Neurocomputing 218 (2016) 17–25

Contents lists available at ScienceDirect

Neurocomputing

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

Brief papers

# Probabilistic neural network based categorical data imputation

Kancherla Jonah Nishanth, Vadlamani Ravi *

*Center of Excellence in Analytics, Institute for Development and Research in Banking Technology (IDRBT), Castle Hills Road #1, Masab Tank, Hyderabad 500057, India*

ARTICLE INFO

*Article history:*
Received 19 July 2015
Received in revised form
21 April 2016
Accepted 11 August 2016
Communicated by Bijaya Ketan Panigrahi
Available online 27 August 2016

*Keywords:*
Categorical data imputation
Probabilistic Neural Network (PNN)
Random Forest (RF)
Decision Tree (DT)

ABSTRACT

Real world datasets contain both numerical and categorical attributes. Very often missing values are present in both numerical and categorical attributes. The missing data has to be imputed as the inferences made from complete data are often more accurate and reliable than those made from incomplete data [15]. Also, most of the data mining algorithms cannot work with incomplete datasets. The paper proposes a novel soft computing architecture for categorical data imputation. The proposed imputation technique employs Probabilistic Neural Network (PNN) preceded by mode for imputing the missing categorical data. The effectiveness of the proposed imputation technique is tested on 4 benchmark datasets under the 10 fold-cross validation framework. In all datasets, except Mushroom, which are complete, some values, which are randomly removed, are treated as missing values. The performance of the proposed imputation technique is compared with that of 3 statistical and 3 machine learning methods for data imputation. The comparison of the mode+PNN imputation technique with mode, K-Nearest Neighbor (K-NN), Hot Deck (HD), Naive Bayes, Random Forest (RF) and J48 (Decision Tree) imputation techniques demonstrates that the proposed method is efficient, especially when the percentage of missing values is high, for records having more than one missing value and for records having a large number of categories for each categorical variable.



## 1. Introduction

Missing data is observed in almost all real world datasets. Missing data forms a serious hurdle for many statistical analyses and data mining techniques. For obtaining accurate inferences from the data, the data should be complete. Imputation is the substitution of a missing data point or missing component of a data point with the most plausible value. In case the dataset contains missing values, the missing values should be imputed before performing any further analysis on the data. Missing data has two major negative effects [1]: (i) it has a negative impact on statistical power; (ii) it may result in biased estimates i.e., the measures of central tendency and dispersion may be biased.

Also, analyses on complete data yield inferences that are more precise and dependable than those from incomplete data. Statistical and computational intelligence techniques for data mining tasks such as classification, regression, association rule mining and outlier analysis require accurate and complete data. Data imputation is of great use in such applications if the data contains missing values.

Almost all fields are replete with datasets having missing values. For instance, in surveys, data may be missing due to variety of reasons such as errors in data entry, disclosure restrictions, failure to complete the entire questionnaire, absence of the respondent at the time of survey and when the response does not apply for an individual (e.g., questions regarding the years of marriage for a respondent who has never been married) [2]. In the geosciences, data items in the observational data sets may be missing altogether, or they may be imprecise in one way or another [3]. Datasets for effort prediction in software project management contain missing values [4]. Geophysical time series datasets also contain missing data [5]. Reasons such as equipment malfunctioning, outliers and incorrect data entry contribute to missing values in many practical observations [6]. Due to faults in the data acquisition process, data tend to be missing in environmental research data sets. In automatic speech recognition speech samples that are corrupted by very high levels of noise are considered to be missing data [7]. Datasets for business and financial applications may also contain missing data. Missing data problems are common in health research (e.g. retrospective and prospective studies). Longitudinal studies which collect data on a set of subjects repeatedly over time are afflicted by attrition;subjects drop out because they move or suffer side effects from drugs, or for other often unknown reasons. In biological research with DNA microarrays, gene data may be missing due to reasons such as a scratch on the slide containing the gene sample and contaminated samples [8].

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2016.08.044
0925-2312/

The standard categorization of missing data mechanism [9], considers data: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR).

1. *Missing Completely At Random (MCAR).* It occurs when the probability of an instance having missing value on some variable X is independent of the variable itself and on the values of any other variables in the dataset. Typical examples of MCAR are when the gender or phone number of a customer is missing in customer's database, when a tube containing a blood sample of a study subject is broken by accident or when a questionnaire of a study subject is accidentally lost. Possible reasons for MCAR include manual data entry procedure, incorrect measurements, equipment error, changes in experimental design etc.
2. *Missing At Random (MAR).* It occurs when the probability of an instance having missing value on some variable X depends on other variables in the database but not the variable itself. For example, if the income level of a customer is missing it can be estimated from other variables like customer's profession, age and qualification.
3. *Not Missing At Random (NMAR).* It occurs when the probability of an instance having missing value on some variable X depends on the variable itself. For instance, if citizens did not participate in a survey, then NMAR occurs.

In real life datasets, numerical and categorical attributes contain missing values. A lot of literature is available for numerical data imputation. To impute the incomplete or missing numerical data, several techniques based on statistical analysis are reported [10]. These methods include mean substitution methods, Hot Deck imputation, regression methods, expectation maximization, and multiple imputation methods. Other machine learning based methods include SOM [11], K-Nearest Neighbor [12], multi layer perceptron [13], fuzzy-neural network [14] and auto-associative neural network imputation with genetic algorithms [15].

Even though numerous studies are reported regarding the imputation of the numerical or continuous data, there is not much research devoted to categorical data imputation with machine learning despite the fact that many real life datasets contain categorical attributes. Categorical data are common in many fields like education (e.g., student responses to an exam question with the categories correct and incorrect), marketing (e.g., consumer preference among the leading brands of a product), banking (e.g., type of loan with categories house loan, vehicle loan, educational loan etc.), social and biomedical sciences, behavioral sciences (e.g., type of mental illness, viz., schizophrenia, depression, neurosis), epidemiology and public health, genetics (type of allele inherited by an offspring) and zoology (e.g., alligators primary food preference with categories as fish, invertebrate, reptile) [16]. The existing methods for categorical data imputation employ techniques that are originally designed for numerical variables. These techniques include listwise deletion, mode imputation, model-based procedures such as factored likelihoods [9] and Bayesian methods [17].

This paper proposes a novel soft computing based imputation technique based on PNN for categorical data. PNN is employed in this study because of its ability to identify complex non-linear relationships between a set of input and output variables. Also, PNN can train fast on sparse data sets and it is a universal approximator for smooth classification problems [18]. Finally, it should be noted the paper proposes a soft computing technique only for categorical data imputation. It does not evaluate the impact of imputation on classification accuracy of a classifier.

The remainder of the paper is organized as follows: A brief review of literature on imputation of missing categorical data is presented in Section 2. Probabilistic Neural Network (PNN) is described briefly in Section 3. The proposed method and experimental setup are described in Section 4. Results and discussions are presented in Section 5, followed by conclusions in Section 6.

## 2. Review of categorical data imputation techniques

The methods for handling missing categorical data can be broadly classified three categories [9]: (1) deletion, (2) Model the distribution of missing data and then estimating them based on certain parameters and (3) Imputation. Each of these is discussed below.

### 2.1. Deletion procedures

The deletion techniques simply delete the cases that contain missing data. Deletion procedures are generally easy to carry out and they are the default choice for many statistical packages. This approach has two forms: (i) Listwize deletion which omits the cases or in-stances 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 determines the extent of missing data on each attribute and instance, and deletes the instances or attributes with high levels of missing data. It is good when the overall sample size is small or when the number of missing data observations are large [19].

### 2.2. Model – based procedures

Maximum likelihood is one of the model-based procedures. It assumes that the observed data is a sample drawn from a multivariate normal distribution [20]. Expectation Maximization (EM) algorithm is an iterative method for finding the maximum likelihood estimates of the data. EM algorithm for missing data consists of estimating the model parameters from the available data and then determining the missing values based on the estimated parameters in the first iteration. The consecutive iterations re-estimate the missing data based on new parameters then recalculate the new parameter estimates based on actual and re-estimated missing data [9]. West and Dawson [21] proposed an approach that employs EM algorithm and Bayesian prior for incomplete categorical data imputation in repeated measures setting that allows missing data to depend on other observed outcomes.

### 2.3. Imputation procedures

The earliest and simplest technique for categorical data imputation is the mode imputation. In mode imputation, the mode of an attribute is used to impute the missing value of the corresponding attribute. The disadvantage of mode imputation is that it leads to underestimation of the population variance. Hot-Deck imputation replaces the missing value in a variable or an attribute in an incomplete observation with the corresponding variable or attribute of the closest complete observation. Coutinho et.al [22] used a calibrated hot-deck donor method for categorical data imputation. They employed univariate Hot Deck, random Hot Deck and K-NN Hot Deck imputation to impute the missing data. In K-NN imputation the K nearest neighbors of an incomplete record are calculated and the mode of the neighbors is used to impute the missing value [23–25].

Buuren and Rijckevorsel [26] proposed an imputation method, which imputes the categorical data by maximizing the internal consistency. The squared correlation ratio $n^2$, proposed by Guttman, was employed by them to measure the internal consistency of the data. A log- linear model based multiple imputation method was proposed by Schafer [27]. Sentas and Angelis [28] employed

Multinominal Logistic Regression (MLR) to impute the missing categorical data in the projects of ISBSG to evaluate the performance of the imputation technique. Ray-del-Castillo and Cardenosa [29] presented an imputation method that uses fuzzy min-max neural networks for imputing the categorical data. Ferrari et. al [30] employed Non Linear Principal Component Analysis (NLPCA) for imputation. The procedure is based on an iterative algorithm where sequentially missing categories for one element are replaced with corresponding values observed for a most similar element from a complete dataset. Wang et.al [31] introduced a new family of distributions, grouped Dirichlet distributions (GDD), which includes the classical Dirichlet distribution (DD) as a special case. First, he developed distribution theory and explored important distribution properties. Second, he proposed a new way to manage the statistical analysis of incomplete categorical data based on the GDD.

Farhangfar [32] studied the effect of missing discrete data imputation using five single imputation methods (a mean method, a Hot deck method, a Naive-Bayes method) and one multiple imputation method (a polytomous regression based method) on classification accuracy for six popular classifiers viz., RIPPER, C4.5, K-nearest-neighbor, support vector machine with polynomial and RBF kernels, and Naive-Bayes on 15 datasets. Gebregziabher and DeSantis [33] proposed a latent class based multiple imputation approach for analyzing missing categorical covariate data in a highly stratified data model. In this approach, the missing data is imputed assuming a latent class imputation model and likelihood methods are used to analyze the imputed data.

Baker et.al [34] employed sensitivity analysis to impute non-randomly missing categorical data arising from a national health disability survey. The sensitivity analysis was based on fitting a range of ignorable missing-data models along with non-ignorable missing-data models that added one or two parameters. Kaiser [35] developed an algorithm that employs association rules for imputing categorical data. The algorithm has three variants. The first variant replaces the missing values with a special value and then it generates the association rules. The association rules with a support lower than the threshold, consequent lower than 1 and a consequent with the special value are removed. The association rules thus obtained is used for imputing the missing values. The second variant is similar to the first except that the missing values are imputed with mode if there is no suitable association rule. The third variant employs the association rules and the most common attribute value method.

A Bayesian model for learning probabilities under graphical models for incomplete categorical data is proposed by Geng et al [36]. The model employs a partial augmentation formula and a recursive formula of posterior means proposed by them. Latent class based approach for imputing categorical data was used by Vermunt et al [37]. Vermunt et al. [37] used Multiple Imputation (MI) along with latent class analysis. Thibaudeau [38] proposed an item imputation method for categorical data based on a MLE derived from a conditional probability model. Wu et al. [39] proposed Missing Categorical Data Imputation Based on Similarity (MIBOS) to impute the missing categorical data. The imputation algorithm proposed by Wu et al. [39] defines a similarity model between objects with incomplete data, then constructs the similarity matrix of objects and computes the nearest undifferentiated object sets of each object to impute the missing data iteratively. Lobato et.al [40] proposed Multi-objective Genetic algorithm for data imputation which is suitable for mixed-attribute datasets.

## 3. Overview of the Probabilistic Neural Network (PNN)

The Probabilistic Neural Network (PNN) was originally proposed by Specht [41]. PNN is a feed-forward neural network involving a one pass training algorithm used for classification and mapping of data. PNN is an implementation of statistical algorithm called kernel discriminant analysis in which the operations are organized into multilayer feed forward network with four layers: input layer, pattern layer, summation layer and output layer. It is a pattern classification network based on the classical Bayes classifier, which is statistically an optimal classifier that seeks to minimize the risk of misclassifications.

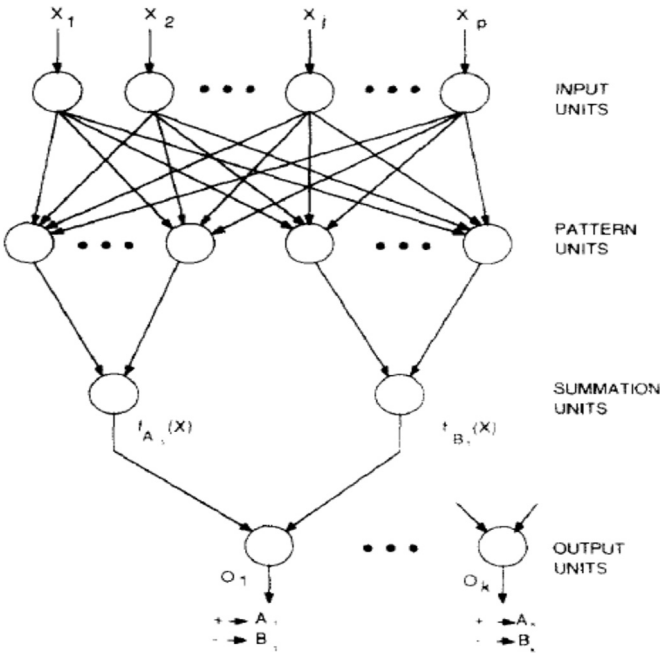


**Fig. 1.** Schematic diagram of PNN architecture.

The topology of PNN developed by Specht [41] is depicted in Fig. 1. It consists of four layers input layer, pattern layer, summation layer and the output layer. Fig. 1 depicts PNN architecture for classification of input patterns X into two categories. The input layer contains input units that are merely distributed units, which provide all predictor variables to the neurons in the second layer, the pattern layer. The pattern unit is dedicated to one cluster center, when dealing with large datasets. In the case of small datasets, all the samples are placed in the pattern layer with one pattern occupying a pattern node. When a new vector is entered into the network, it is subtracted from the stored vector representing each cluster center. Either the squares or the absolute values of the differences are summed and fed into a nonlinear activation function, which is usually taken as gaussian. The pattern unit outputs are passed onto summation units. The summation units simply sum the inputs from the pattern units that correspond to the category from which the training pattern was selected. The output, or decision, units are two-input neurons. These units produce binary outputs.

## 4. Proposed methodology

The architecture of the proposed method, preprocessing steps for dealing with categorical variables, the datasets employed in the experiment and the experimental design are described below.

### 4.1. Architecture of the proposed method

The architecture of the proposed PNN based imputation technique is depictedin Fig. 2. A record for which all the attribute values are observed is termed as a complete record. A record that

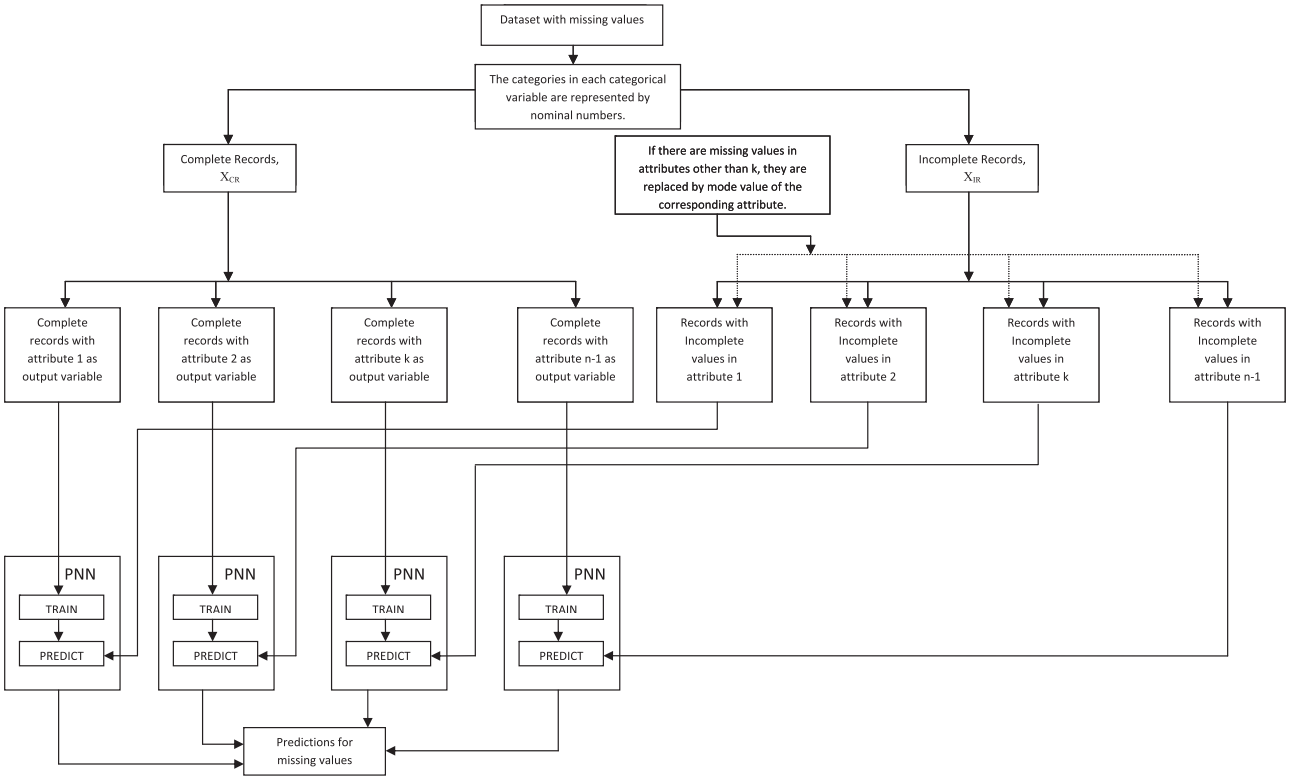


**Fig. 2.** Architecture of the proposed imputation technique.

contains missing values in one or more attributes is an incomplete record. Let $X_D$ and C denote the given dataset and the number of attributes respectively. Let $X_{CR}$ denote the set of complete records and $X_{IR}$ that of incomplete records. Let INR denote the number of incomplete records. Let $\Omega$ denote the set of attributes or variables containing missing values. The algorithm for the proposed imputation technique is depicted in Fig. 3.

Initially, the categories in each categorical variable are represented by nominal numbers and the missing value is represented by '?'. The dataset obtained after conversion of categories into nominal numbers is termed as $X_D$. The set of complete records $X_{CR}$ and incomplete records $X_{IR}$ are separated from $X_D$. The missing value in each attribute of $X_{IR}$ is imputed with the mode of the corresponding attribute and the resulting dataset is termed as $X_{IM}$. The mode of each attribute is calculated from $X_D$ by ignoring the missing values in the corresponding attribute. Now, for each attribute j containing a missing value, the set of records containing missing value in attribute j are derived from $X_{IR}$ and the resulting dataset is termed as $X_1$. The missing values (in $X_1$) in attributes other than j are replaced by mode. Datasets $X_{PR}$ (dataset for prediction) and $X_{TR}$ are obtained from $X_1$ and $X_{CR}$ respectively by designating the attribute j as the output variable. The input variables of the dataset $X_{TR}$ and $X_{PR}$ are converted into the corresponding indicator matrices, by converting the categories in the variables to as many numbers as the number of categories. Now, PNN is trained with the indicator matrix of $X_{TR}$ and then it is employed for predicting the missing values in the indicator matrix of $X_{PR}$. A similar procedure is followed for imputation with Random Forest (RF), DT and Naïve Bayes.

### 4.2. Datasets

The effectiveness of the proposed method for data imputation is tested on 4 Benchmark datasets viz., Car Evaluation, Mushroom, Solar Flare and Spect Heart, taken from UCI machine learning repository. All the datasets used for experimentation contain only categorical attributes. Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX by Bohanec and Rajkovic (1990). The model evaluates cars according to the concept structure such as car acceptability,overall pricing, buying price, price of the maintenance etc.

The Mushroom data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. The Solar Flare database contains 3 potential classes, one for the number of times a certain type of solar flare occurred in a 24 h period. Each instance represents captured features for 1 active region on the sun. The Spect Heart dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The number of attributes and records of datasets utilized for the experiment is provided in Table 1.

### 4.3. Preprocessing step

The datasets used in the experiments contain only categorical variables. The categories in each categorical variable are represented by nominal numbers. The categorical variables and the corresponding nominal numbers for each category for Car Evaluation dataset are presented in Table 2. The Car Evaluation dataset contains six input attributes i.e., buying price, price of maintenance, number of doors, capacity in terms of persons to carry, the size of luggage boot and estimated safety of the car. The buying price is represented in 4 categories i.e., very high, high, medium, low. The categories very high, high, medium and low are denoted with the nominal numbers 4, 3, 2 and 1 respectively. The nominal numbering is assigned in the ascending order of buying price. A similar approach is employed for representing the categories in other categorical variables.

```
1.  Algorithm data_imputation
2.  {
3.  for X_ij in X_D
4.    if X_ij is missing then
5.          Add X_i to X_IR
6.  Else
7.          Add X_i to X_CR
8.
9.  for each j ∈C
10. Mode_j = Calc_Mode(X_D, j)
11.
12. for each i∈ INR
13.   for each j ∈ C
14. If X_ij is missing then
15. X_ij = Mode_j
16. Add X_i to X_IM
17.
18. for each j ∈C
19. {
20.  If j ∈Ω then
21. {
22.   X_TR: Records in X_CR with j as the output variable
23.
24.   For each X_i∈ X_IR
25. If X_ij is missing then
26.         Add corresponding X_i from X_IM to X_1
27.
28. X_PR: Records in X_1 with j as the output variable and output contains the missing value
    "?"
29.
30.  Convert X_TR and X_PR into an indicator matrix
31.
32.  Train PNN using X_TR with SF as α
33.  Now employ the PNN trained in step 32 to predict the missing values in X_PR
34.  }
35. }
36. }
37.
38. Calc_Mode(X_D, j)
39. {
40.  Calculates the mode of the attribute j by ignoring the missing values in j.
41. }
```

**Fig. 3.** Algorithm of the proposed imputation technique.

**Table 1**
Attributes and records in each dataset.

| Dataset | Records | Attributes |
|---|---|---|
| Car Evaluation | 1728 | 7 |
| Mushroom | 8124 | 22 |
| Solar Flare | 324 | 13 |
| Spect Heart | 267 | 23 |

**Table 2**
Categorical variables and the corresponding nominal numbers for Car Evaluation dataset.

| Variable | Nominal number for each category |
|---|---|
| Buying | V-High ⇔ 4, High ⇔ 3, Med ⇔2, Low ⇔1 |
| Maint | V-High ⇔ 4, High ⇔ 3, Med ⇔2, Low ⇔1 |
| Doors | 2⇔2, 3⇔3, 4⇔4, 5-MORE ⇔5 |
| Persons | Less than 4 ⇔1, 4-More ⇔ 2 |
| Lug_Boot | Small⇔1, Med ⇔ 2, Big⇔3 |
| Safety | Low ⇔ 1, Med ⇔ 2, High ⇔ 3 |

### 4.4. Experimental design

The experimental design involves three kinds of experiments involving three percentages of missing data: 5%, 10% and 15%, which are quite realistic for any real world dataset. All the records having missing values are removed from the Mushroom dataset and the resulting dataset is used for experimenting. Since none of the remaining datasets has missing values, we conducted the experiments by deleting some values from the original datasets randomly. The same method is employed for the Mushroom dataset. This is essential in order to assess the accuracy of the proposed method of imputation. Every dataset is divided into 10 folds and 9 folds are used for training while the tenth one is left out for testing. From $i^{th}$ test fold, every time, we deleted approximately (subject to getting an integral number of missing values) 5%, 10% and 15% of the values (cells) randomly. These variants of experiments are conducted over 10 folds of 4 datasets with 7 imputation techniques viz., PNN, RF, DT, Naïve Bayes, HD, K-NN and mode. Therefore, for each of the given datasets, 210 different models are constructed. The accuracy of each experiment is measured by the percentage of correct values on the test set.

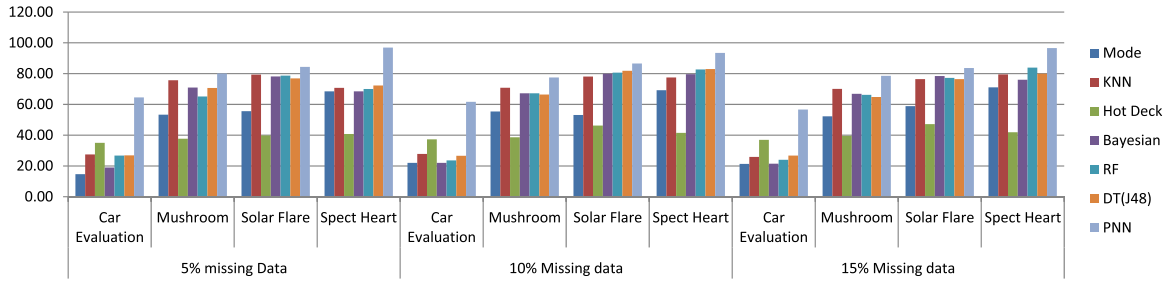


**Fig. 4.** PCP for different percentages of missing values and for different techniques for the datasets used for the experiment.

Initially the missing values are imputed with the mode of the corresponding attribute. These imputations serve as good initial approximations. Subsequently, we employed all the methods in order to refine these approximate values. We predict the missing values in one attribute, by considering it as the target variable and the rest as independent variables. While predicting, we use the initial approximations of the variables as part of test set for predicting the target variable if we have more than one missing value in a record. The experiments are carried out using 10 fold cross validation framework for all the datasets.

## 5. Results and discussions

We developed the code for PNN in Java in Windows environment on a laptop with 2 GB RAM. Weka tool is used for imputation with RF, DT and Naïve Bayes. We measured the performance of the proposed method by percentage of values that are predicted correctly for the missing values.

$$\text{Percentage of Correct Predictions(PCP)} = 100 * \frac{\text{Number of Correct Predictions(NCP)}}{\text{Total number of predictions}}$$

PNN is governed by a parameter called smoothing factor (SF). For a dataset with missing values in 'p' attributes, p different training and test sets are created. However, it is observed that a single SF is not sufficient to impute missing values in the p test sets obtained from p different attributes. Several experiments are conducted and the p different SF that yield the best prediction accuracy for p different attributes are collected. The overall prediction accuracy is calculated from the predictions obtained from p attributes and is tabulated. We performed 10 fold cross validation throughout the experiments and the average Percentage of Correct Predictions (PCP) over 10 folds are presented in Table 3 and Fig. 4. The range of smoothing factor (SF) for the datasets employed in the experimentation is presented in Tables 4–7. The pattern layer of PNN contains a pattern node for each training record. Each

**Table 4**
Range of SF for attributes of Mushroom dataset.

| Attribute | Range | | Attribute | Range | |
|---|---|---|---|---|---|
| **0** | 0.0001 | 0.0041 | **11** | 0.0031 | 0.0441 |
| **1** | 0.0051 | 0.0071 | **12** | 0.0011 | 0.0081 |
| **2** | 0.0001 | 0.0131 | **13** | 0.0011 | 0.0451 |
| **3** | 0.0041 | 0.0131 | **14** | 0.0011 | 0.0741 |
| **4** | 0.0011 | 0.2451 | **15** | 0.0001 | 0.9999 |
| **5** | 0.0011 | 0.0241 | **16** | 0.0011 | 0.0151 |
| **6** | 0.0011 | 0.0521 | **17** | 0.0011 | 0.0141 |
| **7** | 0.0011 | 0.0101 | **18** | 0.0011 | 0.0271 |
| **8** | 0.0001 | 0.2401 | **19** | 0.0071 | 0.0091 |
| **9** | 0.0041 | 0.0441 | **20** | 0.0031 | 0.0141 |
| **10** | 0.0011 | 0.6431 | | | |

**Table 5**
Range of SF for attributes of Car Evaluation dataset.

| Attribute | Range | |
|---|---|---|
| **1** | 0.0001 | 0.5000 |
| **2** | 0.0001 | 0.7000 |
| **3** | 0.0011 | 0.9999 |
| **4** | 0.0081 | 0.1291 |
| **5** | 0.0011 | 0.8161 |
| **6** | 0.2831 | 0.9999 |

**Table 6**
Range of SF for attributes of Solar Flare dataset.

| Attribute | Range | | Attribute | Range | |
|---|---|---|---|---|---|
| **0** | 0.0051 | 0.0341 | **6** | 0.0011 | 0.1141 |
| **1** | 0.0001 | 0.0351 | **7** | 0.0011 | 0.0091 |
| **2** | 0.0011 | 0.0571 | **8** | 0.0001 | 0.0361 |
| **3** | 0.0011 | 0.0471 | **9** | 0.0011 | 0.0471 |
| **4** | 0.0011 | 0.0121 | **10** | 0.0001 | 0.0371 |
| **5** | 0.0001 | 0.0301 | **11** | 0.0001 | 0.0131 |

**Table 3**
Percentage of Correct Predictions (PCP) for various techniques used for experimentation.

| | Dataset | Mode | KNN | Hot Deck | Bayesian | RF | DT (J48) | PNN |
|---|---|---|---|---|---|---|---|---|
| **5% missing data** | **Car Evaluation** | 14.65 | 27.44 | 35.00 | 18.95 | 26.74 | 26.86 | **64.53** |
| | **Mushroom** | 53.36 | 75.73 | 37.70 | 70.95 | 65.12 | 70.71 | **80.09** |
| | **Solar Flare** | 55.63 | 79.38 | 39.85 | 78.13 | 78.75 | 76.88 | **84.38** |
| | **Spect Heart** | 68.46 | 70.72 | 40.77 | 68.46 | 70.00 | 72.31 | **96.92** |
| **10% Missing data** | **Car Evaluation** | 22.00 | 27.84 | 37.32 | 21.97 | 23.60 | 26.57 | **61.68** |
| | **Mushroom** | 55.44 | 70.82 | 38.61 | 67.21 | 67.21 | 66.38 | **77.46** |
| | **Solar Flare** | 53.13 | 78.12 | 46.25 | 79.96 | 80.62 | 81.87 | **86.57** |
| | **Spect Heart** | 69.23 | 77.55 | 41.53 | 79.57 | 82.68 | 83.01 | **93.46** |
| **15% Missing data** | **Car Evaluation** | 21.34 | 25.88 | 36.96 | 21.46 | 24.00 | 26.77 | **56.65** |
| | **Mushroom** | 52.26 | 70.13 | 39.57 | 66.87 | 66.22 | 64.82 | **78.54** |
| | **Solar Flare** | 58.88 | 76.40 | 47.20 | 78.40 | 77.20 | 76.40 | **83.60** |
| | **Spect Heart** | 71.04 | 79.46 | 41.91 | 76.05 | 83.87 | 80.00 | **96.58** |

**Table 7**
Range of SF for attributes of Spect Heart dataset.

| Attribute | Range | | Attribute | Range | |
|---|---|---|---|---|---|
| **0** | 0.0011 | 0.0331 | **11** | 0.0011 | 0.0821 |
| **1** | 0.0001 | 0.2011 | **12** | 0.0001 | 0.1011 |
| **2** | 0.0001 | 0.1011 | **13** | 0.0011 | 0.0641 |
| **3** | 0.0001 | 0.1501 | **14** | 0.0001 | 0.1021 |
| **4** | 0.0001 | 0.1021 | **15** | 0.0001 | 0.0631 |
| **5** | 0.0001 | 0.1361 | **16** | 0.0001 | 0.1151 |
| **6** | 0.0001 | 0.0611 | **17** | 0.0001 | 0.0481 |
| **7** | 0.0001 | 0.0100 | **18** | 0.0001 | 0.0451 |
| **8** | 0.0001 | 0.1521 | **19** | 0.0001 | 0.0691 |
| **9** | 0.0001 | 0.0911 | **20** | 0.0001 | 0.1171 |
| **10** | 0.0001 | 0.1411 | **21** | 0.0001 | 0.1411 |

pattern node forms a product of the weight vector and the given training record for classification. The number of nodes in summation layer is same as the number of classes in the output attribute.

The proposed method for categorical imputation is compared with 3 statistical and 3 machine learning techniques for data imputation. The statistical techniques include mode, K-Nearest Neighbors (K-NN) and Hot Deck (HD) imputation. Naïve Bayes, Random Forest (RF) and J48 (DT) are the machine learning techniques employed for imputation. From the results it is observed that the proposed PNN based imputation technique preceded by mode imputation yields a better prediction for missing values than the other imputation techniques. The datasets used for the experimentation have a considerable number of attributes and records as seen from Table 1. The Spect Heart dataset contains two categories for each categorical attribute and a minimal number of records. So, all the methods used for the experimentation except Hot Deck (HD) imputation has yielded a good percentage of correct predictions. But, even then the proposed mode+PNN imputation technique outperformed other imputation techniques by yielding classification accuracy above 93% for the three percentage of missing values as depicted in Fig. 4. The distribution of records for each category of a categorical variable is not even for the datasets mentioned above. Consequently, the model doesn't have sufficient records to learn the characteristics of the data. So, the PCP for Car Evaluation, Mushroom and Solar Flare datasets is not so good compared to that of the Spect Heart dataset. The distribution of records for various categories in each categorical variable for Solar Flare dataset is depictedin Fig. 5.

The mushroom dataset contains 22 categorical attributes and the number of categories in each attribute is more when compared to other datasets. For the Mushroom dataset the proposed imputation technique yields a better PCP compared to other techniques. The proposed imputation technique yields a PCP of 80.09, 77.46 and 78.54 for 5%, 10% and 15% of missing data respectively. For 5% of missing data a PCP of 53.36, 75.73, 37.70, 70.95, 65.12 and 70.71 is obtained by Mode, K-NN, HD, Naive Bayes, RF and DT imputation techniques respectively. For 10% and 15% of missing data a PCP of 70.82 and 70.13 is obtained by using K-NN. From the results, it is observed that the number of correct predictions decreases for increasing percentage of missing values for all the imputation techniques except the proposed imputation technique. The proposed imputation technique yields a higher number of correct predictions for all the percentage of missing values. From the results it is observed that the proposed imputation technique performs better even when the dataset is large and when the dataset has a large number of categories in each attribute.

For the Solar Flare dataset, all the methods used for the experimentation except HD have yielded a good percentage of correct predictions. The proposed imputation technique outperformed the other techniques by yielding classification accuracy above 83% for the three percentage of missing values. For the Car Evaluation dataset, the proposed imputation technique yields a better percentage of correct predictions than all the other techniques. However, the percentage of the correct predictions is very low compared to that of other datasets. To investigate the reason for the poor classification by all the imputation techniques on Car Evaluation dataset, the association between the attributes of the datasets utilized for the experiment is calculated by employing the Cramer's-V statistic. For the Car Evaluation dataset it is observed that the association or correlation between the attributes is weak. Hence, all the imputation techniques utilized for the experiment yielded poor predictions compared to other datasets. However, the proposed imputation technique yielded a good percentage of good predictions despite the weak association among the attributes of the Car Evaluation dataset.

From the results it is observed that the proposed imputation technique outperforms all the other imputation techniques. Also, the proposed imputation technique yields better % of correct predictions when the percentage of missing data is high as compared to other imputation techniques. The proposed imputation technique performs better especially when the percentage of missing values is high and when the correlation between the attributes is weak. This is the significant outcome of the study.

## 6. Conclusions

We proposed a novel, soft computing technique for categorical data imputation. The proposed technique employs PNN preceded by mode for imputation. The proposed imputation method is tested on four benchmark datasets in the framework of 10 fold cross validation. The performance of the proposed imputation technique is compared with that of RF, DT, Naïve bayes, HD, K-NN and mode imputation. The results indicate that the proposed imputation technique yields better Percentage of Correct Predictions (PCP). Therefore, we conclude that, we can use the proposed

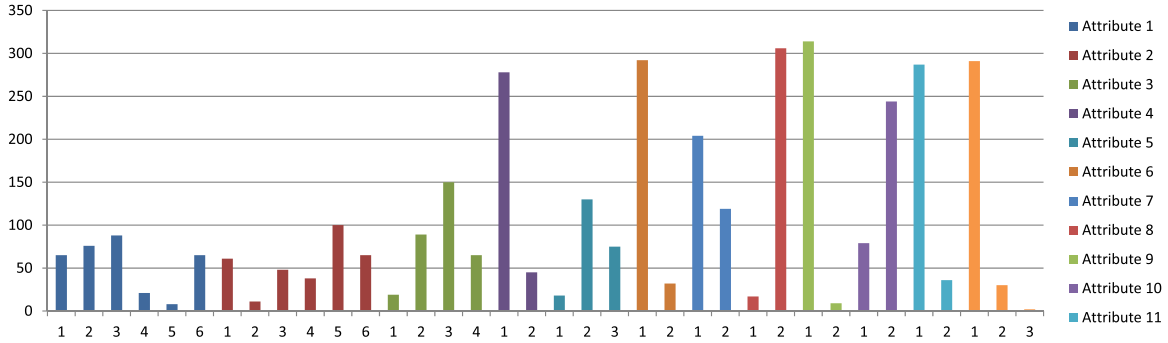


**Fig. 5.** Distribution of records for various categories in each categorical variable for Solar Flare dataset.

approach as a viable alternative to the extant methods for categorical data imputation. In particular, this method is useful for a dataset with a records having more than one missing value.

## References


[1] P.L. Roth, F.S. Switzer, D.M. Switzer, Missing data in multiple item scales: a Monte Carlo analysis of missing data techniques, Organ. Res. Methods 2 (3) (1999) 211–232.
[2] K.J. Nishanth, V. Ravi, A computational intelligence based online data imputation method: an application for banking, J. Inf. Process. Syst. 9 (4) (2013) 633–650.
[3] S. Henley, The problem of missing data in geoscience databases, Comput. Geosci. 32 (9) (2006) 1368–1377.
[4] Q. Song, M. Shepperd, A new imputation method for small software project data sets, J. Syst. Softw. 80 (2007) 51–62.
[5] D.H. Schoellhamer, Singular spectrum analysis for time series with missing data, Geophys. Res. Lett. 28 (16) (2001) 3187–3190.
[6] P.M.T. Broersen, S.D. Waele, R. Bos, Autoregressive spectral analysis when observations are missing, Automatica 40 (2004) 1495–1504.
[7] M. Cooke, P. Green, M. Crawford, Handling missing data in speech recognition, in: Third International Conference on Spoken Language Process (ICSLP), Yokohama, Japan, 1994, pp. 1555–1558.
[8] O. Troyanskaya, M. Cantor, O. Alter, G. Sherlock, P. Brown, D. Botstein, R. Tibshirani, T. Hastie, R. Altman, Missing value estimation methods for DNA microarrays, Bioinformatics 17 (6) (2001) 520–525.
[9] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, second edition, Wiley, New York, 2002.
[10] P.J. García-Laencina, J.L. Sancho-Gómez, A.R. Figueiras-Vidal, Pattern classification with missing data: a review, Neural Comput. Appl. 19 (2) (2010) 263–282.
[11] P. Merlin, A. Sorjamaa, B. Maillet, A. Lendasse, A. X-SOM and L-SOM: a double classification approach for missing value imputation, Neurocomputing 73 (7–9) (2010) 1103–1108.
[12] G.E. Batista, M.C. Monard, A study of K-nearest neighbor as an imputation method, Hybrid. Intell. Syst. 251–260 (2002) 48.
[13] A. Gupta, M.S. Lam, Estimating missing values using neural networks, J. Oper. Res. Soc. 47 (2) (1996) 229–238.
[14] B. Gabrys, Neuro-Fuzzy approach to processing inputs with missing values in pattern recognition problems, Int. J. Approx. Reason. 30 (3) (2002) 149–179.
[15] M. Abdella, T. Marwala, The Use of Genetic Algorithms and Neural Networks to Approximate Missing Data in Database, IEEE 3rd International Conference on Computational Cybernetics (ICCC), Mauritius, 2005, pp. 207–212.
[16] A. Agresti, Categorical Data Analysis, second edition, John Wiley and Sons Inc, Hoboken, NewJersey, 2002.
[17] G. Chen, T. Astebro, How to deal with missing categorical data: test of sample Bayesian method, Organ. Res. Methods 6 (3) (2003) 309–327.
[18] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1990) 109–118.
[19] N.K. Kasabov, Q. Song, DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction, IEEE Trans. Fuzzy Syst. 10 (2) (2002) 144–154.
[20] W.S. Desabro, P.E. Green, J.D. Carroll, Missing data in product-concept testing, Decis. Sci. 17 (2) (1986) 163–185.
[21] C.P. West, J.D. Dawson, Complete imputation of repeated categorical data: one sample applications, Stat. Med. 21 (2) (2002) 203–217.
[22] W. Coutinho, Ton de Waal, N. Shlomo, Calibrated hot-deck donor imputation subject to edit restrictions, J. Off. Stat. 29 (2) (2013) 1–23.
[23] M. Cartwright, M. Shepperd, Q. Song, Dealing with Missing Software Project Data, 9th IEEE International Software Metrics Symposium (METRICS 2003), 3-5 September 2003, IEEE Computer Society, Sydney, Australia 2003, pp. 154–165.
[24] Q. Song, M. Shepperd, M. Cartwright, A short note of safest default missingness mechanism assumptions, Empir. Softw. Eng. 10 (2) (2005) 235–243.
[25] P. Jonsson, C. Wohlin, An Evaluation of k-nearest neighbor imputation using likert data, 10th IEEE International Software Metrics Symposium (METRICS 2004), 11–17 September 2004, Chicago, IL, USA. IEEE Computer Society 2004, pp.108–118.
[26] S.V. Buuren, J.L.A. Van Rijckevorsel, Imputation of missing categorical data by maximizing internal consistency, Psycometrika 57 (4) (1992) 567–580.
[27] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman & Hall, London, 1997.
[28] P. Sentas, L. Angelis, Categorical missing data imputation for software cost estimation by multinominal logistic regression, J. Syst. Softw. 79 (3) (2006) 404–414.
[29] P. Rey-de-Castillo, J. Cardenosa, Fuzzy min–max neural networks for categorical data: application to missing data imputation, Neural Comput. Appl. 21 (6) (2011) 1349–1362.
[30] P.A. Ferrari, P. Annoni, A. Barbiero, G. Manzi, An imputation method for categorical variables with applications to non-linear principal component analysis, Comput. Stat. Data Anal. 55 (7) (2011) 2410–2420.
[31] K. Wang, M.L. Ng, M. Tang, M. Tan, G.L. Tian, Grouped Dirichlet distribution: a new tool for incomplete categorical data analysis, J. Multivar. Anal. 99 (3) (2008) 490–509.
[32] A. Farhangfar, L. Kurgan, Jennifer Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognit. 41 (12) (2008) 3692–3705.
[33] M. Gebregziabher, S.M. DeSantis, Latent class based multiple imputation approach for missing categorical data, J. Stat. Plan. Inference 140 (11) (2010) 3252–3262, 2010.
[34] S.G. Baker, C.W. Ko, B. Graubard, A sensitivity analysis for non randomly missing categorical data arising from a national health disability survey, Biostatistics 4 (1) (2003) 41–56.
[35] J. Kaiser, Algorithm for missing values imputation in categorical data with use of association rules, Int. J. Recent Trends Eng. Technol. 6 (1) (2011) 111–114.
[36] Z. Geng, Y.B. He, X.L. Wang, Q. Zhao, Bayesian method for learning graphical models with incompletely categorical data, Comput. Stat. Data Anal. 44 (1–2) (2003) 175–192.
[37] J.K. Vermunt, J.R. Van Ginkel, L.A. Van Der Ark, K. Sijtsm, Multiple Imputation of incomplete categorical data using latent class analysis, Sociol. Methodol. 38 (2008) 369–397.
[38] Y. Thibaudeau, Model explicit item imputation for demographic categories, Surv. Methodol. 28 (2) (2002) 5–12.
[39] S. Wu, X. Feng, Y. Han, Q. Wang, Missing categorical data imputation approach based on similarity, IEEE International Conference on Systems, Man and Cybernetics – SMC, 14–17 October 2012, Seoul, Korea (South), pp. 2827–2832.
[40] F. Lobato, C. Sales, I. Araujo, V. Tadaiesky, L. Dias, L. Ramos, A. Santana, Multi-objective genetic algorithm for missing data imputation, Pattern Recognit. Lett. 68 (1) (2015) 126–131.
[41] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1990) 109–118.


**Mr. Kancherla Jonah Nishanth** is currently a Senior Technology Manager at Andhra Bank, Hyderabad. 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, data imputation and machine learning. He has two publications to his credit in reputed International Journals.

**Prof. Vadlamani Ravi** is a Professor at the Institute for Development and Research in Banking Technology, Hyderabad since June 2014. 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 (1991) and M.Sc. (Statistics & Operations Research) from IIT, Bombay (1987). At IDRBT, he spearheads the Center of Excellence in Analytics, first-of-its-kind in India and evangelizes Analytical CRM and Non-CRM related analytics in a big way by conducting customized training programs for bankers on OCRM & ACRM; Data Warehousing, Data and Text Mining, Big Data Analytics, Fraud Analytics, Risk Analytics, Social Media Analytics, Credit Recovery Analytics, Business Analytics and conducting POC for banks etc. He has 176 papers to his credit with the break-up of 74 papers in refereed International Journals, 6 papers in refereed National Journals, 77 papers in refereed International Conferences and 3 papers in refereed National Conferences and 16 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, Engineering Application of Artificial Intelligence, Fuzzy Sets and Systems, IEEE Transactions on Fuzzy Systems, IEEE Transactions on Reliability, Information Sciences, Journal of Systems and Software, Knowledge Based Systems, Neurocomputing, IJUFKS, IJCIA, IJAEC, IJDMMM, IJIDS, IJDATS, IJISSS, IJECRM, IJISSC, IJCIR, IJCISIM, IJBIC, JIPS, Computers and Chemical Engineering, Canadian Geotechnical Journal, Biochemical Engineering Journal, Computers in Biology and Medicine, Applied Biochemistry and Biotechnology, Bioinformation, 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 and the Proceedings of 5th Fuzzy and Neuro Computing Conference, 2016 held at Hyderabad, India. Some of his research papers are listed in Top 25 Hottest Articles by Elsevier and World Scientific. He has an H-index of 28 and more than 3178 citations for his papers (http://scholar.google.co.in/). His profile was among the Top 10% Most Viewed Profiles in LiknkedIn in 2012. He is recognized as a Ph.D. supervisor at Department of Computer and Information Sciences, University of Hyderabad and Department of Computer Sciences, Berhampur University, Orissa. He is an invited member in Marquis Who's Who in the World, USA in 2009, 2015. He is also an invited member in 2000 Outstanding Intellectuals of the 21st Century 2009/2010- published by International Biographical Center,

Cambridge, England, He is an Invited Member of "Top 100 Educators in 2009" published by International Biographical Center, Cambridge, England. So far, 3 Ph.D. students graduated under his supervision and 5 more are currently working towards Ph.D. So far, he advised more than 50 M.Tech./MCA/M.Sc projects and 20 Summer Interns from various IITs. He currently supervises 3 M.Tech students. He is on the IT Advisory Committee of Canara Bank for their DWH and CRM project; IT Advisor for Indian Bank for their DWH and CRM project and Principal Consultant for Bank of India for their CRM project; Expert Committee Member for IRDA for their Business Analytics and Fraud Analytics projects. He is a referee for 40 International Journals of repute. Moreover, he is a member of the Editorial Review Board for the "International Journal of Information Systems in Service Sector" published by IGI Global, USA; "International Journal of Data Analysis Techniques and Strategies" published by Inderscience, Switzerland; International Journal of Information and Decision Sciences (IJIDS), Inderscience, Switzerland; International Journal of Strategic Decision Sciences (IJSDS), IGI Global, USA and International Journal of Information Technology Project Management (IJITPM), IGI Global, USA. International Journal of Data Science (IJDS), Inderscience, Switzerland; Editorial Board Member for Book Series in Banking, Inderscience Switzerland, He is on the PC for some International Conferences and chaired many sessions in International Conferences in India and abroad. His research interests include Fuzzy Computing, Neuro Computing, Soft Computing, Evolutionary Computing, Data Mining, Text Mining, Web Mining, Big Data Analytics, Privacy Preserving Data Mining, Global/Multi-Criteria/Combinatorial Optimization, Bankruptcy Prediction, Risk Measurement, Customer Relationship Management (CRM), Fraud Analytics, Sentiment Analysis, Social Media Analytics, Big Data Analytics, Churn Prediction in Banks and firms and Asset Liability Management through Optimization. In a career spanning 28 years, Dr. Ravi worked in several cross-disciplinary areas such as Financial Engineering, Software Engineering, Reliability Engineering, Chemical Engineering, Environmental Engineering, Chemistry, Medical Entomology, Bioinformatics and Geotechnical Engineering. At IDRBT, he held various administrative positions such as Coordinator, IDRBT-Industry Relations (2005-06), M. Tech (IT) Coordinator, (2006–2009), Convener, IDRBT Working Group on CRM (2010-11); Ph.D.Coordinator (2014–2016). As the convener, IDRBT Working Group on CRM, he coauthored a Handbook on Holistic CRM and Analytics (http://www.idrbt.ac.in/PDFs/Holistic%20CRM%20Booklet_Nov2011.pdf), where a new framework for CRM, best practices and new organization structures apart from HR issues for Indian banking industry are all suggested. He has 28 years of research and 15 years of teaching experience. He designed and developed a number of courses in Singapore and India at M. Tech level in Soft Computing, Data Warehousing and Data Mining, Fuzzy Computing, Neuro Computing, Quantitative Methods in Finance, Soft Computing in Finance etc. Further, he designed and developed a number of short courses for Executive Development Programs (EDPs) in the form of 2-week long CRM for executives, Data Mining, Big Data and its relevance to Banking, Fraud Analytics etc. He conducted ACRM proof of the concept (POC) for 14 banks on their real data. He established excellent research collaborations with University of Hong Kong, University of Ghent, Belgium, IISc., Bangalore and IIT Kanpur. He co-ordinated Two International EDPs in University of Ghent, Belgium on ACRM to banking executives jointly with Prof Dr Dirk Van den Poel, University of Ghent, Belgium in 2011 and 2012. As part of academic outreach, he is a Guest Speaker in IIM Kolkata's PGP program and an invited Resource Person in various National Workshops and Faculty development programs on Soft Computing, Data Mining, Big Data funded by AICTE and organized by SRM University, JNTU, UoH and some Engineering Colleges in India. He is an invited Panel Member/Chair in several National forums like IBA on Big Data Analytics in Banking in India. He is a member of task force set up to suggest changes to curriculum by NSDC and administered by NASSCOM o to develop Data Scientists. Further, he contributed to the Roadmap for Big Data in India developed by DST. He is an External examiner for Ph.D. in Auckland University of Technology, New Zealand and Christ University, India. Further, he is an External Expert to review Research Project Proposals on Analytics and ACRM submitted by the Belgian Academics to Belgian Government for funding. He was the Associate Professor at IDRBT between February 2010-June 2014. 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 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. Earlier, he worked as Assistant Director (Scientist E1) from 1996–2002 and Scientist C from 1993 to 1996 respectively at the Indian Institute of Chemical Technology (IICT), Hyderabad. He was deputed to RWTH Aachen (Aachen University of Technology) Germany under the DAAD Long Term Fellowship to carry out advanced research during 1997–1999. He earlier worked as Scientist B and Scientist C at the Central Building Research Institute, Roorkee (1988–1993) and was listed as an expert in Soft Computing by TIFAC, Government of India.