bras and Menezes [7] introduced a modification of the weighted k-nearest neighbors imputation method (k-NN Impute) for missing values (MVs) estimation in microarray data based on the reuse of estimated data. The model called iterative k-NN imputation (IKNN-Impute) as the estimation is performed iteratively using the recently estimated values. Huang and Lee [31] presented a gray-based nearest neighbor approach to predict accurately missing attribute values. First, gray relational analysis was employed to determine the nearest neighbors of an instance with missing attribute values. Accordingly, the known attribute values derived from these nearest neighbors were used to infer those missing values. Cartwright et al. [9] examined sample mean imputation (SMI) and k-NN on two industrial data sets with real missing data. They found that both methods improved the model fit but k-NN gave better results than SMI.

Hrydziuszko and Viant [30] demonstrated that missing data estimation algorithms have a major effect on the outcome of data analysis when comparing the differences between biological sample groups, including ANOVA. They assessed eight algorithms for their ability to impute known, but labelled as missing, entries and found that the k-nearest neighbor imputation method (KNN) as the optimal missing value estimation approach for direct infusion mass spectrometry datasets.

Sentas et al. [52] used multinomial logistic regression (MLR) method for the categorical missing data imputation for software cost estimation by multinomial logistic regression. The common practice is to replace missing data using predictions from a regression model. For binary variables, the prediction is a probability of 1 versus 0 and the imputed value is 1 or 0 according to that probability. Since the logistic regression is used to model the probability of binary outcomes, a reasonable extension is to use MLR for more than two categories, as a generalization of the logistic regression.

In Reis and Saraiva [43], the problem of extending the multiscale decomposition framework based upon the wavelet, transformed to situations where datasets contain any type of missing data patterns (e.g., random, multivariate). Their proposed approaches integrate data uncertainty information into their algorithms to explore all knowledge available about data during the decomposition stage. These frameworks, called generalized multi-resolution decomposition frameworks (GMRD), also lead to new developments in data-analysis tools based upon the information they provide.

Sehgal et al. [51] presented an Ameliorative Missing Value Imputation (AMVI) technique which has ability to exploit global/local and positive/negative correlations in a given dataset by automatic selection of the optimal number of predictor genes k using a wrapper nonparametric method based on Monte Carlo simulations. The AMVI technique has CMVE strategy at its core since CMVE has demonstrated improved performance in comparison to low variance methods like BPCA, LLS-Impute, and high variance methods such as k-NN and Zero-Impute, as CMVE exploits positive/negative correlations. Zhang et al. [67] proposed a sequential local least squares imputation (SLLS-Impute) method estimating missing values sequentially from the gene containing the fewest missing values and partially utilizing these estimated values. Imtiaz and Shah [32] considered the pre-treatment and data analysis as a collective problem and proposed data conditioning methods in a multivariate framework in their study inasmuch as most modeling and data analysis methods are developed to analyze regularly sampled and well conditioned data sets. Qin et al. [41] contend that one way to deal with missing values is to use correlations between the attributes of the data; yet identifying relations within data containing missing values is difficult. Hence, they developed a kernel-based missing data imputation, parameter optimization method (POP algorithm), attempting to make an optimal inference on statistical parameters: mean, distribution function, and quantile after missing data are imputed. They demonstrated that their POP algorithm (random regression imputation) is much better than deterministic regression imputation in efficiency and generating an inference on the above parameters.

Ragel and Cremilleux [42] proposed an external method, MVC (Missing Values Completion), to improve performances of completion and also declarativity and interactions with the user. The core of MVC is the Robust Association Rules (RARs) algorithm. Shen et al. [53] presented a Fast Recycle Combined Association Rules (FRCARs) method to fill in the missing values applying a technique to recycle sub-frequent item sets and bit-arrays to discover more association rules than the Missing Value Completion (MVC) approach. An efficient algorithm, X-Miner, for mining association rules and frequent item sets in databases with missing values was introduced by Calders et al. [8]. They empirically evaluated X-Miner and showed that it gains over a straightforward baseline-algorithm.

Two commonly used approached are *Maximum Likelihood and Multiple Imputation*. There are alternative maximum likelihood estimators that assume an underlying model (usually the multivariate normal distribution) for the distribution of variables with missing data. Then maximum likelihood estimates of the variance, covariance and means of the variables are taken from the existing data, perhaps using listwise deletion. Maximum likelihood then uses these estimates to solve for the model parameters (e.g. regression coefficients) and then estimate missing data based on the estimated model. An iterative approach usually adopted to have the optimal parameters estimation.

Schneider [50] presented a regularized EM algorithm in which the expectation maximization (EM) algorithm for Gaussian data, an iterative method both for the estimation of mean values and covariance matrices from incomplete datasets and for the imputa-