J Inf Process Syst, Vol.9, No.4, pp.633-650, December 2013
http://dx.doi.org/10.3745/JIPS.2013.9.4.633
pISSN 1976-913X
eISSN 2092-805X

A Computational Intelligence Based Online Data Imputation Method: An Application For Banking

Kancherla Jonah Nishanth*, and Vadlamani Ravi*

Abstract—All the imputation techniques proposed so far in literature for data imputation are offline techniques as they require a number of iterations to learn the characteristics of data during training and they also consume a lot of computational time. Hence, these techniques are not suitable for applications that require the imputation to be performed on demand and near real-time. The paper proposes a computational intelligence based architecture for online data imputation and extended versions of an existing offline data imputation method as well. The proposed online imputation technique has 2 stages. In stage 1, Evolving Clustering Method (ECM) is used to replace the missing values with cluster centers, as part of the local learning strategy. Stage 2 refines the resultant approximate values using a General Regression Neural Network (GRNN) as part of the global approximation strategy. We also propose extended versions of an existing offline imputation technique. The offline imputation techniques employ K-Means or K-Medoids and Multi Layer Perceptron (MLP) or GRNN in Stage-1 and Stage-2 respectively. Several experiments were conducted on 8 benchmark datasets and 4 bank related datasets to assess the effectiveness of the proposed online and offline imputation techniques. In terms of Mean Absolute Percentage Error (MAPE), the results indicate that the difference between the proposed best offline imputation method viz., K-Medoids+GRNN and the proposed online imputation method viz., ECM+GRNN is statistically insignificant at a 1% level of significance. Consequently, the proposed online technique, being less expensive and faster, can be employed for imputation instead of the existing and proposed offline imputation techniques. This is the significant outcome of the study. Furthermore, GRNN in stage-2 uniformly reduced MAPE values in both offline and online imputation methods on all datasets.

Keywords—Data Imputation, General Regression Neural Network (GRNN), Evolving Clustering Method (ECM), Imputation, K-Medoids clustering, K-Means clustering, MLP

1. INTRODUCTION

Missing data is observed in almost all of the real world datasets. Missing values result in less efficient estimates because of sample bias and reduced sample sizes. Most data mining algorithms cannot work with incomplete datasets. For analyzing the available data, its completeness and quality play a major role, because the inferences made from complete data are more accurate and reliable than those made from incomplete data [1]. Hence, missing values are to be imputed

Manuscript received November 13, 2012; first revision March 6 2013; second revision May 13 2013; accepted August 25, 2013.
Corresponding Author: Vadlamani Ravi (rav_padma@yahoo.com)
* Center of Excellence in CRM and Analytics, Institute for Development and Research in Banking Technology (IDRBT), Hyderabad, India (jonah.nishanth@gmail.com, rav_padma@yahoo.com)

before performing any further analysis on the data. In statistics, imputation is the substitution of some value for a missing data point or a missing component of a data point. Once all of the missing values have been imputed, the dataset can then be analyzed using standard techniques for complete data. Data imputation seeks to improve the quality of the data and to make the data more reliable for mining purposes.

Datasets with incomplete observations are broadly experienced in almost all of the areas of research. Many reasons may lead to why data is missing. In surveys, data may be missing due to procedural factors such as errors in data entry, disclosure restrictions, failure to complete the entire questionnaire and when the response does not apply for an induvidual (e.g., questions regarding the years of marriage for a respondent who has never been married). In the geosciences, data items in the observational data sets may be missing altogether, or they may be imprecise in one way or another [15].Practical observations are often incomplete because of equipment malfunctioning, outliers, or incorrect data entry. In environmental research, data may be missing due to faults in data acquisition. Speech samples that are corrupted by very high levels of noise are considered to be missing data in automatic speech recognition [7]. Incomplete data may also appear in business and financial applications. In biological research with DNA microarrays, gene data may be missing due to reasons such as a there being a scratch on the slide containing the gene sample and contaminated samples [38]. Missing data can also occur as a result of dropouts. For example, when an experiment is run on a group of individuals over a period of time as in clinical studies. According to Roth et al. [30], missing data has two major negative effects. First, it has a negative impact on statistical power. Second, missing data may result in biased estimates in several ways. Missing data biases the measures of central tendency upward or downward depending upon where in the distribution they appear. The measures of dispersion may also be affected depending upon which part of distribution has missing data. Missing data may result in biased correlation coefficients [23].

According to Little and Rubin [20], missing data is categorized into 3 categories: (i) Missing Completely At Random (MCAR), (ii) Missing At Random (MAR) and (iii) Not Missing At Random (NMAR). MCAR occurs if the probability of missing value on some variable X is independent of the variable itself and on the values of any other variables in the dataset. For example, if the gender of a customer is missed in the customer's database then it does not depend on the any other variable in the database. Possible reasons for MCAR include manual data entry procedure, incorrect measurements, equipment error, changes in experimental design etc. MAR occurs when the probability of missing data on a particular variable (i.e., income level) depends on other variables (i.e., profession) in the database but not the variable itself. NMAR occurs when the probability of missing data on a particular variable depends on the variable itself. For instance, if citizens did not participate in a survey, then NMAR occurs. MCAR and MAR data are recoverable, whereas NMAR is irrecoverable.

Data imputation techniques are categorized into deletion, imputation, model-based and machine learning or computational intelligent or soft computing procedures. The machine learning based methods include SOM [26], K-Nearest Neighbor [4], MLP [13], Fuzzy-neural network [11], Auto-Associative Neural Network (AANN) imputation with genetic algorithms [1] etc. All the methods mentioned above require a lot of iterations to learn the characteristics of the data. As such, these methods are termed as being offline techniques for data imputation.

In this paper, we propose a novel, online, two stage imputation technique. Recently, Ankaiah and Ravi [2] employed a soft computing hybrid for data imputation. In this paper, we extend

their work and we also propose other offline hybrid imputation methods. The work presented here is different from that of Ankaiah and Ravi [2] in that we employed: (i) the Evolving Clustering Method (ECM) which is a fast, one-pass algorithm for a dynamic estimation of the number of clusters in a dataset for clustering in Stage-1; (ii) the General Regression Neural Network (GRNN) instead of Multi Layer Perceptron (MLP) that they used ; and (iii) offline methods, such as K-Medoids for clustering instead of K-Means which they used. We used GRNN because unlike MLP, GRNN is a one pass algorithm and it can outperform other methods even with sparse number of points. The most interesting aspect of the present work is that ECM and GRNN are one pass learning algorithms. Hence, they are employed for performing online data imputation. K-Means is replaced by K-Medoids because the former has the following disadvantage: it is sensitive to noise and outlier data points, as a small number of this type of data can substantially influence the mean value. The performance of the proposed imputation techniques is compared with a least squares approximation method viz., Iterative Majorization Least Squares (IMLS) algorithm and a nearest neighbor based hybrid data imputation algorithm viz., IMLS-NN-IMLS (INI) [39] [40].

The remainder of this paper is organized as follows: a brief review of the literature on imputation of missing data is presented in Section 2. A brief overview of the techniques used in this paper for online and offline imputation is presented in Section 3. The proposed online and offline methods are presented in Section 4. Experimental setup is described in Section 5. Results and discussions are presented in Section 6, followed by conclusions in Section 7.

2. OVERVIEW OF TECHNIQUES USED

2.1 Online Evolving Clustering Method (ECM)

The online ECM proposed by Kasabov and Song [17] is a fast, one-pass algorithm for a dynamic estimation of the number of clusters in a dataset and for finding their current centers in an input data space. It is a distance based connectionist clustering method. ECM is based on the concept of dynamically adding and modifying the clusters as new data is presented to the ECM algorithm, where the modification to the clusters affects both the position of the clusters and the size of the cluster in terms of a radius parameter associated with each cluster that determines the boundaries of that cluster. ECM has only one parameter, which drives the addition of clusters, known as the distance threshold *Dthr*. The ECM algorithm is described as follows:

1. Create the first cluster center C_1 by simply taking the first example X_1.
2. For all the subsequent vectors X_k do the following
   2.1 Calculate the distance D_{kj} between X_k and existing cluster centers C_j where j= 1 to n
   2.2 Find D_{min} = min D_{kj}, j = 1 to n
   2.3 If D_{min} is less than the radius of any of the already created clusters R_j, j = 1 to n then
      2.3.1 Add to X_k the nearest cluster
      Else
      2.3.1 Find a cluster center C_t with a minimum S_{kj} where $S_{kj} = D_{kj+}R_j$, j = 1 to n
      2.3.2 If S_{kt} 〉 2*Dthr then
         2.3.2.1 Create a new cluster with cluster center as X_k

Else Update the cluster C_t as follows

2.3.2.1 $R_t^{new} = S_{kt}/2$

2.3.2.2 the new center $\mathbf{C}_t^{new}$ is located at the point on the line that connects X_k the and C_t, and the distance from the new center $\mathbf{C}_t^{new}$ to the point is equal to R_t^{new} .

2.2 K-Means clustering

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. The K-Means clustering [22] method takes the input parameter k, and partitions a set of n objects into k clusters such that the intra-cluster similarity is high but the inter-cluster similarity is low. The K-Means algorithm works as follows: first, it randomly selects k objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster [14]. The process iterates until the criterion function converges. Generally the square-error criterion is used as a convergence function, which is defined as

$$E=\sum_{i=1}^{k} \sum_{p\in C_i} |p-m_i|^2 \quad (1)$$

$\boldsymbol{p}$ is the point in space representing a given object in cluster C_j and m_i is the representative object of cluster C_i.

2.3 K-Medoids clustering

The K-Medoids clustering method takes the actual objects to represent the clusters instead of taking the mean value of the objects in a cluster as the reference point. Each remaining object is clustered with the representative object to which it is most similar. The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities between each object and its corresponding reference point. That is, an error criterion is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p\in C_i} |\mathbf{p}-\mathbf{o_j}| \quad (2)$$

Where E is the sum of the absolute error for all objects in the data set; $\boldsymbol{p}$ is the point in space that represents a given object in a cluster C_j and o_j is the representative object of C_j. The K-Medoids method iterates until, each representative object is actually the medoid [14].

2.4 Multi Layer Perceptron (MLP)

Since MLP is too popular to be described here, we go ahead with explaining the GRNN. The reader can refer to Rumlhart et al. [31] for a detailed description of MLP.

2.5 General Regression Neural Network (GRNN)

The General Regression Neural Network (GRNN) was originally proposed and developed by

Specht [37]. This class of network paradigm has the distinctive features of learning swiftly, working with a simple and straight forward training algorithm, and being discriminative against infrequent outliers and erroneous observations. As its name implies, GRNN is capable of approximating any arbitrary function from historical data. In GRNN, each training sample is operated as a kernel during the training process. The regression surface is established by using the Parzen window estimator. The estimation of GRNN is based on non-parametric regression analysis to create the best fit for the observed data. As such, GRNN does not require prior knowledge of the regression function.

The regression of a dependent variable, Y, on an independent variable, X, is the computation of the most probable value of Y for each value of X based on a finite number of possibly noisy measurements of X and the associated values of Y. The variables X and Y are usually vectors. In order to implement system identification, it is usually necessary to assume some functional form. In the case of linear regression, for example, the output Y is assumed to be a linear function of the input, and the unknown parameters, a_i, are linear coefficients. The method does not need to assume a specific functional form. A Euclidean distance (D_i^2) is estimated between an input vector and the weights, which are then rescaled by the spreading factor. The radial basis output is then the exponential of the negatively weighted distance. The GRNN equation can be written as:

$$D_i^2 = \left(X - X^i\right)\left(X - X^i\right)^T \tag{3}$$

$$Y(X) = \frac{\sum_{i=1}^{n} Y_i \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{D_i^2}{2\sigma^2}\right)} \tag{4}$$

Where, σ is the Smoothing Factor (SF).

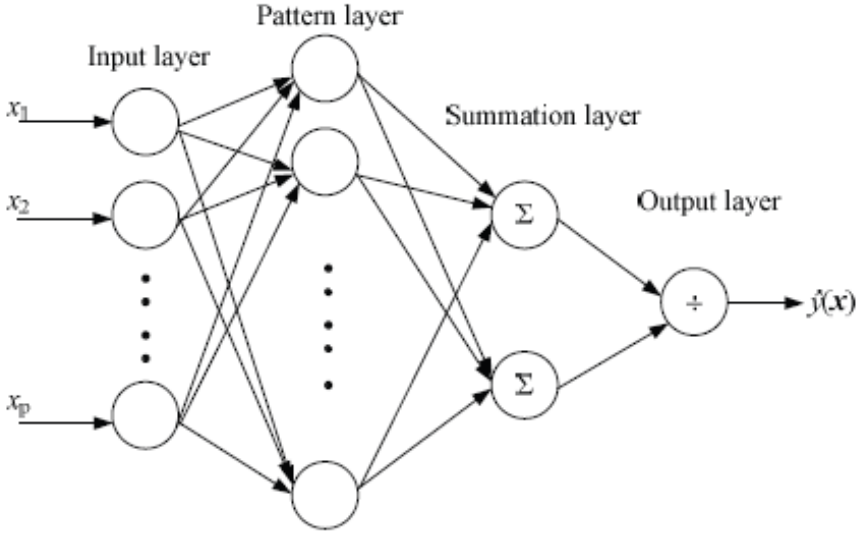

Fig. 1. Schematic diagram of GRNN architecture

The estimate Y(X) can be visualized as a weighted average of all of the observed values, Y^i, where each observed value is weighted exponentially according to its Euclidian distance from X. Y(X) is simply the sum of Gaussian distributions that are centered at each training sample. The topology of GRNN developed by Specht [37] is shown in Figure 1. It consists of the following four layers: the input layer, the pattern layer, the summation layer and the output layer. The input layer contains input units that are merely distributed units, which provide all of measurement variables to all the neurons in the second layer, which is the pattern layer. The pattern unit is dedicated to one cluster center. When a new vector is entered into the network, it is subtracted from the stored vector that represents each cluster center. Either the squares or the absolute values of the differences are summed and are fed into a nonlinear activation function. The activation function normally used is the exponential. The pattern unit outputs are passed onto summation units. The summation units perform a dot product between a weight vector and a vector composed of the signals from the pattern units. It includes two units. One is the denominator summation unit, and the other is the numerator summation unit. The first unit adds up the weight values coming from each of the hidden neurons. The numerator summation unit adds up the weight values that are multiplied by the actual target value for each hidden neuron. The output layer generates the desired estimation of output, denoted by Y(X). It divides the value accumulated in the numerator summation unit by the value in the denominator summation unit, and uses the result as the estimated value.

3. Review of Data Imputation Techniques

Missing data handling methods can be broadly classified into four categories [18]: (a) deletion, (b) imputation (c) modeling the distribution of missing data and then estimating them based on certain parameters and (d) machine learning methods. Each of these techniques is discussed below.

3.1 Deletion procedures

The deletion techniques simply delete the cases that contain missing data. Deletion procedures are generally easy to carry out and may be a good choice for datasets with small amounts of missing data. This approach has two forms: (i) Listwise deletion which omits the cases or instances containing missing values. This method may lead to serious biases when there are a large number of missing values and if the original dataset is too small. (ii) Pairwise deletion which considers each feature separately. For each feature, all recorded values in each observation are considered and missing data are ignored. It is good when the overall sample size is small or when the number of missing data observations are large [36].

3.2 Imputation procedures

In imputation based procedures the missing values are filled-in and the resulting complete data is used for further analysis. The advantages of these procedures are the retention of the sample size and statistical power in subsequent analysis. The simplest and earliest method of imputation is the mean imputation. The mean imputation replaces the missing value of a variable with the average of all the remaining records of that variable [20]. The disadvantage is that it leads to an

underestimation of the population variance and it ignores the correlations between variables. When the variables are correlated, data imputation can be done with regression imputation. In regression imputation the missing variables in a record are replaced by the predicted values on the regression for the known variables for that record. The disadvantage of regression imputation is that it assumes a linear relationship between the predictors and the missing variable. Hot and cold deck imputation replaces the missing variable or attribute in an incomplete observation with the corresponding variable or attribute of the closest complete observation [33]. The drawback of hot deck imputation is that the estimation of missing data is based on single complete vector and thus it ignores the global properties of the dataset. The drawback of cold deck imputation is that missing values are replaced with the different dataset values [20].In the multiple imputation procedure, the missing data is filled in M times to yield M complete datasets. The M complete datasets are analyzed and the results from the datasets are combined for inference [12].

3.3 Model-based procedures

Maximum likelihood is one of the model-based procedures. The maximum likelihood approach to analyzing missing data assumes that the observed data is a sample drawn from a multivariate normal distribution [8]. The parameters are estimated by available data and then missing values are determined based on the estimated parameters. The expectation maximization algorithm is an iterative process [19]. The first iteration estimates missing data and then it estimates the parameters using maximum likelihood. The second iteration re-estimates the missing data based on new parameters then recalculates the new parameter estimates based on actual and re-estimated missing data [20].

3.4 Machine learning methods

In K-Nearest Neighbor (K-NN) approach the missing values are replaced by their nearest neighbors. The nearest neighbors are selected from the complete cases which minimize the distance function. For numerical variables the mean of K neighbours is used to replace the missing value, whereas for categorical variables the mode of K neighbors is used to replace the missing value. Jerez et al. [16] used K-NN for breast cancer prognosis. Batista and Monard [4][5] also used K-NN for missing data imputation. Liu and Zhang[21] developed a mutual K-NN algorithm for classifying incomplete and noisy data. Samad and Harp [32] implemented the SOM approach for handling the missing data. Austin and Escobar [3] used Monte Carlo simulations to examine the performance of three Bayesian methods that imputed missing data by placing a simple prior distribution upon the variable that was subject to being missing. In the neural network approach, MLP should be trained as a regression model by using the complete cases and by choosing one variable as a target each time. By using the appropriate MLP model, each incomplete pattern value is predicted. Several researchers Sharpe and Solly [34], Nordbotten [28], Gupta and Lam [13], Yoon and Lee [41], Ramirez et al. [35] and Nkuna and Odiyo [27] used MLP for missing data imputation. In Auto-Associative Neural Network (AANN) imputation, the network is trained to predict the inputs by taking same input variable as a target [24] [25]. Ragel and Cremilleux [29] proposed a missing value completion method. This method extends the concept of the Robust Association Rules Algorithm (RAR) for databases with multiple missing values. Chen et.al, [6] employed a selective Bayes classifier to classify incomplete data with a simpler formula for computing the gain ratio. Nouvo [9] employed fuzzy a c-means for data

Table 1. Techniques for Data Imputation

TECHNIQUE	DESCRIPTION
DELETION PROCEDURES	
Listwise Deletion [20]	Eliminates all the instances with missing values.
Pairwise Deletion [20]	Analysis with all cases in which the variables of interest are present.
IMPUTATION PROCEDURES	
Hot-deck Imputation [33]	Replaces the missing data with values from a similar complete data vector.
Mean Imputation [20]	The missing value is replaced by the mean.
Multiple Imputation [20]	Replaces each missing value with a set of plausible ones that represent uncertainty about the right value to impute.
Regression Imputation [20]	Estimates the relationships among the variables and then uses coefficients to estimate the missing values.
MODEL BASED PROCEDURES	
Expectation Maximization [20]	An iterative procedure that continues until there is a convergence in parameter estimates.
MACHINE LEARNING METHODS	
Genetic algorithms and neural networks [24][25]	A genetic algorithm is used to minimize an error function derived from an AANN.
Imputation with K-Nearest Neighbors [4]	K-Nearest Neighbors are selected from completed cases. The replacement value depends on type of data: the mode can be used for discrete data and mean for continuous data.
Imputation using a soft computing hybrid [2]	A two stage imputation technique where K-Means and MLP are used for imputation in Stage-1 and Stage-2 respectively.
Missing Value Completion(MVC) method [29]	This method extends the concept of the Robust Association Rules Algorithm (RAR) for databases with multiple missing values.
MLP Imputation [13][28][34]	MLP is trained using only the complete cases as a regression model by taking an incomplete variable as a target and the remaining variables as input.
SOM Imputation [32]	The value to be imputed is computed based on the activation group of nodes in the missing dimensions.

imputation. Elshorbagy et al. [10] employed the principles of the chaos theory to estimate the missing stream flow data. Various imputation techniques appeared in literature for data imputation is presented in Table 1.

4. Proposed Online and Offline Imputation Techniques

4.1 Architecture of the proposed online imputation technique

The architecture of the proposed online imputation technique is shown in Figure 2. The proposed online imputation technique is a 2-stage imputation technique. The problem with K-Means and K-Medoids is that the number of clusters to be formed must be specified beforehand and a number of iterations are required for convergence. For an online imputation technique a fast one pass algorithm must be used in Stage-1 and Stage-2. So ECM and GRNN, which are fast one pass learning algorithms, are employed in Stage-1 and Stage-2 respectively. A complete rec-

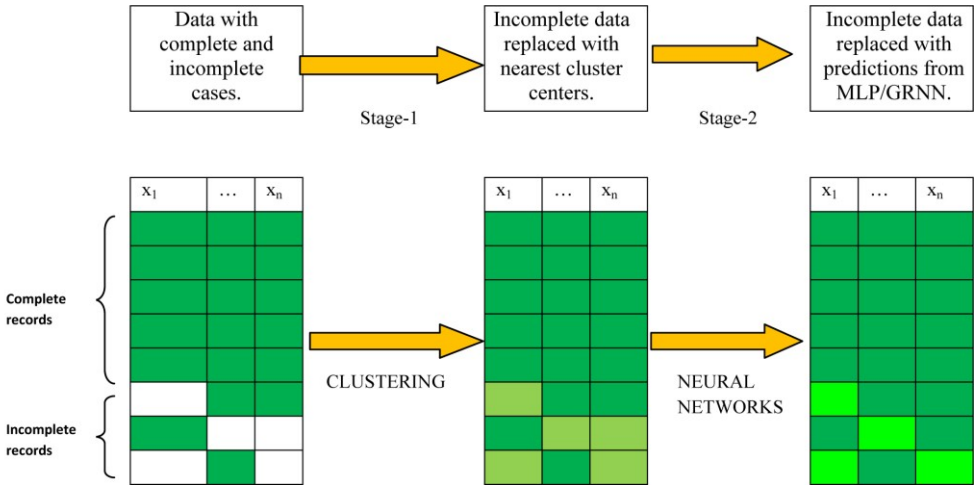

Fig. 2. Architecture of the proposed 2-stage imputation

ord is a record for which all the attribute values are observed. A record that contains missing values in one or more attributes is an incomplete record. Let X_D and p denote the given dataset and the number of attributes respectively. Let X_{CR} denote the complete records and X_{IN} denote incomplete records. Let Ω denote the set of attributes or variables containing missing values. The proposed online imputation technique is described as follows

1. Separate the complete records, X_{CR} and the incomplete records, X_{IN} in X_D.
2. Cluster the complete records X_{CR} with online ECM.
3. For each incomplete record $X_t \in X_{IN}$

   For each cluster center C_j

   - Measure the distance d_j, between the complete components of an incomplete record and cluster centers obtained from Step-2.
   - $d_j = \|X_t - C_j\|$

     Here d_j is the distance between complete components t^{th} (the observed values of an incomplete record) incomplete record and j^{th} cluster center (attribute values corresponding to observed values of an incomplete record). X_t is the t^{th} incomplete record. C_j is the j^{th} cluster center. $\|\ \|$ indicates the Euclidean distance.
   - Find the smallest d_j, i.e the cluster center C_j closest to the incomplete record X_t.
   - Replace the missing values in X_t with nearest cluster center.
4. For each variable k

   If $k \in \Omega$ (if variable k contains missing values), then

   I. Select the records containing missing values in the variable k (i.e., X_k from the set of incomplete records X_{IN}.)

   II. If the records in X_k contain missing values in variables other than k, use the estimations from step 3 (imputed values from Stage-1) to fill those missing values.

   III. Train GRNN with the complete records X_{CR}, by considering the variable 'k' as predictor (output) and all other variables as input.

   IV. Employ the GRNN trained in step III to obtain the predictions for X_k, which in other words, are the refinements.

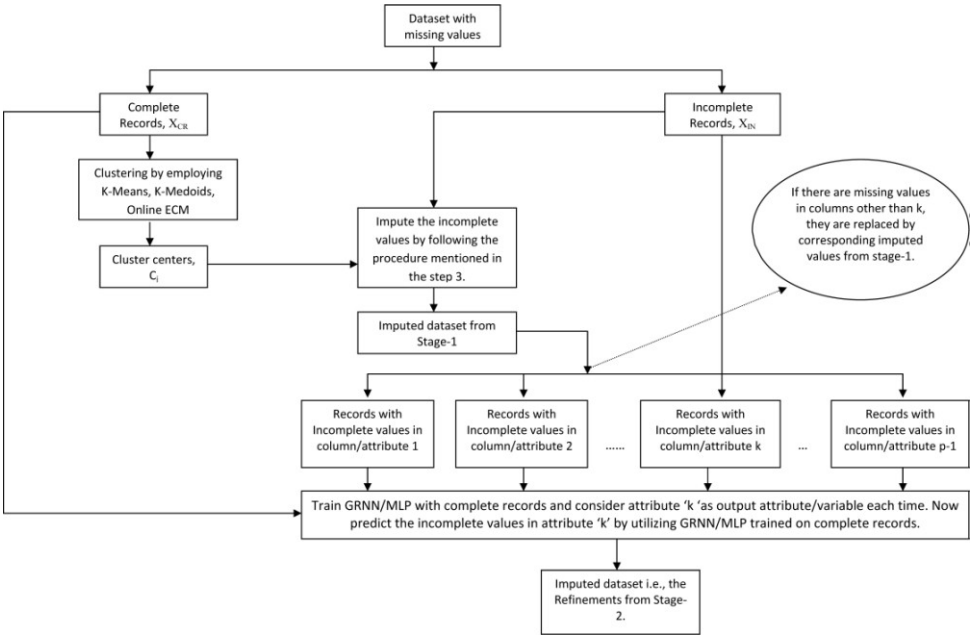

Fig. 3. Detailed architecture of proposed imputation techniques.

4.2 Architecture of the proposed offline imputation technique

The architecture of the proposed offline imputation technique is shown in Figure 2. The architecture is similar to that of the online imputation. K-Means/K-Medoids is employed for imputation in Stage-1. MLP/GRNN is applied for imputation in Stage-2. The detailed architecture of the proposed online and offline imputation techniques is described in Figure 3.

5. EXPERIMENTAL DESIGN

The effectiveness of the proposed method for data imputation has been tested on 8 benchmark and 4 banking datasets. All of the datasets are taken from UCI Machine learning repository. None of the datasets that have been considered for the experimentation have missing values. Hence, we conducted the experiments by randomly deleting some values from the original datasets. First, every dataset was divided into 10 folds and 9 folds were used for training and the tenth one was left out for testing. From the i^{th} test fold, 10% of the values (cells) are deleted randomly. We ensured that at least one cell from every record was deleted. The 12 resulting dataset combinations of training and test sets were analyzed by 5 offline imputation techniques viz., ECM+GRNN, K-Means+MLP, K-Means+GRNN, K-Medoids+MLP and K-Medoids+GRNN. Thus, 600 models in all were constructed from the training sets and the accuracy of each one was measured by the Mean Absolute Percentage Error (MAPE) on the test set, where imputation took place in 2 stages. In the Stage-1 of data imputation K-Means/K-Medoids clustering was performed by using only the complete set of records (training data comprising of 9 folds). The number of clusters (K) in K-Means and K-Medoids were chosen by a systematic procedure. The number of clusters obtained by each of the methods for all datasets is presented in Table 2.

Table 2. - Number of clusters (K) formed for different datasets

NUMBER OF CLUSTERS		
DATASET	K-MEANS	K-MEDOIDS
Boston Housing	3	3
Forest Fires	2	3
Auto MPG	2	4
Body Fat	2	3
Wine	3	3
Prima Indain	2	3
Iris	3	3
SpectF	2	4
UK Credit	2	2
Spanish	2	4
Turkish	2	4
UK Bankruptcy	3	5

In Stage-1, the missing values of incomplete records were replaced by the closest cluster centers. Thus, in Stage-1, missing values were replaced by approximate values obtained through local learning via clustering. In Stage-2, we use MLP/GRNN for approximating the values closest to actual values by using the initial estimates from Stage-1. MLP/GRNN is trained as regression models by considering the attribute that has missing values as a predictor. For training MLP/GRNN we use data of complete records (training data comprising 9 folds). The missing values are predicted by utilizing MLP/GRNN that was trained on complete data. While predicting the missing values, we used the initial approximations yielded by the K-Means/K-Medoids clustering from Stage-1 as part of test set for predicting the target variable if we had more than one missing value in a record. Thus, the proposed imputation techniques involve local learning followed by global approximation. In the online imputation, ECM which is a one-pass algorithm for estimating the number of clusters in a dataset was employed for clustering in Stage-1. To design a 2-stage online imputation method, GRNN was employed in Stage-2 which also employs a one-pass learning algorithm. The effectiveness of the online imputation method (Online ECM+GRNN) was tested on 12 datasets. The experiments were carried out using 10 fold cross validation for all datasets.

6. Results and Discussion

We developed the code for ECM, K-Means, K-Medoids, MLP and GRNN in Java in a Windows environment on a PC with 2 GB RAM and then later integrated them. We measured the performance of the proposed approach by using the Mean Absolute Percentage Error (MAPE) as the measure of accuracy. MAPE is defined as

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i}\right| \tag{5}$$

Where n is the number of missing values in a given dataset. $\hat{x}_i$ is the value predicted (imputed) by the hybrid model for the missing value and x_i is the actual value. The average MAPE val-

ues and standard deviation of MAPE values are computed over 10 fold cross validation experiments on all datasets and are presented in Table 3.

For online data imputation, the number of clusters obtained by ECM is dictated by a parameter known as the distance threshold *Dthr*. The *Dthr* value that yields the best reduction in MAPE is obtained by conducting several experiments and the least MAPE value thus obtained is tabulated. Similarly, GRNN which was employed in Stage-2 has a parameter known as smoothing factor (σ). The value of σ that yields the least MAPE is found and the corresponding MAPE is finally tabulated.

For offline data imputation, the number of clusters (K) for K-Means/K-Medoids is found by using a systematic procedure. MLP which was employed in Stage-2 has three parameters viz., the number of hidden nodes, the Learning Rate (LR) and the Momentum Rate (MR). The combination of these parameters that yields the least MAPE is obtained and the corresponding MAPE is tabulated. The average and standard deviation of the MAPE values of stage 1 and stage 2 using K-Means/K-Medoids in Stage-1 and MLP/GRNN in Stage-2 for different datasets that were used in the experiment are presented in Table 3.

For the Boston housing, dataset the MAPE is reduced from 26.55% to 21.01% by employing the K-Means and MLP in Stage-1 and Stage-2 respectively. The value of MAPE is further reduced to 19.57% when GRNN is used instead of MLP in Stage-2. A reduction of 2.6% in MAPE is observed when K-Means is replaced by K-Medoids in Stage-1. The MAPE is reduced from

Table 3. Average MAPE values and Standard Deviation (SD) (in parenthesis) values

DATASET	K-MEANS+MLP		K-MEDOIDS+MLP		K-MEANS+GRNN		K-MEDOIDS +GRNN		ECM+GRNN		IMLS-4	INI
	S -1	S-2	S -1	S-2	S -1	S-2	S -1	S -2	S-1	S-2		
Boston Housing	26.55 (4.48)	21.01 (4.16)	23.95 (6.71)	17.69 (5.27)	26.55 (4.48)	19.57 (5.44)	23.95 (6.71)	**17.68** (4.75)	24.34 (7.58)	18.08 (4.56)	35.38 (15.44)	49.14 (14.53)
Forest Fires	37.58 (6.12)	26.61 (5.23)	29.17 (5.69)	24.46 (4.64)	37.58 (6.12)	26.21 (5.05)	29.17 (5.69)	**22.97** (4.09)	32.74 (11.5)	24.38 (4.57)	37.07 (5.78)	48.26 (10.62)
Auto MPG	35.14 (7.26)	23.75 (4.52)	30.54 (7.65)	20.70 (4.16)	35.14 (7.26)	20.27 (3.61)	30.54 (7.65)	**16.66** (3.56)	20.5 (3.56)	17 (3.56)	84.53 (20.93)	134.74 (31.83)
Body Fat	10.93 (3.46)	7.83 (1.64)	9.81 (2.28)	6.46 (1.01)	10.93 (3.46)	6.96 (2.60)	9.81 (2.28)	**5.37** (2.60)	7.44 (2.57)	5.56 (2.58)	11.27 (4.39)	80.54 (9.68)
Wine	28.84 (5.42)	21.58 (3.87)	18.54 (4.53)	15.73 (2.82)	28.84 (5.42)	16.21 (3.56)	18.54 (4.53)	**14.75** (3.58)	19.33 (3.18)	15.61 (3.58)	39.31 (8.74)	38.97 (5.97)
Prima Indian	33.60 (2.85)	29.70 (3.39)	30.80 (4.35)	26.63 (3.0)	34.68 (2.85)	28.3 (2.25)	30.80 (4.35)	**26.33** (2.19)	41.64 (8.19)	26.51 (6.52)	42.14 (5.51)	68.02 (9.28)
Iris	11.91 (2.68)	9.41 (1.97)	11.47 (2.64)	9.17 (2.12)	11.91 (2.68)	8.79 (2.79)	11.47 (2.64)	8.04 (3.0)	7.59 (3.1)	**6.30** (2.84)	13.90 (14.78)	106.23 (21.06)
SpectF	13.48 (3.19)	12.14 (2.68)	12.38 (2.87)	10.65 (1.41)	13.48 (3.19)	10.61 (2.11)	12.38 (2.87)	**10.22** (2.0)	13.11 (5.96)	10.35 (2.72)	13.90 (3.56)	106.23 (4.5)
UK Credit	46.45 (1.06)	32.17 (11.56)	39.76 (9.65)	25.42 (10.58)	46.45 (1.06)	29.8 (10.82)	39.76 (9.65)	24.04 (7.76)	35.60 (16.1)	**21.93** (7.33)	40.68 (9.14)	46.45 (13.33)
Spanish	62.25 (36.06)	39.91 (13.06)	53.13 (28.43)	32.45 (12.5)	62.25 (36.06)	37.96 (21.9)	53.13 (8.08)	**26.01** (14.59)	59.64 (43.2)	34.11 (19.85)	67.32 (35.3)	106.88 (50.3)
Turkish	48.56 (24.75)	33.01 (21.34)	39.66 (21.62)	26.9 (18.57)	48.56 (24.75)	25.9 (15.12)	39.66 (24.67)	**19.34** (13.03)	36.48 (15.1)	22.34 (14.76)	64.13 (17.72)	92.23 (49.87)
UK Bankruptcy	46.39 (13.56)	30.96 (10.58)	39.28 (16.09)	29.69 (9.63)	46.39 (13.56)	29.06 (9.94)	39.28 (2.59)	**28.39** (8.76)	42.85 (24.1)	29.07 (10.12)	204.18 (106.9)	279.42 (136.1)

23.95% to 17.69% by using K-Medoids and MLP in Stage-1 and Stage-2 respectively. The MAPE is further reduced to 17.68% when GRNN is used in Stage-2.

For Forest fires dataset, the MAPE is reduced from 37.58% to 26.61% by using K-Means and MLP in Stage-1 and Stage-2 respectively. There is not much reduction in the value of MAPE when GRNN is used instead of MLP in Stage-2. A huge reduction of 8.41% in the value of MAPE is observed when K-Means is replaced by K-Medoids in Stage-1. The MAPE is reduced from 29.17% to 24.46% by using K-Medoids and MLP in Stage-1 and Stage-2 respectively. The MAPE is reduced to 22.97% when GRNN is used instead of MLP in Stage-2.

In regards to the Auto mpg dataset, a reduction of 11.39% (from Stage-1 to Stage-2) in MAPE is observed by using K-Means and MLP in Stage-1 and Stage-2 respectively. The MAPE reduced from 23.75% to 20.27% by using GRNN instead of MLP in Stage-2. The MAPE reduced from 35.14% to 30.54% when K-Means is replaced by K-Medoids in Stage-1. The MAPE reduced from 30.54% to 20.70% by using K-Medoids and MLP in Stage-1 and Stage-2 respectively. The MAPE reduced nearly by 14% (from30.54% to 16.66%) when MLP is replaced by GRNN in Stage-2.

A MAPE that is less than 10% is observed for all the proposed imputation techniques with the Body fat dataset. The MAPE is reduced from 10.93% in Stage-1 to 7.83% in Stage-2 by using K-Means and MLP in Stage-1 and Stage-2 respectively. The value of MAPE is reduced from 7.83% to 6.96%when GRNN is employed instead of MLP in Stage-2. The MAPE is reduced by 1.12%when K-Medoids is employed instead of K-Means in Stage-1. The MAPE reduced from 9.81% in Stage-1 to 6.46% in Stage-2 by using K-Medoids and MLP in Stage-1 and Stage-2 respectively. The MAPE is reduced from 6.46% to 5.37% when MLP is replaced by GRNN in Stage-2.

For the Wine dataset, the MAPE is reduced from 28.84% in Stage-1 to 21.58 % in Stage-2 by using K-Means and MLP in Stage-1 and Stage-2 respectively. A reduction of 5.37% in MAPE is observed by using GRNN in place of MLP in Stage-2. A drastic reduction of 10.3% in MAPE is observed when K-Medoids is employed instead of K-Means in Stage-1. The MAPE is reduced from 18.54% in Stage-1 to 15.73% in Stage-2 by using K-Medoids and MLP in Stage-1and Stage-2 respectively. The MAPE is further reduced to 14.75% when GRNN is used in Stage-2.

For the Prima Indian dataset, the MAPE is reduced from 33.6% to 29.7% by using K-Means and MLP in Stage-1 and Stage-2 respectively. The MAPE is further reduced to 28.3% when GRNN is used instead of MLP in Stage-2. A reduction of 2.8% in MAPE is observed when K-Means is replaced by K-Medoids in Stage-1. The value of MAPE is reduced from 30.8% in Stage-1 to 26.63 % in Stage-2 by using K-Medoids and MLP in Stage-1 and Stage-2 respectively. The MAPE is reduced to 26.33% when GRNN is used in Stage-2.

For the Iris dataset, the MAPE value is reduced from 11.91% to 9.41 % by using K-Means and MLP in Stage-1 and Stage-2 respectively. The MAPE is reduced to 8.79% by using GRNN in Stage-2.There is not much reduction in the MAPE value when K-Medoids is used instead of K-Means in Stage-1. The MAPE is reduced from 11.47% to 9.17% by using K-Medoids and MLP in Stage-1and Stage-2 respectively. The MAPE is reduced to 8.04% when GRNN is used in Stage-2.

For the SpectF dataset, the MAPE is reduced from 13.48% to 12.14% by using K-Means and MLP in Stage-1 and Stage-2 respectively. The value of MAPE is reduced to 10.61% when GRNN is used instead of MLP in Stage-2. The MAPE is reduced from 12.38% in Stage-1 to 10.65% inStage-2 by using K-Medoids and MLP in Stage-1 and Stage-2 respectively. The MAPE is re-

duced to 10.22% when GRNN is used in Stage-2.

For the UK Credit dataset a huge reduction of 14.28% (46.45% to 32.17%) in MAPE from Stage-1 toStage-2 is observed by using K-means and MLP in Stage-1 and Stage-2 respectively. The value of MAPE is reduced from 32.17% to 29.8% when MLP is replaced by GRNN in Stage-2. A reduction of 6.69% in MAPE is observed by using K-Medoids instead of K-Means in Stage-1. A huge reduction of 14.34% (39.76% to 25.42%) in MAPE is obtained by using K-Medoids and MLP in satge-1 and Stage-2 respectively. A reduction of 15.72% (39.76% to 24.04%) in MAPE from Stage-1 to Stage-2 is obtained by using K-Medoids and GRNN in Stage-1 and Stage-2respectively.

A massive reduction in the value of MAPE from Stage-1 to Stage-2 is observed for the Spanish Bankruptcy dataset. A reduction of 22.34% (62.25% to 39.91%) in MAPE is obtained by using K-Means and MLP in Stage-1 and Stage-2 respectively. A reduction of 20.68% (53.13% to32.45%), 24.29% (62.25% to 37.96%) is obtained by using K-Medoids and MLP, K-Means and GRNN in Stage-1 and Stage-2 respectively. The MAPE is reduced by 27.12% (53.13% to 26.01%) with K-Medoids and GRNN in Stage-1 and Stage-2 respectively. The MAPE is reduced from 62.25% to 53.13% when K-Means is replaced by K-Medoids in Stage-1.

For the Turkish Bankruptcy dataset a reduction of 15.55% (48.56% to 33.01%) is obtained by using K-Means and MLP in Stage-1 and Stage-2 respectively. A reduction of 12.76% (39.66% to26.9%), 22.66% (48.56% to 25.9%) is obtained by using K-Medoids and MLP, K-Means and GRNN in Stage-1 and Stage-2 respectively. A huge reduction of 20.32% (39.66% to 19.34%) in the value of MAPE is obtained by employing K-Medoids and GRNN in Stage-1 and Stage-2respectively.

For the UK Bankruptcy dataset, a reduction of 15.43% (46.39% to 30.96%) is obtained by using K-Means and MLP in Stage-1 and Stage-2 respectively. A reduction of 9.59% (39.28% to 26.69 %),17.33% (46.39% to 29.06%) is obtained by using K-Medoids and MLP, K-Means and GRNN inStage-1 and Stage-2 respectively. The MAPE is reduced from 39.28% to 28.39% by using K-Medoids and GRNN in Stage-1 and Stage-2 respectively.

From the experiments, it is observed that the reduction in MAPE is least when K-Medoids and GRNN are employed in Stage-1 and Stage-2 respectively in all datasets. Thus, it is the best imputation technique out of all the offline techniques employed here. Furthermore, the proposed offline and online imputation techniques outperform other imputation techniques viz., IMLS and INI. The results of the online imputation method, ECM+GRNN, are also presented in Table 2. For the Iris and the UK bankruptcy datasets a MAPE of 6.3% and 21.93% is obtained by using online imputation technique viz., ECM+GRNN, which is better than the MAPE obtained by using the best offline technique viz., K-Medoid+GRNN. For the Boston housing dataset, a MAPE of 18.08% is obtained by using the online imputation which is nearly equal to 17.68% obtained by using the best offline imputation technique, K-Medoid+GRNN. For Auto mpg, Bodyfat, Wine, Pima Indian and SpectF datasets, the MAPE values of 17%, 5.56%, 15.61%, 26.51% and10.35% respectively are obtained using the online imputation, ECM+GRNN. For the Auto mpg ,Body fat, Wine, Pima Indian and SpectF datasets MAPE values of 16.66%, 5.37%, 14.75%, 26.33%and 10.22% respectively are yielded by the best offline imputation, K-Medoid+GRNN.

Since the MAPE values obtained by the online imputation technique and the best offline imputation technique are close for the Auto mpg, Body fat, Wine, Pima Indian and SpectF datasets, we investigated the statistical significance by performing the t-test at 1% level of significance.

Table 4. t-Test Values for various techniques.

Technique / Dataset	WINE	AUTO MPG	FOR-EST FIRES	IRIS	SPEC-TF	BODY FAT	PRIMA INDI-AN	U K CRED-IT	SPAN-ISH	TURKI-SH	US-BANK-RUPT-CY	BOS-TON HOU-SI-NG
K-Means+MLP vs K-Means+GRNN	0.855	1.921	0.794	0.568	1.360	0.901	1.094	0.184	0.243	0.836	0.986	1.726
K-Means+MLP vs K-Medoids+MLP	0.255	1.588	1.618	0.254	1.358	2.253	2.145	0.699	1.304	0.664	0.281	2.526
K-Means+MLP vs K-Medoids+GRNN	1.457	1.805	2.412	1.199	1.415	2.531	2.646	1.099	3.321	0.887	0.592	2.991
K-Means+GRNN vs K-Medoids+MLP	0.302	0.247	0.807	0.344	0.049	0.565	1.402	0.916	0.690	0.134	0.743	0.660
K-Means+GRNN vs K-Medoids+GRNN	0.778	0.142	1.575	0.577	0.423	1.363	1.982	1.369	2.236	0.017	0.466	0.926
K-Means+GRNN vs ECM+GRNN	0.353	2.040	0.847	1.975	0.241	1.203	0.817	1.905	0.412	0.539	0.589	0.768
K-Medoids+MLP vs K-Medoids+GRNN	0.543	0.116	0.762	0.971	0.553	1.231	0.259	0.332	2.157	0.157	0.316	0.233
K-Medoids+MLP vs ECM+GRNN	0.081	2.138	0.038	2.556	0.311	1.020	0.053	0.857	0.223	0.615	0.139	0.058
K-Medoids+GRNN vs ECM+GRNN	0.424	2.195	0.727	1.331	0.119	0.166	0.085	0.624	1.895	0.561	0.162	0.192
K-Means+MLP vs ECM+GRNN	1.256	3.729	1.662	2.836	1.392	2.347	1.372	1.605	0.772	1.273	0.407	2.869

Thus, the t-test is performed on 10-folds (10 experiments) of all the datasets to see whether the difference in performance between offline and online imputation methods is statistically significant. The t-test values are presented in Table 4. Since the table value of the t distribution with 18 degrees of freedom (10+10-2=18) at 1% level of significance is 2.87, the computed t-test values for all the datasets indicate that the difference between the online method and the best offline method (indicated in bold) is not statistically significant at 1% level of significance. Furthermore, the t-test values (see Table 4) indicate that there is no significant difference between the offline and online imputation techniques (indicated in italics). A t-test is not performed with IMLS and INI as the proposed imputation techniques clearly outperform them by a large margin. Therefore, we infer that the proposed online imputation method can be used as a viable alternative since it is faster and involves single iteration in both stages 1 and 2. This is a significant outcome of the study. Furthermore, another important point to be noted is that the GRNN working in stage 2 always improved (reduced) the MAPE values in both the offline and online methods in all datasets, which is another achievement of the study.

7. CONCLUSIONS

We have proposed a computational intelligence hybrid for fast online data imputation and for the extended version of offline data imputation. The effectiveness of the proposed techniques has been demonstrated on 8 benchmark datasets and 4 bank datasets. The results demonstrate that there is a significant reduction in MAPE from Stage-1 to Stage-2 in all of the methods and that the best offline imputation technique is K-Medoids+GRNN as it gives best reduction in MAPE. In addition, the dif-

ference between the best offline imputation, viz., K-Medoids+GRNN and the proposed online imputation method viz., ECM +GRNN is statistically insignificant. This is demonstrated by t-test conducted on the 10 folds at 1% level of significance on all datasets used for the experiment. So, we can conclude that the proposed online imputation technique (ECM+GRNN) is a viable alternative to the existing methods of offline data imputation. The proposed online data imputation technique uses a fast and one pass algorithms but it needs user intervention for fine tuning two parameters, i.e., *dthr* for ECM and *smoothing factor (σ)* for GRNN. The next stage of research will focus on enhancing the proposed imputation technique, so that it doesn't need user intervention for parameter tuning, while retaining its predictive efficiency.

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, 2005, pp. 207-212.

[2] N. Ankaiah, and V. Ravi, "A novel soft computing hybrid for data imputation", DMIN, Las Vegas, USA, 2011.

[3] P. C. Austin and M. D. Escobar, "Bayesian modeling of missing data in clinical research", Computational Statics & Data Analysis, vol. 49, no. 3, 2005, pp. 821-836.

[4] G. Batista and M. C. Monard, "A study of K-nearest neighbor as an imputation method", Abraham A et al (eds) Hybrid Intelligent Systems, Ser Front Artificial Intelligence Applications,vol. 87, 2002, pp. 251–260.

[5] G. Batista and M. C. Monard, Experimental comparison of K-nearest neighbor and mean or mode imputation methods with the internal strategies used by C4.5 and CN2 to treat missing data, December,2003, Technical Report, University of Sao Paulo.

[6] J. Chen, H. Huang, F. Tian and S. Tian, "A selective Bayes Classifier for classifying incomplete data based on gain ratio,"Knowledge Based Systems, vol. 21, no. 7, 2008,pp. 530-534.

[7] M. Cooke, P. Green and M. Crawford,"Handling missing data in speech recognition,"International Conference on Spoken Language Process, 1994, pp. 1555– 1558.

[8] W. S. Desabro, P. E. Green and J. D. Carroll, "Missing data in product-concept testing," Decision Sciences, vol. 17, no. 2, 1986,pp. 163-185.

[9] A.G. Di Nuovo, "Missing data analysis with fuzzy C-Means: A study of its application in a psychological scenario," Expert Systems With Applications, vol. 38, no. 6, 2011, pp. 6793-6797.

[10] A.Elshorbagy, S. P. Simonovic and U. S. Panu, "Estimation of missing streamflow data using the principles of chaos theory,"Journal of Hydrology, vol. 255, no. 1-4, 2002,pp.123-133.

[11] B.Gabrys,"Neuro-Fuzzy approach to processing inputs with missing values in pattern recognition problems,"International Journal of Approximate Reasoning,vol. 30, no. 3, 2002,pp. 149-179.

[12] P.J. Garcı´a-Laencina, J. L. Sancho-Go´mez andA. R. Figueiras-Vidal, "Pattern classification with missing data: a review,"Neural Computing & Applications, vol. 19, no. 2, 2012, pp. 263–282.

[13] A.Gupta and M. S. Lam, "Estimating missing values using neural networks", Journal Of Operational Research Society,vol. 47, no. 2, 1996, pp. 229–238.

[14] J. Han and M. Kamber,Data Mining concepts and techniques, Morgan Kufmann Publishers, San Francisco, 2008.

[15] S. Henley,"The problem of missing data in geoscience databases,"Computers & Geosciences, vol. 32, no. 9, 2006, pp. 1368-1377.

[16] J. Jerez,I. Molina, J. Subirates and L. Franco, "Missing data imputation in breast cancer prognosis,"BioMed'06 Proceedings of the 24th IASTED International Conference on Biomedical Engineering, USA, 2006, pp. 323-328.

[17] N. K. Kasabov andQ. Song, "DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Appli-

cation for Time-Series Prediction,"IEEE transactions on fuzzy systems, vol. 10, no. 2, 2002, pp. 144-154.
[18] R. B. Kline, Principles and Practice of Structural Equation Modelling, Guliford Press, New York.
[19] N. M. Laird, "Missing data in longitudinal studies," Statistics in Medicine, vol. 7, no. 1-2, 1988, pp. 305-315.
[20] R. J. A. Little and D. B. Rubin, Statistical analysis with missing data, second edition, Wiley, New York, 2002, pp. 2-250.
[21] H. LiuandS. Zhang, "Noisy data elimination using mutual k-nearest neighbor for classification mining," Journal of Systems and Software, vol. 85, no. 5, 2012, pp. 1067-1074.
[22] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations,"Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1967, pp. 281-297.
[23] W. G. Madlow, H. Nisselson and H. Olkin, Incomplete data in sample surveys: Report and case Studies volume 1, Academic Press, New York, 1983.
[24] M. Marseguerra andA. Zoia,"The autoassociative neural network in signal analysis. II. Application to on-line monitoring of a simulated BWR component,"Annals of Nuclear Energy, vol. 32, no.11, 2002, pp. 1207– 1223.
[25] T. Marwala and S. Chakraverty, "Fault classification in structures with incomplete measured data using auto associative neural networks and genetic algorithm,"Current Science India, vol. 90, no 4, 2006, pp. 542–548.
[26] P. Merlin, A. Sorjamaa, B. Maillet andA. Lendasse, "X-SOM and L-SOM: A double classification approach for missing value imputation," Neurocomputing,vol. 73, no. 7-9, 2010, pp. 1103-1108.
[27] T. R. Nkuna andJ. O. Odiyo, "Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks,"Physics and Chemistry of the Earth, Parts A/B/C,vol. 36, no. 14-15, 2011, pp. 830-835.
[28] S. Nordbotten, "Neural network imputation applied to the Norwegian 1990 population census data,"Journal of Official Statistics,vol. 12, no. 4, 1996,pp. 385–401.
[29] A.Ragel andB. Cremilleux, "MVC—a preprocessing method to deal with missing values,"Knowledge Based Systems,vol. 12, no. 5-6, 1999, pp. 285-291.
[30] P. L. Roth, F. S. Switzer&D. M. Switzer, "Missing data in multiple item scales: a Monte Carlo analysis of missing data techniques,"Organizational research methods,vol. 2, no. 3, 1999, pp. 211-232.
[31] A.E. Rumlhart, G. E. Hinton and R. J. Williams, "Learning internal representations by error propagation,"Parallel distributed processing: explorations in the microstructure of cognition, vol. 1, 1986, pp. 318-362.
[32] T. Samad andS. A. Harp, "Self-organization with partial data,"Network: Computation in Neural Systems,vol. 3, no. 2, 1992, pp. 205-212.
[33] J.L. Schafer, Analysisof incomplete multivariate data, Chapman & Hall,Florida, 1997.
[34] P. K. Sharpe and R. J. Solly, "Dealing with missing values in neural network-based diagnostic systems,"Neural Computing & Applications,vol. 3, no. 2, 1995,pp. 73–77.
[35] A.L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello andM. D. Cubiles-de-la-Vega, "Missing value imputation on missing completely at random data using multilayer perceptrons,"Neural Networks, vol. 24, no. 1, 2011, pp. 121-129.
[36] Q. Song andM. Shepperd, "A new imputation method for small software project data sets,"Journal of Systems and Software,vol. 80, no. 1, 2007, pp. 1-62.
[37] D. F. Specht, "A General Regression Neural Network,"IEEE transactions on neural networks, vol. 2, no.6, 1991, pp. 568-576.
[38] O. Troyanskaya, M. Cantor, O. Alter, G. Sherlock, P. Brown, D. Botstein, R. Tibshirani, T. Hastie and R. Altman, "Missing value estimation methods for DNA microarrays,"Bioinformatics,vol. 17, no. 6, 2001, pp. 520–525.
[39] I. Wasito and B. Mirkin, "Nearest Neighbor approach in the least-squares data imputation algorithms", Information Sciences, vol. 169, no. 1-2, 2005, pp. 1-25.
[40] I. Wasito and B. Mirkin, "Nearest Neighbor in the least-squares data imputation algorithms with different missing patterns", Computational Statistics and Data Analysis, vol. 50, no. 4, 2005, pp. 926-949.
[41] S. Y. Yoon and S. Y. Lee, "Training algorithm with incomplete data for feed-forward neural net-

works,"Neural Processing Letters,vol. 10, no. 3, 1999, pp. 171–179.

Kancherla Jonah Nishanth

He is currently a Technology Manager at Andhra Bank. He received M.Tech degree in Information Technology from University Of Hyderabad (UoH), Hyderabad, India in 2012. His research interests include data mining, data analytics and machine learning.

Dr. Vadlamani Ravi

Dr. Vadlamani Ravi is Associate Professor, in the Institute for Development and Research in Banking Technology, Hyderabad since February 2010. He obtained his Ph.D. in the area of Soft Computing from Osmania University, Hyderabad and RWTH Aachen, Germany (2001); MS (Science and Technology) from BITS, Pilani and M.Sc. (Statistics & Operations Research) from IIT, Bombay. At IDRBT, he spearheads the CRM lab, first-of-its-kind in India and evangelizes CRM in a big way by conducting customized training programmes for bankers on CRM subsuming OCRM & ACRM; Data Warehousing & Data Mining and conducting POC for banks etc.

Prior to joining IDRBT as Assistant Professor in April 2005, he worked as a Faculty at the Institute of Systems Science (ISS), National University of Singapore (April 2002 - March 2005). At ISS, he was involved in teaching M.Tech. (Knowledge Engineering) and research in the areas of Fuzzy Systems, Neural Networks, Soft Computing Systems and Data Mining & Machine Learning. Further, he consulted for Seagate Technologies, Singapore and Knowledge Dynamics Pte. Ltd., Singapore, on data mining projects. Before leaving for Singapore, he worked as Assistant Director (Scientist E1) from 1996 -2002 and Scientist C from 1993-1996 respectively at the Indian Institute of Chemical Technology (IICT), Hyderabad.

He has 132 papers to his credit with the break-up of 64 papers in refereed International Journals, 6 papers in refereed National Journals, 47 papers in refereed International Conferences and 3 papers in refereed National Conferences and 14 invited book chapters. His papers appeared in Applied Soft Computing, Soft Computing, Asia-Pacific Journal of Operational Research, Decision Support Systems, European Journal of Operational Research, Expert Systems with Applications, Fuzzy Sets and Systems, IEEE Transactions on Fuzzy Systems, IEEE Transactions on Reliability, Information Sciences, Journal of Systems and Software, Knowledge Based Systems, IJUFKS, IJCIA, IJAEC, IJDMMM, IJIDS, IJDATS, IJISSS, IJCIR, IJCISIM, IJBIC, Computers and Chemical Engineering, Canadian Geotechnical Journal, Biochemical Engineering Journal, Bio information, Journal of Services Research etc. He also edited a Book entitled "Advances in Banking Technology and Management: Impacts of ICT and CRM" (http://www.igi-global.com/reference/details.asp?id=6995), published by IGI Global, USA, 2007.