Information Sciences xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Information Sciences

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

Counter propagation auto-associative neural network based data imputation

Q1 Chandan Gautam[a,b], Vadlamani Ravi[a,*]

[a] Center of Excellence in CRM and Analytics, Institute for Development and Research in Banking Technology, Castle Hills Road No. 1, Masab Tank, Hyderabad, 500057, AP, India
[b] School of Computer & Information Sciences, University of Hyderabad, Hyderabad, 500046, AP, India

ARTICLE INFO

Article history:
Received 26 July 2014
Revised 31 May 2015
Accepted 4 July 2015
Available online xxx

Keywords:
Data imputation
Auto-associative neural network
Counter propagation auto associative neural network (CPAANN)
Grey system theory (GST)

ABSTRACT

In this paper, we propose two novel methods viz., counterpropagation auto-associative neural network (CPAANN) and grey system theory (GST) hybridised with CPAANN for data imputation. The effectiveness of these methods is demonstrated on 12 datasets and the results are compared with that of various extant methods. Wilcoxon signed rank test conducted at 1% level of significance, indicated that the proposed methods are statistically significant against all methods. The spectacular success of CPAANN can be attributed to the local learning, global approximation and auto-association that take place in tandem in a single architecture. Furthermore, significantly CPAANN turned out to be the best in the class of AANN architectures used for imputation. The reason could be the competitive learning that is intrinsic to the CPAANN architecture, but conspicuously absent in other auto-associative neural network architectures.

1. Introduction

In many real world datasets, occurrence of missing data is a common phenomenon. There could be several reasons for data to be missing such as non-response by respondents to some fields during data collection because of privacy concerns, data entry errors, system or machine failure, ambiguity of the survey questions, etc. While survey data are just one example of the prevalence of missing data, it is a common problem in various other fields too as discussed below. A longitudinal study [25,89] is an observational research method in which data are gathered for the same subjects repeatedly over several years or even decades. In this case, subjects drop out because they move or suffer side effects from drugs or for other often unknown reasons resulting in missing data. In wireless sensor network [55,56], missing data occurs because some sensors do not respond properly. In geosciences, data items in the observational datasets may be missing altogether, or imprecise [35]. Datasets for effort prediction in software project management contain missing values [84]. Geophysical time series datasets also contain missing data [79]. Reasons such as equipment malfunctioning, outliers and incorrect data entry contribute to missing values in many practical observations [13]. Due to faults in the data acquisition process, data tend to be missing in environmental research data sets. In automatic speech recognition, speech samples, corrupted by very high levels of noise, are considered to be missing data [18]. Datasets for business and financial applications may also contain missing data. Missing data problems are common in health research (e.g. retrospective and prospective studies). In biological research with DNA microarrays, gene data may be missing due to reasons such as a scratch on the slide containing the gene sample and contaminated samples [88].

* Corresponding author. Tel.: +914023294042; fax: +914023535157.
E-mail addresses: induindu31@gmail.com (C. Gautam), rav_padma@yahoo.com, padmarav@gmail.com (V. Ravi).

http://dx.doi.org/10.1016/j.ins.2015.07.016
0020-0255/

The missing data poses major challenges for analysts because the fundamental requirement of analysing any data is its completeness. Furthermore, most of the data mining algorithms cannot work with incomplete datasets. Therefore, imputation, the process of substituting a missing data point or a missing component of a data point by a suitable value, becomes essential [1,2,28,65].

According to Kline [51], the remedial methods include deletion, imputation, model based and machine learning methods, which are briefly described as follows:

(1) Deletion procedures
These techniques simply delete the cases or records that contain missing data. There are two types in deletion approaches: list wise deletion and pair wise deletion [84]. The former ignores the records containing missing values, while the latter considers each feature separately and missing data is ignored for that feature.
(2) Imputation procedures
These techniques include regression imputation, hot and cold deck imputation, multiple imputation and mean imputation. The regression equations are computed each time by considering the feature containing incomplete value as the target variable. In hot and cold deck imputation, the missing values are replaced by the closest components that are present in both vectors for each case with a missing value [78]. In mean imputation, however, the missing values of a variable are replaced by the mean value of all the remaining records of that variable. In multiple imputations, on the other hand, we can make combined inferences by analysing N complete datasets after replacing each value N times.
(3) Model-based procedures
The maximum likelihood estimation method and expectation maximisation fall in this category. The maximum Likelihood approach assumes that the observed data are a sample drawn from a multivariate normal distribution and the parameters are estimated based on available data. Then, the missing values are imputed based on these parameters [21].
(4) Machine learning methods
These include multilayer perceptron (MLP), K-nearest neighbour (K-NN) and many others, which are reviewed in Section 2.

In this paper, we propose two novel imputation methods under machine learning category viz., (i) counter propagation auto association neural network (CPAANN) and (ii) a hybrid of CPAANN and grey system theory (GST) and test their effectiveness on a host of 12 datasets used for this purpose in the literature.

The rest of the paper is organised as follows: a brief review of literature on imputation of missing data is presented in Section 2. The motivation for the proposed architecture is presented in Section 3 followed by overview of the proposed architecture in Section 4. The description of the datasets and experimental design are presented in Section 5. Results and discussions are presented in Section 6 while Section 7 concludes the paper.

2. Literature review

Literature abounds with several imputation methods under the machine learning and soft computing category. These include self-organising map (SOM) [62], K-nearest neighbour [9], MLP [33], fuzzy neural network [27], auto-associative neural network hybridised with the genetic algorithm [1], etc. While Batista and Monard [9,10] and Jerez et al. [48] used K-NN for imputing missing data, Liu and Zhang [57] developed the mutual K-NN algorithm for classifying incomplete and noisy data. Samad and Harp [77] implemented SOM for imputing the missing data. Austin and Escobar [4] used Monte Carlo simulations to examine the performance of three Bayesian imputation methods. Many studies [33,66,67,80,82,94] employed MLP for the purpose, where MLP is trained as the regression model by using the complete cases and choosing one variable as target each time. When auto-associative neural network (AANN) is used for imputation, the network is trained for predicting the inputs by taking the same input variables as target variables [58, 59]. Ragel and Cremilleux [71] proposed a missing value completion method, which extends the concept of Robust Association Rules Algorithm (RAR) for databases with multiple missing values. Chen et al. [17] employed selective Bayes classifier for classification on incomplete data. Nouvo [68] employed Fuzzy C-means for data imputation. Elshorbagy et al. [24] employed the principles of chaos theory to estimate the missing stream flow data. Dempster et al. [19] designed the Expectation Maximisation (EM) algorithm by using correlated data variables to estimate missing observations in multivariate data. Figueroa et al. [28] proposed a method to impute missing observations in multivariate data using a genetic algorithm that minimises an error function derived from their covariance matrix and vector of means. Ankaiah and Ravi [2] proposed a hybrid two stage imputation method, where the K-means algorithm and MLP are employed in stage 1 and stage 2 respectively. Zhang [57] employed grey system theory (GST) based imputation by using the K-NN algorithm.

Recently, Nishanth et al. [65] extended the two-stage soft computing approach of Ankaiah and Ravi [2] for data imputation to assess the severity of phishing attacks obtaining improved results. Dove et al. [22] used recursive partitioning; Kang [50] proposed locally linear construction; Garcia-Laencina et al. [29] proposed a multi-task learning based method for training and operating a modified MLP; Duma et al. [23] proposed the hybrid multi-layered artificial immune system and GA; Nelwamondo et al. [63] combined dynamic programming, neural networks and GA; Rahman et al. [72] imputed both categorical and numerical missing values using decision trees and forests; Aydilek et al. [5] hybridised Fuzzy C-means (FCM), support vector regression (SVR) and GA for imputation. Tian et al. [87] proposed the hybrid imputation technique involving GST and Entropy based clustering. Nishanth and Ravi [64] proposed four hybrid imputation methods; one online and 3 offline methods. They employed ECM with

Please cite this article as: C. Gautam, V. Ravi, Counter propagation auto-associative neural network based data imputation, Information Sciences (2015), http://dx.doi.org/10.1016/j.ins.2015.07.016