Knowledge-Based Systems 000 (2017) 1–15

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases

R. Devi Priya[a,*], R. Sivaraj[b], N. Sasi Priyaa[c]

[a] Department of Information Technology, Kongu Engineering College, Perundurai, Erode, Tamil Nadu 638 060, India
[b] Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Erode, Tamil Nadu, India
[c] Department of Computer Science and Engineering, Kongu Engineering College, Perundurai, Erode, Tamil Nadu 638 060, India

ARTICLE INFO

Article history:
Received 15 May 2016
Revised 23 June 2017
Accepted 26 June 2017
Available online xxx

Keywords:
Missing values
Heterogeneous attributes
Ant colony optimization
Bayesian methods
Repopulation

ABSTRACT

The incomplete datasets with missing values are unsuitable for making strategic decisions since they lead to biased results. This problem is even worse when the dataset is large and collected from many heterogeneous sources. The paper deals with missing scenarios which were not dealt together earlier. The proposed Dual Repopulated Bayesian Ant Colony Optimization (DPBACO) handles both ignorable and non-ignorable missing values in heterogeneous attributes of large datasets The DPBACO integrates Bayesian principles with Ant Colony Optimization technique since both are simple and efficient to implement. After pheromone updation, repopulation of the solution pool is done by dividing the population into two based on their fitness values and generating new offsprings by performing crossover operation. The DPBACO algorithm is implemented on six large mixed-attribute datasets for imputing both kinds of missing values. The empirical and statistical results show that DPBACO performs better than other existing methods at variable missing rates ranging from 5% to 50%.

1. Introduction

1.1. Missing data

Today's database capacity of some organizations is in PetaBytes (PB) with enormous amount of data. Data being the significant asset for many organizations, its quality is expected to be high in order to use it for making strategic decisions. With much advancement in information sciences, size of the data grows very rapidly and data analysts encounter serious threats in handling huge volume, velocity, variety and verocity of data. There issues in their storage and retrieval also increases day by day. The common challenges encountered by researchers in handling large databases are data with low quality, scalability, high dimensionality, heterogeneity, complex nature of data, etc. Data for such large databases is collected from multiple heterogeneous sources in different formats. While integrating multiple sources into a single source, many problems arise including data loss and missing values. The values may also be lost or missing due to instrumental errors, manual mistakes, non- response in surveys, etc. Whatever may be the reason for missingness, missing values adversely affect the knowledge extracted from them and the main consequences are: (i) the incomplete dataset looks different from the original one and the samples may often become non-representative. (ii) Due to limited available information, statistical power is lost to a great extent. (iii) Most of the commonly used statistical methods require complete data for analysis and hence they cannot be applied successfully in incomplete datasets and (iv) Missing values may negatively distort the estimates.

Depending upon its nature, the missingness is divided into ignorable and non-ignorable missingness. Ignorable missingness may further include Missing At Random (MAR) and Missing Completely At Random (MCAR) values. When the values are MAR, the missing values can be recovered using other dependent attributes in the dataset whereas when the values are MCAR, the missingness ie entirely random and does not have any influence over other attributes. Not Missing At Random values fall under non-ignorable pattern where the missingness cannot be ignored and the missing values need to be imputed. Here, the values which are missing are themselves reasons for missingness.

Many researchers have proposed methods for dealing with missing values in either discrete or continuous attributes. But, real datasets often contain mixed type of attributes and there is no standard solution till now. Since nowadays all domains have huge databases with different formats of inputs from multiple sources, missing values are very common and hence an efficient method

* Corresponding author.
E-mail addresses: scrpriya@gmail.com (R. Devi Priya), rsivarajcse@gmail.com (R. Sivaraj), nsasipriyaa@gmail.com (N. Sasi Priyaa).

http://dx.doi.org/10.1016/j.knosys.2017.06.033
0950-7051/

to deal with them is absolutely essential. Many literatures have successfully used Bayesian methods for treating both discrete and continuous missing values under all missing scenarios.

Estimation of missing values is like an approximation and the evolutionary algorithms like Genetic Algorithms (GA), Ant Colony optimization (ACO), Particle Swarm optimization (PSO) can be effectively used. Among the evolutionary algorithms, ACO, a population based meta-heuristic algorithm used to find approximate solutions for optimization problems is observed to be better for some problems ([28]; [42]). ChandraMohan and Baskaran [11] have done a detailed study about the implementation of ACO for different kinds of problems in engineering domain. The searching capability of ACO can be improved if it is hybridized with other searching methods. In Aghdam et al. [2], ACO and Bayesian methods are combined for feature selection in a large bioinformatics dataset. ACO is applied to estimate the missing values in Betechuoh and Marwala [7]. Borrottia et al. [8] have hybridized Naïve Bayes classifier with Ant Colony Optimization (NACO) for solving problems with high dimensions. They have implemented NACO for designing artificial enzyme structures. Galan et al. [20] have imputed the missing values in questionnaires using a method based on Genetic Algorithms. They have evaluated the fitness of solutions using Bayesian and Akaike's information. Its imputation results are better than that of existing Multivariate Imputation by Chained Equations (MICE) algorithm.

1.2. Contributions of the paper

This paper introduces a new method called as Repopulated Bayesian Ant Colony Optimization where ACO is hybridized with Bayesian principles to impute both discrete and continuous missing values under MAR, MCAR and NMAR patterns. Bayesian methods are simple and efficient in missing value imputation. ACO is successfully implemented for solving many optimization problems. Since missing data imputation is a kind of optimization where accurate values to be substituted are searched, it is used in our methodology. Yet sometimes, they lack in reaching the global optima due to insufficient exploration and exploitation. In order to provide improve exploration and exploitation performance of DPBACO, the solution space is repopulated with new inputs using local beam search during the imputation if samples in the solution space are not satisfactory. The proposed algorithm is applied to impute missing values in real datasets taken from standard repositories under MAR, MCAR and NMAR patterns at different missing proportions and the results are found to be better than the existing algorithms which are compared.

1.3. Organization of the paper

The rest of the paper is organized as follows. Section 2 gives an overview about various literatures that have dealt with missing value imputation. Section 3 provides a brief introduction about Ant Colony Optimization algorithm which is used in this paper. Section 4 briefly enumerates about the proposed Repopulated Bayesian Ant Colony Optimization methodology. Section 5 shows the implementation results and discusses about them. Section 6 concludes the paper and throws directions for future research.

2. Background

Acuna and Rodriguez [1] and Sanders et al. [46] have alerted analysts that missing data in large datasets adversely affects the accuracy of results. In order to avoid biased inferences, missing data is generally handled by any of these following methods: (i). Ignoring the records with missing values (ii). Single value substitution (iii). Statistical techniques or (iv). Evolutionary algorithms.

2.1. Ignoring missing records

Litwise deletion and pairwise deletion are the common techniques under this category where records with missing values are removed from the dataset and only the records with complete values are retained. Ignoring the records will result in loss of valuable information and only less amount of information will be available for analysis which reduces the statistical power and hence not recommended by researchers [47].

2.2. Single value substitution

A single value (zero, default, random or mean/mode value) is substituted in all the missing places. It assumes that all the missing instances behave similarly and tries to fill the same value. The actual values in different places would have been different and this deviation causes distortion in variance of the values. These methods are simple to implement but when accuracy is concerned, they are not generally preferred [47].

2.3. Statistical methods

Statisticians often handle missing values rather than making decisions from incomplete datasets. After filling the missing values, analysis is done on the imputed dataset using standard techniques that are used for complete dataset. But while making any analysis, it is to be considered that there is always some degree of uncertainty associated with the imputed values and improvements need to be made in standard data analysis methods. Depending upon the type of attributes also, the technique needed may vary. For example, the technique applied for discrete case cannot be easily applied for continuous case. It stands as a risk factor for analysts because most of the available techniques were designed for handling either discrete or continuous case. The methods and assumptions used for missing value imputation are discussed in Soley-Bori [55].

Multiple Imputation (MI) is widely accepted as the standard imputation method and is freely available in many statistical packages and by default, it assumes MAR condition. However, recent versions of MI are available for MCAR and NMAR conditions [62]. In Hardt et al. [23], the authors have found that the performance of MI depends on percentage of missing data, number of samples and variables in the model used. Expectation Maximization (EM) algorithm is also an appreciated Maximum Likelihood (ML) imputation method for MAR with 2 iterative steps: (i). Expectation (estimation of the possible values) and (ii). Maximization (maximizing likelihood of the estimated values) [17]. However, the scope of basic EM algorithm is very limited since it demands more computational resources and are not suitable for large datasets. The algorithm needs to be enhanced in different perspectives for dealing with different missing scenarios [57].

kNN (k Nearest Neighbor) is yet another important method commonly used for missing value imputation. It analyzes k nearest neighbors of the missing instance to identify the value to be substituted. Hron et al. [27] have used different versions of kNN for imputing incomplete compositional data. Liao et al. [37] have discussed about the cases where missing values need to be imputed and proposed four versions of kNN for imputation of mixed variables in large phenomic data. In Tutz and Ramzan [60], the authors have proposed a weighted kNN (wNN) to estimate the missing values. The distance between predictor variables is used to evaluate the closest neighbors. By cross validation, the tuning parameter values are updated. Datta et al. [15] have recently proposed a method to perform kNN classification by using feature weighted

Please cite this article as: R. Devi Priya et al., Heuristically repopulated Bayesian ant colony optimization for treating missing values in large databases, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.06.033