International Journal of Trend in Scientific Research and Development (IJTSRD)

International Open Access Journal

ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

Data Imputation by Soft Computing

Ritesh Kumar Pandey[1], Dr Asha Ambhaikar[2]
[1]M.Tech Scholar, [2]Professor
Department of CSE, Kalinga University, Naya Raipur, Chhattisgarh, India

ABSTRACT

Data imputing uses to posit missing data values, as missing data have a negative effect on the computation validity of models. This study develops a genetic algorithm (GA) to optimize imputing for missing cost data of fans used in road tunnels by the Swedish Transport Administration (Trafikverket). GA uses to impute the missing cost data using an optimized valid data period. The results show highly correlated data (R- squared 0.99) after imputing the missing data. Therefore, GA provides a wide search space to optimize imputing and create complete data. The complete data can be used for forecasting and life cycle cost analysis.

Keywords: *data imputing, genetic algorithms (GA), R-Squared*

1 INTRODUCTION

Data imputing uses to posit the existence of missing values to decrease the computational process, estimate model variables and derive the results that would have been seen if the complete data were used. The common practice is to impute the missing data using the average of the observed values. With imputing, no values are sacrificed, thus precluding the loss of analytic results [1].

Genetic algorithm (GA) is a widely used evaluation technique to optimize and predict missing data by finding an approximate solution interval that minimizes the error prediction function [2]. Several studies of imputing data have used GAs to understand and improve data to avoid bias in decision-making.

Ibrahim Berkan Aydilek et al. [3] proposed a hybrid approach that utilizes fuzz c-means clustering with combination between support vector regression and a genetic algorithm. This approach used to optimize cluster size and weight factor and estimating missing values. The proposed lustering technique used to estimate the missing values based on the similarity and Root Mean Standard Errors (RMSE) used to estimate the imputing accuracy. The authors found that clustering makes missing value a member of more than one cluster centroids, which yields more sensible imputation results.

Mussa Abdella et al. [4] introduced a new method by combing genetic algorithm (GA) and neural networks to approximate the missing data in database. The authors use GA to minimize an error function derived from an auto-association neural network. They used a standard method (Se) to estimate the imputing accuracy of the missing data that investigated using the proposed method. The authors found that the model approximates the missing values with higher accuracy.

Missing data creates various problems in many research fields like data mining, mathematics, statistics and various other fields [1]. The process of replacing or estimating missing data is called data imputation. Data imputation is very useful for data mining applications for getting completeness in the data. For analyzing the data through any technique completeness and quality of data are very important things. For example researchers rarely find the survey data set that contains complete entries [3]. The respondents may not give complete information because of negligence, privacy reasons or ambiguity of the survey questions. But the missing parts of variables may be important things for analyzing the data. So in this situation data imputation plays a major role. Data imputation is also very useful in the control based applications like traffic monitoring, industrial process, telecommunications and computer networks, automatic speech recognition, financial and business applications, and medical diagnosis etc.

To impute with incomplete or missing data, several techniques are reported based on statistical analysis [4]. These methods include like mean substitution methods, hot deck imputation, regression methods, expectation maximization, multiple imputation methods. Some other techniques proposed based on machine learning methods include SOM, K-Nearest Neighbor, multi layer perceptron, recurrent neural network, auto-associative neural network imputation with genetic algorithms, and multi-task learning approaches.

2. Methods

2.1. Data collection

The cost data are for Swedish tunnel fans in Stockholm. The data were collected over ten years from the Swedish Transport Administration (Trafikverket) and stored in the MAXIMO computerized maintenance management system (CMMS). In CMMS, the cost data are recorded based on work orders of tunnel fans and contain labour cost. It is important to mention that labour cost data used in this study are real costs without inflation. Due to company regulations, labour cost data are encoded and expressed as currency units (cu) for this study.

2.2. Genetic algorithm (GA)

GA is widely applied in imputing because of its ability to optimize valid imputing period in a large space of random populations [6]. The GA operates with a population of chromosomes containing data of work orders. The chromosomes are proportional to the case and problem statement [7] as seen in figure 1. GA is applied longitudinally to the data. GA operates with a population of chromosomes that contains labour cost. Forty percent of each cost object is selected randomly at two different times.

3.Proposed Soft Computing Architecture

The proposed missing data imputation approach is a 2 stage approach. The block diagram (Fig 1) depicts the schema of the proposed imputation method. In this novel hybrid we using K-means [19] clustering for stage 1. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure for stage 1 imputation as follows:

1. Identify K cluster centers by using K-means clustering algorithm with complete records.
2. Fill the incomplete records with the corresponding features of the nearest cluster center by measuring the Euclidean distance of complete components of an incomplete record and cluster centers.

In the second stage, we used multilayer perceptron (MLP) for imputation. MLP is trained by using only complete cases. We have to train as a regression model by taking one incomplete variable as target and remaining variables as inputs. So that we have to form different regression models that are equal to the number of incomplete variables in a given dataset. The steps for MLP imputation (Stage 2) scheme as follows:

1. For a given incomplete dataset , separate the records that contain missing values from the set of those without missing values (or with complete values). Let us take the set of complete records as known values and incomplete records as unknown records
2. For each incomplete variable, construct an MLP by considering the remaining variables in as inputs for training.
3. Predict the missing values in the variable, which is the target variable in MLP. While predicting we use the initial approximate which are given by K-means clustering from stage 1 as part of

4 Experimental Design

The effectiveness of the proposed method is tested on 2 classification and 2 regression datasets. Since none of these datasets has missing values, we conducted the experiments by deleting some values from the original datasets randomly. Every dataset is divided into 10 folds and 9 folds are used for training and the tenth one is left out for testing. From th test fold, every time, we deleted nearly 10% of the values (cells) randomly. We ensured that at least one cell from every record is deleted. In the stage 1 of data imputation, K-means clustering is performed by using only complete set of records (training data comprising 9 folds). The value of K in K-means is set equal to the number of classes in case of classification datasets. In the case of Wine data the number of classes is 3, so we have chosen K-value as 3. Similarly, in the case of UK banks dataset the number of clusters are chosen as

2. However, in the case of regression datasets, the number of clusters, K, is chosen by visualizing the data using principle component analysis (PCA). By visualizing the plot of PC1 vs PC2, we can set the approximate number of clusters. Thus, the number of clusters is taken as 2 for Boston housing dataset and 3 for forest fires dataset. We can see the plots of PCA visualization for Boston housing and forest fires dataset in Figures 3 and 4 respectively.

5. Results and Discussion

The amount of missing data in the labour cost 56.84% as seen in figure 2. Missing data cause a substantial amount of bias, make the analysis of the data more arduous, and reduce analysis efficiency. GA is implemented to impute the missing data. The imputation will help to provide complete data that can be used for forecasting or life cycle cost analysis

6. Datasets Description

In this paper we analyzed 4 datasets. Those include two regression datasets viz., Forest fires, Boston housing and two classification datasets viz., Wine and UK banks. The benchmark datasets, Wine, Boston housing, and Forest fires are taken from UCI machine learning repository. Forest fires dataset contains 11 predictor variables and 517 records, whereas Boston housing dataset contains 13 predictor variables. Another two datasets we used are Wine and UK bank bankruptcy datasets. Both these datasets are classification datasets. Wine dataset contains 13 predictor variables and 248 records. UK banks dataset contains 10 predictor variables and 60 records. The predictor variables of UK banks dataset are *(i)* Sales *(ii)* Profit Before Tax / Capital

7. Conclusion

The techniques proposed for missing data imputation in the literature used either local learning or global approximation only. In this paper, we replaced the missing values by using both local learning and global approximation. The proposed hybrid is tested on four datasets in the framework of 10 fold cross validation. In all the data sets some values are randomly removed and we treated those values as missing values. In stage 1, by using K-means clustering we replaced missing values by local approximate values. In stage 2 by using the local approximate values which are resulting from stage 1 and trained MLP from complete records, we further approximate the missing value to the actual value. The missing values are replaced by using proposed novel hybrid approach, and then we compared predicted values with actual values by using MAPE. We observed that MAPE value decreased from stage 1 to stage 2. t- test is performed on four datasets, and from the values of t-test we can say that the reduction in MAPE from stage 1 to stage -2 is statistically significant. We conclude that, we can use the proposed approach as a viable alternative to the extant methods for data imputation. In particular, this method is useful for a dataset with a records having more than one missing values.

REFERENCES

1) M. Abdella and T. Marwala, "The use of genetic algorithms and neural networks to approximate missing data in database,"*Computational Cybernetics, ICCC 2005. IEEE 3rd International Conference*, pp. 207-212, 2005.

2) R.J.A. Little and D.B. Rubin, "Statistical analysis with missing data", *Wiley*, 2nd ed., New Jersey, 2002.

3) W. Hai , W. Shouhong, "The Use of Ontology for Data Mining with Incomplete Data", *Principle Advancements in Database Management Technologies*, *pp.* 375-388, 2010.

4) Abdella M, Marwala T (2005) The use of genetic algorithms and neural networks to approximate missing data in database. In: Anonymous Computational Cybernetics, 2005. ICCC 2005. IEEE 3rd International Conference on. IEEE, p 207

5) Ni D, Leonard JD, Guin A et al (2005) Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. J Transp Eng 131(12):931-938

6) Deb K, Pratap A, Agarwal S et al (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation 6(2):182-197

7) Cordón O, Herrera F, Gomide F et al (2001) Ten years of genetic fuzzy systems: current framework and new trends. In: Anonymous IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th, 3 vol. IEEE, p 1241

8) J.L. Schafer, "Analysis of incomplete multivariate data", *Chapman & Hall*, Florida, 1997.