Information Sciences 233 (2013) 25–35

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

A hybrid method for imputation of missing values using optimized fuzzy *c*-means with support vector regression and a genetic algorithm

Ibrahim Berkan Aydilek *, Ahmet Arslan

Department of Computer Engineering, Selçuk University, Konya, Turkey

ARTICLE INFO

Article history:
Received 1 May 2011
Received in revised form 21 October 2012
Accepted 19 January 2013
Available online 1 February 2013

Keywords:
Missing data
Missing values
Imputation
Support vector regression
Fuzzy *c*-means

ABSTRACT

Missing values in datasets should be extracted from the datasets or should be estimated before they are used for classification, association rules or clustering in the preprocessing stage of data mining. In this study, we utilize a fuzzy *c*-means clustering hybrid approach that combines support vector regression and a genetic algorithm. In this method, the fuzzy clustering parameters, cluster size and weighting factor are optimized and missing values are estimated. The proposed novel hybrid method yields sufficient and sensible imputation performance results. The results are compared with those of fuzzy *c*-means genetic algorithm imputation, support vector regression genetic algorithm imputation and zero imputation.

1. Introduction

Missing values are highly undesirable in data mining, machine learning and other information systems [33]. In recent years, much research has been regarding missing value estimation and imputation has been performed [3,9,24,33,35,47,49]. To deal with missing values in datasets: ignoring, deleting, zero or mean estimation methods might be used instead of imputation methods [7,30]. However, the primary disadvantages of these estimation methods are the loss of efficiency due to discarding incomplete observations and biases in estimates when data are missing in a systematic manner [35]; these disadvantages reduce data quality. Quality data mining results can be obtained only with high quality data [37,41]. Therefore, missing values should be estimated to increase data quality. Missing values typically occur because of sensor faults, a lack of response in scientific experiments, faulty measurements, data transfer problems in digital systems or respondents' unwillingness to respond to survey questions [1,27,31,32,36]. In scientific research, especially in psychology, data for some variables in the database to be analyzed may be missing. If the missing values are not treated correctly, they may decrease or even jeopardize the validity of the research [3,5,14,22,34].

2. Literature review

This section presents a brief summary of the studies related to support vector regression imputation and fuzzy *c*-means imputation.

* Corresponding author. Tel.: +90 3322233333.
E-mail addresses: berkan@selcuk.edu.tr (I.B. Aydilek), ahmetarslan@selcuk.edu.tr (A. Arslan).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.01.021

Abdella et al. studied the use of genetic algorithms and neural networks to approximate missing data in databases [1]. Deogun et al. utilized the clustering method with soft computing, which is more tolerant of inaccuracy and uncertainty, and they applied a fuzzy clustering algorithm to treat incomplete data [9]. Liao et al. presented a fuzzy k-means clustering algorithm that uses a sliding window for the imputation of incomplete data to improve data quality [24]. Pelckmans et al. proposed an alternative approach, in which no attempt is made to reconstruct the values that are missing, but the impact of the missing data on the result and the expected cost are modeled using support vector machines. The approach is to assume some models for the covariates of missing values and then use a maximum likelihood approach to obtain the estimates for these models. The advantage of this approach is that classification rules can be learned from observational data even when missing values occur amongst the input variables, whereas the disadvantage is that the proposed model aims for high classification accuracy rather than high imputation accuracy for the missing values [35]. Lim et al. proposed a hybrid neural network that uses fuzzy ARTMAP and fuzzy *c*-means clustering for pattern classification using incomplete training and test data. One of the disadvantages of fuzzy ARTMAP is that it is very sensitive to the arrangement of the training data. Fuzzy ARTMAP is also acutely sensitive to the selection of the vigilance parameter because determining the optimal value for the vigilance parameters can be quite difficult [25]. Hathaway et al. introduced an approach for clustering that is based on incomplete dissimilarity data. An advantage of the method is that fuzzy *c*-means are regarded as a reliable clustering algorithm for incomplete data [20]. Feng et al. also presented a support vector regression (SVR)-based imputation method that uses an orthogonal coding scheme to estimate missing values for DNA microarray gene expression data. A comparative study of their method with those previously developed, such as the K nearest neighbor and Bayesian principal component analysis imputation methods indicated that the SVR method is effective in imputation. A significant advantage of the SVR model is that it requires less computational time, but our hybrid SVR clustering technique yields more sensible results for outlier values [47]. Timm et al. noted that incomplete datasets are a significant problem in data analysis. They introduced a class-specific probability for missing values to assign incomplete data points to clusters appropriately [44]. Farhangfar et al. aimed to provide a comprehensive review of representative imputation techniques. The use of a low-quality single-imputation method yielded imputation accuracy comparable to the accuracy achieved when one utilizes some other advanced imputation techniques [11]. Li et al. assumed that missing attributes are represented as intervals, and they proposed a novel fuzzy *c*-means algorithm for incomplete data based on nearest neighbor intervals. The disadvantage of this method is that there is no theoretical basis for the selection of the cluster (c) number, so further research is needed to investigate this problem [23]. Nuovo compared fuzzy *c*-means (FCM) imputation with case deletion imputation. The comparison was made in a psychological research environment, using a database of mentally retarded in patients. The results indicated that completion techniques, in particular the technique based on FCM, yield effective data imputation and help avoid the deletion of elements with missing data that diminish the power of the research. Fuzzy *c*-means (FCM) imputation is more accurate than regression imputation (RI) and expectation maximization estimation (EME). However, a major disadvantage is that the FCM implementation uses a weighting factor (m) parameter value, which is equal to 2, and the value should be adapted to the dataset type [34].

2.1. Missing data

There are three types of missing data described in the literature [26]:

1. Missing completely at random (MCAR) – The missing value has no dependency on any other variable.
2. Missing at random (MAR) – The missing value depends on other variables. The missing value can be estimated using other variables.
3. Missing not at random (MNAT) – The missing value depends on other missing values, and thus missing data cannot be estimated from existing variables.

In this paper, we assume that the data are MAR, which implies that the missing values are deducible in some complex manner from the remaining data [38]. In Table 1, we present a section of a dataset with missing values. In this paper, we aim to estimate missing values using fuzzy *c*-means optimized with support vector regression and a genetic algorithm. These notations will be used in the rest of the paper: Y1, Y2, Y3, Y4, Y5 and Y6 are records (rows). X1, X2, X3, X4 and X5 are attributes (columns). Y2, Y5 and Y6, which do not have any missing values, are 'complete' rows, and Y1, Y3 and Y4, which have missing values, are called 'incomplete' rows.

Table 1
A section of a dataset with missing values.

	X1	X2	X3	X4	X5
Y1	0.113524	0.084785	?	0.625473	0.06385
Y2	0.112537	0.138211	0.15942	0.625473	0.068545
Y3	0.110563	?	0.144928	0.624212	0.083568
Y4	0.110563	0.170732	0.146998	0.623581	?
Y5	0.108588	0.129501	0.144928	0.624212	0.076056
Y6	0.108588	0.082462	0.112836	0.626103	0.015023