Chapter 2
Missing Value Imputation Using ANN Optimized by Genetic Algorithm

Anjana Mishra
Department of IT, C.V.Raman College of Engineering, Mahura, India

Bighnaraj Naik
Department of Computer Application, Veer Surendra Sai University of Technology, Burla, India

Suresh Kumar Srichandan
Department of IT, Veer Surendra Sai University of Technology, Burla, India

ABSTRACT

Missing value arises in almost all serious statistical analyses and creates numerous problems in processing data in databases. In real world applications, information may be missing due to instrumental errors, optional fields and non-response to some questions in surveys, data entry errors, etc. Most of the data mining techniques need analysis of complete data without any missing information and this induces researchers to develop efficient methods to handle them. It is one of the most important areas where research is being carried out for a long time in various domains. The objective of this article is to handle missing data, using an evolutionary (genetic) algorithm including some relatively simple methodologies that can often yield reasonable results. The proposed method uses genetic algorithm and multi-layer perceptron (MLP) for accurately predicting missing data with higher accuracy.

INTRODUCTION

Missing data are universal in research work and the research of missing value imputation has gain momentum since first few decades (Little, 1987; McCleary, 2002; Royston, 2004; Sehgal et al., 2005). Still the missing imputation is a key term among academic and statistic research (Aydilek and Arslan, 2013; D'Souza, 2015; Deb and Liew, 2016; Sahri et al., 2014). By missing data, mean data that are missing

DOI: 10.4018/978-1-7998-8048-6.ch002

for some (but not all) variables and for some cases. If data are missing on a variable for all cases, then that variable is said to be latent or unobserved. The predictive models are used by many decision-making processes that take observed data as inputs. Such prototypes breakdown when one or more inputs are missing. But simply ignoring the incomplete record is not an option in many applications. This is mainly because, the fact that ignorance can lead to biased results in statistical modeling or even damages in machine control (Roth and Switzer, 1995). For this reason, it is often essential to make the decision based on available data. In real world applications, information may be missing due to instrumental errors, power system failure, omitted entries in databases, noise environment factor (humidity, temp), human error in measurements problems of data transfer in digital system, low quality of sensor, nonresponse to some questions in surveys, data entry errors, etc. (Amiri & Jensen, 2016). There are three missing data mechanisms: not missing at random (NMAR), missing at random (MAR), and missing completely at random (MCAR).

LITERATURE REVIEW

In this section, we will provide a brief literature survey of missing values imputation methods with merits and demerits.

Amiri and Jensen (2016) have proposed three missing value imputation methods based on fuzzy-rough sets; namely, implicator/t-norm based fuzzy-rough sets, vaguely quantified rough sets and also ordered weighted average based rough sets which combined with the nearest neighbor algorithm to get benefit from both the simplicity and accuracy of nearest neighbor prediction with the robustnesss and noise tolerance of fuzzy-rough sets. The three algorithms are Fuzzy-Rough Nearest Neighbor Imputation algorithm (FRNNI), Ordered Weighted Average-based nearest neighbor Imputation algorithms (OWANNI) and Vaguely Quantified Nearest Neighbor Imputation (VQNNI). All algorithms compared with each other and found that FRNNI performs better than the other two methods and 11 other existing methods – Bayesian PCA(BPCAI), Concept Most Common (CMCI), Fuzzy K Means(FKMI),K Means(KMI),KNN impute(KNNI), LLS Impute(LLSI), Most Common(MCI),SVD impute(SVDI), SVM impute(SVMI), WKNN impute(WKNNI)and finally Expectation Maximization(EMI) on 27 benchmark datasets.

Deb and Liew (2016) have proposed an algorithm which is used to find the missing value in the traffic accident databases of numerical or categorical values. For estimating, this algorithm has considered four publicly available traffic accident databases from the United States. The first data set is (explore.data.gov) an Largest open federal database, the second is (data.opencolorado.org) National Incident Based Reporting System (NIBRS) of the city and county of Denver, third is (MotorVehicleCrashes-caseinformation:2011 and fourth is MotorVehicleCrashes-individualinformation:2011, data.ny.gov from New York open data portal. The proposed algorithm used the decision tree to find the set of interrelated records and this sampling based missing value imputation algorithm is named as DSMI. The large data set horizontally divides based on non –missing attributes of the record, followed by the missing values are imputed by the link between the missing and non-missing attributes using the IS measure and direct and transitive relationship of attribute value across two records using weighed similarity measures. The proposed algorithm has better accuracy than the existing algorithm where a large number of attributes are categorical in the datasets.

Andrea D'Souza (2015) have compared the three data mining techniques such as Artificial Neural Network, K Means Clustering and Frequent ItemSet Generation Using Apriori Algorithm to solve predic-