Chapter 45
Auto Associative Extreme Learning Machine Based Hybrids for Data Imputation

Chandan Gautam
Institute for Development and Research in Banking Technology, India

Vadlamani Ravi
Institute for Development and Research in Banking Technology, India

ABSTRACT

This chapter presents three novel hybrid techniques for data imputation viz., (1) Auto-associative Extreme Learning Machine (AAELM) with Principal Component Analysis (PCA) (PCA-AAELM), (2) Gray system theory (GST) + AAELM with PCA (Gray+PCA-AAELM), (3) AAELM with Evolving Clustering Method (ECM) (ECM-AAELM). Our prime concern is to remove the randomness in AAELM caused by the random weights with the help of ECM and PCA. This chapter also proposes local learning by invoking ECM as a preprocessor for AAELM. The proposed methods are tested on several regression, classification and bank datasets using 10 fold cross validation. The results, in terms of Mean Absolute Percentage Error (MAPE,) are compared with that of K-Means+Multilayer perceptron (MLP) imputation (Ankaiah & Ravi, 2011), K-Medoids+MLP, K-Means+GRNN, K-Medoids+GRNN (Nishanth & Ravi, 2013) PSO_Covariance imputation (Krishna & Ravi, 2013) and ECM-Imputation (Gautam & Ravi, 2014). It is concluded that the proposed methods achieved better imputation in most of the datasets as evidenced by the Wilcoxon signed rank test.

INTRODUCTION

Missing data can be observed in many datasets, which have been collected in real time. It can occur due to many reasons like sometimes people don't answers all query during surveydue to privacy or sometimes data entry operator leave blank space due to lack of concentration or some other reasons etc. Failure of any system or snsor nodes in wireless sensor network can also lead to missing data. Missing data is a

DOI: 10.4018/978-1-7998-2460-2.ch045

very challenging issue in the field of analytics because the completeness and quality of the data always plays a crucial role in analyzing the available data. Replace the missing value by an appropriate value is called imputation. In general, data mining algorithms are not capable of handling data incompleteness on its own. So, it is necessary to impute those missing value by some appropriate vaue using some suitable data imputation algorithm (Ankaiah & Ravi, 2011; Abdella & Marwala, 2005; García & Kalenatic, 2011; Nishanth, Ravi, Ankaiah & Bose, 2012).

Kline (1988) proposed following procedure to handle missing data:

1. Deletion procedure viz., Listwise deletion and Pairwise deletion (Song & Shepperd, 2007),
2. Imputation procedure (Schafer, 1997),
3. Model based procedure, and
4. Machine learning methods.

The remainder of this chapter is organized as follows: first, a brief review of literature on imputation of missing data is presented. Further, proposed method is explained. Then, description of the dataset and Experimental design is described in next section. Results and discussions are presented in second last section and last section states about conclusion.

BACKGROUND

In case of numerical attributes, missing data can be handled in various ways. Numerous type of imputation is possible like: machine learning (ML) based, deletion of missing values, model based approaches etc. There are various ML based approaches like auto-associative neural network imputation with genetic algorithms (Abdella & Marwala, 2005), SOM (Merlin, Sorjamaa, Maillet & Lendasse, 2010), multi-layer perceptron (Gupta & Lam, 1996), K-Nearest Neighbor (Batista & Monard, 2002), fuzzy-neural network (Gabrys, 2002) etc. Batista and Monard (2002, 2003) and Jerez, Molina, Subirates and Franco (2006) employed K-nearest neighbour (K-NN) for handling missing data. Mutual K-NN method proposed by Liu and Zhang (2012) to classify noisy and incomplete data. For handling missing data, Samad and Harp (1992) employed SOM based approach, Austin and Escobar (2005) employed Monte Carlo simulations. Several studies employed Multi-layer perceptron (MLP) for imputation, we train MLP using data without missing attribute as autoassociative model and furthet pass data with missing attribute to trained model for imputation Sharpe and Solly (1995), Nordbotten (1996), Gupta and Lam (1996), Yoon and Lee (1999), Silva-Ramírez, Pino-Mejías, López-Coello and Cubiles-de-la-Vega (2011) and Nkuna and Odiyo (2011). The authors used MLP for data imputation. Auto-associative neural network (AANN) has also been employed for this task by keeping input and output variable identical (Marseguerra & Zoia, 2002; Marwala & Chakraverty, 2006). Ragel and Cremilleux (1999) employed Robust Association Rules Algorithm (RAR) to address multiple missing values in database. Chen, Huang, F. Tian and S. Tian (2008) proposed selective Bayes classifier to handle missing data. Fuzzy c-means algorithm has been employed by Nouvo (2011) to handle incomplete data. Principles of chaos theory has been employed by Elshorbagy, Simonovic and Panu (2002) to handle missing data in stream flow data. Expectation maximization (EM) algorithm has been employed by Dempster, Laird and Rubin (1977) to handle missing values in multivariate data. García and Kalenatic (2011) also proposed Genetic algorithm (GA) based approach to handle missing attribute in multivariate data. Ankaiah and Ravi (2011) handled missing