Journal of Computational and Applied Mathematics 311 (2017) 704–717

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

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

Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions

Celestino Ordóñez Galán [a], Fernando Sánchez Lasheras [b,*], Francisco Javier de Cos Juez [a], Antonio Bernardo Sánchez [b]

[a] Department of Mining Exploitation and Prospecting, University of Oviedo, c/Independencia 13, University of Oviedo, 33004 Oviedo, Spain

[b] Department of Construction and Manufacturing Engineering, University of Oviedo, 33204 Gijón, Spain

HIGHLIGHTS

- A genetic algorithm for missing data imputation is proposed.
- The algorithm is tested in the context of the item response theory.
- Optimum parameters of the algorithm are analyzed.
- The proposed algorithm performs better than MICE algorithm.

ARTICLE INFO

Article history:
Received 9 November 2015
Received in revised form 7 August 2016

Keywords:
Imputation method
Item response theory
Genetic algorithms
Multivariate imputation by chained equations (MICE)
Missing data

ABSTRACT

This article proposes a new missing data imputation method based on genetic algorithms. The algorithm presented in this paper is a useful tool for the completion of missing data in knowledge and skills tests. This algorithm uses both Bayesian and Akaike's information criterions as fitness functions and applies them to the classical item response theory models of one, two and three parameters. The results obtained by this new algorithm have been compared with those achieved by means of the Multivariate Imputation by Chained Equations (MICE) algorithm. For all the missing data ratios checked, the average incorrect imputation percentages obtained with the GA algorithm were, statistically, significantly lower than the results obtained with the MICE method. The most favorable frameworks for the use of the algorithm developed in the present research are those questionnaires in which missing answers would be considered as missing completely at random (MCAR). In other words, those questionnaires in which the same questions are present for all the examinees, but not necessarily in the same order.

1. Introduction

Nowadays, most of the empirical studies that are performed in any branch of science and technique must deal with the problem of missing data. In many cases, researchers opt for either removing observations with missing values without reporting them in their studies or imputing them by means of simple techniques such as replacing the unknown values for the median or mean of the corresponding variables, or even creating a specific category for missing values [1]. However, in recent years more accurate and sophisticated methods have been developed [2–4].

* Corresponding author.
E-mail address: sanchezfernando@uniovi.es (F. Sánchez Lasheras).

http://dx.doi.org/10.1016/j.cam.2016.08.012
0377-0427/

In psychometric paradigms such as Item Response Theory, missing data are quite common and pose a problem. For example, in order to perform the calculation of the difficulty and discrimination values of a set of questions in any model of the item response theory [5], it is necessary to substitute missing data by reasonable estimates of the missing values. The alternative to the missing data imputation involves the removal of either those students that do not answer all the items, or the removal of those items that were not answered by all the students. Any mixed procedure that removes both students and items in order to avoid missing values is also possible, but these methods usually lead to the loss of a great amount of valuable information. In general, the loss of information has a pernicious effect on the statistical analyses of the database and would cause biased parameter estimates and the inflation of standard errors [6].

The Item Response Theory (IRT) is an approach to test theory that solves certain problems that were impossible to solve by classical test theory, such as the inadequacy of classical test procedures to detect item bias. IRT was first proposed in the field of psychometrics for the purpose of ability assessment. In other words, IRT is the study of test and item scores based on assumptions concerning the mathematical relationship between abilities and item responses. IRT postulates that an examinee test performance can be explained by a set of factors (called traits or abilities) and that the relationships between examinee item performance and these factors can be described by a monotonically increasing function called item characteristic. This function describes the probability of a given response as a function of a person's true standing on a latent trait or ability. Then, one of the main purposes of IRT is modeling the relationships between ability and a set of items. In IRT persons and items are part of the same continuum. Nowadays it is widely used in education to calibrate and evaluate items in tests, questionnaires, and other instruments and to score subjects on their attitudes and abilities. The basic building block for IRT are the item response functions. The present research employs one, two and three parameters item response models, which are briefly explained in Section 2.1.

The present research proposes a new missing data imputation method based on genetic algorithms (GA) that would be applied to any test created with the purpose of measuring students' knowledge and skills to complete the available data set. In other words, the algorithm presented in this paper is a useful tool for the completion of missing data in knowledge and skills tests. The proposed methodology is able to predict the answers of the students to those questions that have not been answered by them. The results obtained by this new algorithm are compared with those achieved by means of the Multivariate Imputation by Chained Equations (MICE) algorithm [7], which is considered as the benchmark technique. The reason why the MICE algorithm has been chosen is twofold: on the one hand it has demonstrated good performance for this kind of problem [8] and on the other hand, it is currently one of the newest, most promising and widely-applied methods [9].

A classic classification of missing data problems [10] divided them into three categories: missing completely at random (MCAR), in which all data have the same probability of being missing and therefore the probability of being missing is not linked to the data value; missing at random (MAR), in which the probability of being missing is only the same within groups defined by the observed data; and finally, not missing at random (NMAR), which includes all those cases that cannot be classified as either MCAR nor MAR. An example of MCAR would be the case in which, for example, a group of students have to cope with a large question database. In this database, questions are presented to each student in a different, random order and therefore all the questions of the database have the same chance of being presented to the students. This is not true in those cases where questions are always presented in the same order, as due to the large number of questions usually available (please note that some of these databases include more than 5000 questions), it is quite likely that most of the students will not be able to answer all the questions. Most of the computer-based systems that prepare students for many official test-based exams all around the world are more and more frequently becoming computer-based tests. These systems usually present questions in a random order. In this context, missing data is considered unintentional. The reason for this consideration is that during the study process, students are obliged to answer any question that is presented to them, but they are not required to answer all the questions of the database. In other words, and from the point of view of the authors of the present research, in this situation questions that are not answered by a student would be considered as MCAR as they are going to be exposed to them at different moments of the preparation process and they have different levels of knowledge.

The paper is structured as follows: Section 2 describes the models normally used in the item response theory to analyze the responses to questionnaires, as they are the models adjusted in the missing imputation data problem we were trying to solve. Also, a brief summary of the principles of GA is presented, followed by an explanation of our GA-based method to impute missing data. Then, the MICE algorithm, used to contrast the results of our method, is briefly explained. Finally, the two sets of data used to test our method are introduced. Section 3 shows the results obtained with our GA-based missing imputation method, and they are compared with those obtained using the MICE algorithm. Finally, Section 4 presents the conclusions of our study.

2. Materials and methods

2.1. The one, two and three parameters item response models

The models proposed by IRT assume that there is a functional relationship between the values of the variable that is measured by the questions and the likelihood of giving a correct answer. The function that represents this probability is called item characteristic curve. In other words, the probability of hitting a question only depends on the parameters measured by the question. Therefore, subjects with different scores on this variable will determine different odds of giving a correct answer to the question.