Neurocomputing 156 (2015) 134–142

Contents lists available at ScienceDirect

Neurocomputing

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

Brief papers

Data imputation via evolutionary computation, clustering and a neural network

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

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

ARTICLE INFO

Article history:
Received 29 August 2014
Received in revised form
28 November 2014
Accepted 26 December 2014
Communicated by: S. Mitra
Available online 8 January 2015

Keywords:
Covariance matrix
Particle swarm optimization (PSO)
Data imputation
Evolving clustering method (ECM)
Extreme learning machine (ELM)

ABSTRACT

In this paper, two novel hybrid imputation methods involving particle swarm optimization (PSO), evolving clustering method (ECM) and autoassociative extreme learning machine (AAELM) in tandem are proposed, which also preserve the covariance structure of the data. Further, we removed the randomness of AAELM by invoking ECM between input and hidden layers. Moreover, we selected the optimal value of *Dthr* using PSO, which simultaneously minimizes two error functions viz., (i) mean squared error between the covariance matrix of the set of complete records and that of the set of total records, including imputed ones and (ii) absolute difference between the determinants of the two covariance matrices. The proposed methods outperformed many existing imputation methods in majority of the datasets. Finally, we also performed a statistical significance testing to ensure the credibility of our obtained results. Superior performance of one of the hybrids is attributed to the power of hybrid of local learning, global optimization and global learning. Both methods resolved a nagging issue of the difficult choice of *Dthr* value and its dominant influence on the results in ECM based imputation. We conclude that the proposed models can be used as a viable alternative to the existing ones for the data imputation.

1. Introduction

Knowledge extraction from the database is always cumbersome for researchers in various disciplines in the presence of missing data. Missing data is an inevitable problem in many disciplines. Because most of the data mining algorithms cannot work with incomplete datasets, imputation of missing data became mandatory [1,6–8]. Various methods have been proposed by many researchers to resolve the missing data problem. According to Kline [9], the methods used for dealing with missing data are (i) Deletion procedure viz., listwise deletion and pairwise deletion [10], (ii) Imputation procedure [11], (iii) Model based procedure [12] and (iv) Machine learning methods.

The remainder of the paper is organized as follows: a brief review of literature on imputation of missing data is presented in Section 2. The proposed method is explained in Section 3. The datasets and experimental design are described in Section 4. Results and discussions are presented in Section 5 followed by the conclusion and future work in Section 6.

2. Literature review

Literature abounds with several methods to handle missing data of numerical attributes. Data imputation techniques are categorized into deletion, imputation, model-based and machine learning/soft computing procedures. The machine learning based methods include self organizing feature map (SOM) [13], K-nearest neighbor [14], multi-layer perceptron [15], fuzzy-neural network [16], auto-associative neural network imputation with genetic algorithms [6] etc. Then, Batista and Monard [14,17] and Jerez et al. [18] used K-NN for imputing missing data. Liu and Zhang [19] developed mutual K-NN algorithm for classifying incomplete and noisy data. Samad and Harp [20] implemented SOM imputing the missing data. Austin and Escobar [21] used Monte Carlo simulations to examine the performance of three Bayesian imputation methods. Many studies [22–27] employed MLP by training it as a regression model on the set of complete records and choosing one variable as a target variable 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 variable as the target [28,29]. Ragel and Cremilleux [30] proposed extended Robust Association Rules Algorithm (RAR) for databases with multiple missing values. Chen et al. [31] employed selective Bayes classifier for classification of incomplete data. Nouvo [32] employed fuzzy c-means for data imputation. Elshorbagy et al.

* Corresponding author. Tel.: +91 4023294042; fax: +91 4023535157.
E-mail addresses: induindu31@gmail.com, rav_padma@yahoo.com (V. Ravi).

http://dx.doi.org/10.1016/j.neucom.2014.12.073
0925-2312/

[33] employed the principles of chaos theory to estimate the missing stream flow data. Dempster et al. [34] proposed the expectation maximization (EM) for the same purpose. Figueroa et al. [7] proposed the use of GA that minimizes an error function derived from their covariance matrix and vector of means. Then, Ankaiah and Ravi [1] proposed a hybrid two stage imputation method involving K-means algorithm and multi-layer perceptron (MLP) in stage 1 and stage 2 respectively. Recently, Nishanth et al. [8] also employed the same two-stage soft computing architecture for imputation in order to assess the severity of phishing attacks with improved results. Most recently, Nishanth and Ravi [2] proposed four hybrid methods – one online and 3 offline methods for imputation. They employed ECM with general regression neural network (GRNN) for online imputation, K-means and K-medoids with GRNN and K-medoids with MLP for offline imputation. Dove et al. [35] used recursive partitioning; Kang [36] proposed locally linear construction; Garcia-Laencina et al. [37] proposed a modified MLP; Duma et al. [38] proposed hybrid multi-layered artificial immune system and GA; Nelwamondo et al. [39] combined dynamic programming, neural networks and GA; Rahman et al. [40] imputed both categorical and numerical missing values using decision trees and forests; Aydilek et al. [41] hybridized fuzzy c-means, support vector regression and GA for imputation.

While Nishanth and Ravi [2] and Gautam and Ravi [5] proposed ECM for the imputation task, their results were influenced by the choice of *Dthr* value. Recently, Krishna and Ravi [3] proposed imputation technique based on PSO and Covariance structure of matrices. Most recently, Ravi and Krishna [4] proposed various online and offline techniques for imputation viz., particle swarm optimization trained auto associative neural network (PSOAANN), particle swarm optimization trained auto associative wavelet neural network (PSOAAWNN), radial basis function auto associative neural network (RBFAANN), general regression auto associative neural network (GRAANN).

3. Proposed methodology

We proposed two imputation methods based on PSO, ECM and extreme learning machine (ELM).

3.1. Overview of particle swarm optimization (PSO), covariance matrix and determinant

Particle Swarm Optimization, a population-based evolutionary computation algorithm based on flocking of birds, was proposed by Kennedy and Eberhart in 1995 [42,43]. PSO operates in three phases: (1) Initialization, (2) Velocity and particle position updation (3) Termination. Two matrices are similar if their covariance matrices are similar. Further, two matrices are similar if the determinants of their covariance matrices are similar also. For more information see [44-47]. So, we used the covariance matrix and the determinant of that covariance matrix in the fitness function of PSO.

3.2. Overview of evolving clustering method

ECM is a one-pass, fast clustering method [48,49], where the number of clusters need not be specified upfront. In ECM, a threshold value, *Dthr* is a user-defined parameter which affects the number of clusters to be estimated [48]. Too large or too small values of *Dthr* adversely impact the number of clusters to be found. Therefore, we need a method to select an optimum *Dthr* value. The method proposed in the next section resolves this issue.

3.3. First proposed method: PSO–ECM

Total data records (X_t) are divided in two parts viz., set of complete records (X_c) to train the model and set of incomplete records (X_{ic}) to test it. As we discussed above, our proposed algorithm will have the same fitness function as in [3]. However, our work is completely different in two ways:

(i) Obtaining the missing value for imputation: Here, we apply the ECM method for imputing missing values whereas they employed PSO for that purpose. Missing values are imputed by the nearest cluster centres, which are in turn obtained by applying ECM on the set of complete records.
(ii) Deployment of PSO for selection of optimum value of *Dthr* in ECM: in Krishna and Ravi [3], the role of PSO is to impute missing values optimally, whereas here PSO is employed to obtain the optimum *Dthr* value.

The algorithm of the proposed method is as follows:

1) Compute the covariance matrix of the set of complete records (X_c).
2) Employ ECM on the set of complete data records (X_c) with *Dthr* value randomly initialized by PSO.
3) Perform ECM based imputation for the set of incomplete records (X_{ic}) as follows:
   Attribute value, say x_k, in an incomplete record is imputed by the corresponding value of the attribute in the centre of the nearest cluster by measuring the Euclidean distance between the incomplete record excluding the missing value and the cluster centres excluding the value in the same position. The Euclidean distance is measured by using the following formula:

$$D_j = \sum_{i=1; i\neq k}^{n} \left|x_i - c_j\right|^2$$

   Where j is the number of cluster centres, and n is the number of complete components in each record.
4) Compute the covariance matrix of the set of total records (X_t) after imputation. If total records (X_t) is the order of ($m \times n$) matrix, then its covariance matrix (T_{cov}) is an $n \times n$ matrix.
   If (MSE (X_{cov}, T_{cov}) $< \varepsilon$) and ($|\mathrm{Det}(X_{cov}) - \mathrm{Det}(T_{cov})| < \varepsilon$) then exit. Otherwise, invoke the PSO for selecting improved *Dthr* value. Where, ε is the prespecified small positive value, MSE (X_{cov}, T_{cov}) is the mean squared error computed between the elements of X_{cov} and T_{cov}, Det(X_{cov}) is the determinant of X_{cov} and Det(T_{cov}) is the determinant of T_{cov}.
5) Repeat the above steps until convergence.
   Compute the mean absolute percentage error (MAPE) [50] value: $\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{x_i - \widehat{x_i}}{x_i}\right|$

Where, x_i is the actual value, $\widehat{x_i}$ is the predicted value and n is the total number of missing values.

Thus, in this paper, PSO is used to minimize the above mentioned two error functions in a nested form. The algorithm is designed to stop only when these two errors become very small across two consecutive iterations. After completion of the process, the model yields an optimum *Dthr* value which not only yields the best imputation but also preserves the covariance structure. Then, replace the missing values using ECM imputation with the optimized *Dthr* value.

3.4. Second proposed method: (PSO–ECM)+MAAELM

3.4.1. Overview of AAELM

Extreme learning machine, proposed by Huang et al. [51,52], is a novel feed forward neural network that requires no updation of weights. We employed AAELM (See Fig. 1), an auto associative version of ELM, on 12 datasets and observed that AAELM yielded different results in different runs for teh same dataset. Sometimes,