Global Journal of Technology & Optimization

Darryl, Global J Technol Optim 2016, 7:6
DOI: 10.4172/2229-8711.S1:113

Research Open Access

Missing Value Imputation Using Stratified Supervised Learning for Cardiovascular Data

Darryl ND[1*] **and Rahman MM**[2]

[1]*Director of Research, Computer Science, University of Hull, United Kingdom*

[2]*Manufacturing Technology Centre, United Knigdom*

[*]**Corresponding author**: Darryl ND, Director of Research, Computer Science, University of Hull, United Kingdom, Tel: +4401482466469; Fax: +4401482466666; E-mail: d.n.davis@hull.ac.uk

Received date: May 23**, 2016; Accepted date:** June 20**, 2016; Published date:** June 27**, 2016**

Abstract

Legacy (and current) medical datasets are rich source of information and knowledge. However, the use of most legacy medical datasets is beset with problems. One of the most often faced is the problem of missing data, often due to oversights in data capture or data entry procedures. Algorithms commonly used in the analysis of data often depend on a complete data set. Missing value imputation offers a solution to this problem. This may result in the generation of synthetic data, with artificially induced missing values, but simply removing the incomplete data records often produces the best classifier results. With legacy data, simply removing the records from the original datasets can significantly reduce the data volume and often affect the class balance of the dataset. A suitable method for missing value imputation is very much needed to produce good quality datasets for better analysing data resulting from clinical trials. This paper proposes a framework for missing value imputation using stratified machine learning methods. We explore machine learning technique to predict missing value for incomplete clinical (cardiovascular) data, with experiments comparing this with other standard methods. Two machine learning (classifier) algorithms, fuzzy unordered rule induction algorithm and decision tree, plus other machine learning algorithms (for comparison purposes) are used to train on complete data and subsequently predict missing values for incomplete data. The complete datasets are classified using decision tree, neural network, K-NN and K-Mean clustering. The classification performances are evaluated using sensitivity, specificity, accuracy, positive predictive value and negative predictive value. The results show that final classifier performance can be significantly improved for all class labels when stratification was used with fuzzy unordered rule induction algorithm to predict missing attribute values.

Keywords Data mining; Missing value; Imputation; FURIA; Classifier

Introduction

Legacy medical datasets are rich source of information and knowledge, and there is a growing trend with research funders expecting the data resulting from clinical trials to be used beyond the originating study. However, real-life data sets are often found to be incomplete. This is true for both legacy and current, in use, datasets. Causes for values to be missing vary; ranging from oversights in data capture or data entry procedures to systematic flaws in the studies that led to the data being generated. Often the cause of missing values is due to legacy data being extended with further trials where the information profile being captured has changed. Missing attribute values is already been identified as an important issue in data mining and analytics [1]. In medical data mining and analysis missing values has become a challenging issue, predominantly as legacy data can be a valuable source of information and knowledge. In many clinical trials, the medical report pro-forma allow some attributes to be left blank, because they are inappropriate for some cases or the person providing the information feels that it is not appropriate to record the values of some attributes [2].

According to Roderick and Donald [3] missing data can be classified in to two ways. Data is termed missing completely at random (MCAR) when the response indicator variables R, are independent of the data variables X and the latent variables Z. The MCAR condition can be briefly expressed by P (R|X, Z, μ) = P (R|μ). The second category of missing data is called missing at random or MAR. The MAR condition is often written as P (R = r|X = x, Z = z, μ) = P (R = r|X ° = x°) for all x^{μ}, z and μ [4].

Generally, methods to handle missing values belong either to sequential methods like leastwise deletion, assigning most common values for categorical attributes, arithmetic mean or median for the numeric attribute or parallel methods where algorithms are used to predict missing attribute values [5]. There are some reasons for which leastwise deletion is considered to be a good method [3], but a number of works [2, 3, 6] have shown that the application of these methods on the incomplete data can corrupt the construal of the data and mislead the subsequent analysis through the introduction of bias.

Several techniques for missing value imputation are proposed by researchers; most of the techniques are single imputation approaches [7]. The most commonly used missing value imputation techniques are deleting cases, mean value imputation and other statistical methods [7]. In recent years, research has explored machine learning techniques as a method for missing values imputation; artificial neural network (ANN), self-organising maps (SOM), decision tree and k-nearest neighbors (K-NN) were used as missing value imputation methods in many different domains [6, 8-15]. In many cases machine learning

methods like ANN, SOM, K-NN and decisions tree have been found to perform better than the traditional statistical methods [6, 16].

Machine learning methods can be used for predicting missing values; for example by using rule induction algorithm in which rules are induced from the original complete data set, with missing attribute values ignored. The decision tree can be produced by splitting cases with missing attribute values into fractions and adding these fractions to new case subsets [5]. Other methods of handling missing attribute values were presented in [17]. Jerez et al. [6] presented comparison results of missing data imputation using statistical and machine learning methods in a real breast cancer problem. They used imputation methods based on statistical techniques, e.g., mean, hot-decking and multiple imputations, and machine learning techniques, e.g., multi-layer perceptron (MLP), SOM and K-NN and applied them to the cancer data. The results were then compared to those obtained from the list wise deletion (LD) imputation method. K-NN has been used by many researchers for imputing missing value [18, 19]. Every time a missing value is found in a current instance, K-NN computes the K nearest neighbours and a value from them is imputed. For categorical values, the most common value among all (k) neighbours is taken, and for numerical values, the average value is used [19]. Gajawada and Toshniwal [18] proposed a modified version of imputing missing value with K-NN. Here, the dataset is divided into two sets records with missing value and records without missing value. K-Means clustering is applied to the complete instances set to obtain clusters of complete instances. This was then used to impute the missing values in the incomplete dataset.

In most cases highlighted above, the machine learning based missing value imputation found to be better than conventional statistical methods. However, none of the research considered the class label as of factor that might affect the learning from pattern of the complete dataset. Our contention is that a data pattern of one class is not similar to other class label records, and so stratified learning may give better results.

In this paper we examine stratified supervised learning for predicting missing values. In our proposed approach we used FURIA, fuzzy unordered rules induction algorithm [20], with stratification as a missing values imputation for real life incomplete cardiovascular datasets. The results are compared with some other non-stratified machine learning based missing value imputation methods using decision tree, SVM, K-NN, and conventional statistical mean-mode imputation methods.

Overview of Furia

Fuzzy Unordered Rule Induction Algorithm (FURIA) is a novel rule-based classification method, which is a modification and extension of the state-of-the-art RIPPER rule learner algorithm. The main difference between FURIA and RIPPER is that FURIA learns fuzzy rules and unordered rule sets instead of conventional rules and rule lists. Moreover, FURIA uses a rule stretching method to deal with uncovered examples [20]. A fuzzy interval of that kind is specified by four parameters and will be written:

$$IF = (\varphi^{s,L}, \varphi^{c,L}, \varphi^{c,U}, \varphi^{s,U}):$$

$$I^F(v) \stackrel{def}{=} \begin{cases} 1 \\ \dfrac{v - \varphi^{s,L}}{\varphi^{c,L} - \varphi^{s,L}} \\ \dfrac{\varphi^{s,U} - v}{\varphi^{s,U} - \varphi^{c,U}} \\ 0 \end{cases}$$

$$\begin{array}{l} \varphi^{c,L} \le v \le \varphi^{c,U} \\ \varphi^{s,L} < v < \varphi^{c,L} \\ \varphi^{c,L} < v < \varphi^{s,U} \\ else \end{array} \quad (1)$$

Where $\varphi^{c,L}$ and $\varphi^{c,U}$ are, respectively, the lower and upper bound of the core (elements with membership 1) of the fuzzy set; likewise, $\varphi^{s,L}$ and $\varphi^{s,U}$ are, respectively, the lower and upper bound of the support (elements with membership >0).

For an instance x = (x1……xn) the degree of the fuzzy membership can be found using the formula [20]:

$$\mu_r F(x) = \prod_{i\,=\,1....k} i_i^F(x_i) \qquad (2)$$

For fuzzification of a single antecedent only relevant training data is considered and data are partitioned into two subsets and rule purity is used to measure the quality of the fuzzification [20]:

$$D_T^i = \left\{ x = (x_1 ... x_k) \in D_T^i \,\middle|\, I_j^F(x_j) > 0 for\ all\ j \ne i \right\} \subseteq D_T \qquad (3)$$

$$Pur = \frac{p_i}{p_i + n_i} \qquad (4)$$

Where

$$p_i \stackrel{def}{=} \sum_{x \in D_{T+}^i} \mu_{A_i}(A)$$

$$n_i \stackrel{def}{=} \sum_{x \in D_{T-}^i} \mu_{A_i}(A)$$

The fuzzy rules $r_1^{(j)} .. r_k^{(j)}$ have learned for the class λ_j, the support of this class is defined by [20]:

$$S_j(x) \stackrel{def}{=} \sum_{i\,=\,1..k} \mu_{r_i^{(j)}}(x) \,.\, CF(r_i^{(j)}) \qquad (5)$$

where, the certainty factor of the rule is defined as

$$CF(r_i^{(j)}) = \frac{2\dfrac{\left|D_T^{(j)}\right|}{D_T} + \sum_{x \in D_T^{(j)}} \mu_{r_i^{(j)}}(x)}{2 + \sum_{x \in D_T^{(j)}} \mu_{r_i^{(j)}}(x)} \qquad (6)$$

Fuzzy rule are generated by FURIA by following two steps:

(1) For every single class λc a rule set is learnt, using a one-versus-all decomposition. The RIPPER algorithm is used, which consists of two fundamental steps (building and the optimization phase) described in Sun and Xu [21].

(2) Rules from above step are fuzzified to obtain fuzzy rules. Each rule is fuzzified remembering the same structure as the non fuzzified rule just replacing original intervals in the antecedent with fuzzy intervals (complete procedure is described in Hühn and Hüllermeier [20])

More use of FURIA in different areas of data mining can be found in [20, 22, 23]. Stratified machine learning based missing value imputation

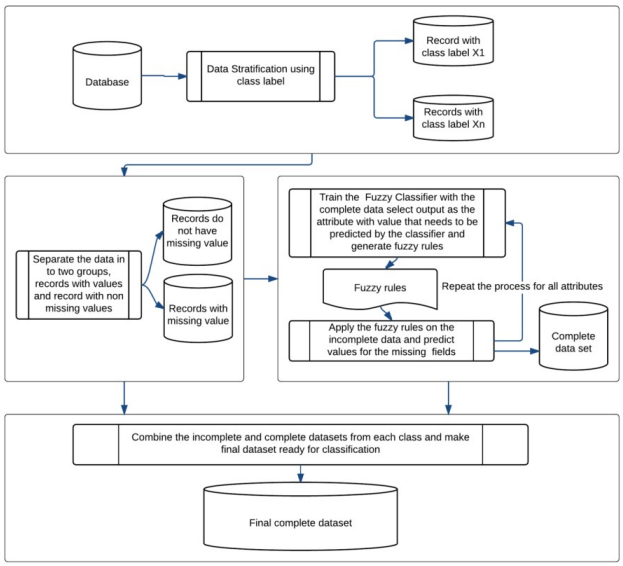

Figure 1: Flow diagram for the imputation process.

This research presents a new way of imputing missing value using machine learning methods. The original data set can be first stratified using the intended class label. It is then partitioned into groups of missing and non-missing; the records having missing values in their attributes are in one group and the records without any missing values are placed in a separate group. **Figure 1** depicts the flow for the imputation process. Below we explain this process in terms of using the FURIA fuzzy rule based classifier to find suitable values for imputation. The process is very similar when using other classifiers. The difference being the flow in the right center of **Figure 1** is modified according to the classifier used. The other classifiers used in the experiments are briefly described in section 6.

The fuzzy rule based classifier FURIA is trained with the complete data sets and optimum fuzzy rules are obtained. The rules are later applied to the incomplete data for predicting the missing attribute values. The process is repeated for the entire set of attributes that have missing values. At the end of training, this training dataset and the missing value imputed datasets are combined to make the complete data. The final dataset is then fed to the selected classifier for classification on the true outcome.

The stratified fuzzy rule based imputation scheme developed in this study can be described as follows:

(1) Given an incomplete data set X, Stratify data based on the class label (for two class problem xa and xb) (2) For all data records of each class do the following:

a. Separate the input vectors that do not contain any missing data from the ones that have missing values.

b. Train the FURIA Classifier with the complete data (having no missing value). Select the output as the attribute whose value needs to be predicted by the classifier for imputation and build up the model with classifiers' best accuracy. Obtain optimum fuzzy rules.

c. For each incomplete pattern apply the fuzzy rules to predict unknown value of the missing fields.

d. Repeat for all attributes with missing value.

Cardiovascular Data

Two data sources for cardiovascular patients are used: the Hull site of 498 patients and the Dundee site of 341 patients. The patients in the Hull site are described by 98 attributes. The patients in the Dundee site are described by 57 attributes. As a dataset, a combination from both sites is used. This gives a group of 823 instances (cardiovascular patients) classified into two levels of risk and described by 22 attributes. After the combination, 18 out of 22 attributes have missing values from 1% to 30%; and 613 out of 839 instances have 4% to 56% missing values in their describing attributes. All instances having 20% or more missing values and relating to live patients 30 days after an operation are removed. The data is described in full in Nguyen [24].

Data description

The description of instances and their summary is given in **Table 1**, showing the percentage of missing values for each attribute. This data is symptomatic of much legacy clinical data, in that it is flawed in data capture, with patient records coming from multiple trials and each data record cannot be replicated (for obvious reasons).

ASA grade is used to classify the patient into categorical values one, two, three or four according to the American Society of Anesthesiologists classification [25]. Value one means the patient is fit and well for her/his age. Value two means the patient's cardiovascular disease is mild, i.e. it does not hamper enjoyment of daily activities. Value three means the patient's cardiovascular disease is severe, i.e. it restricts the patient's daily activities. Value four means the patient's cardiovascular disease is life-threating [25].

Aspirin indicates if the patient takes aspirin. Blood loss represents the blood loss in surgery in milliliters. Coronary artery bypass surgery indicates if coronary artery bypass surgery is present. Carotid status indicates a patient's health status related to carotid arteries. Congestive cardiac failure indicates if heart failure has occurred and when it occurred. Diabetes indicates if and what kind of diabetes is present. Value impaired glucose tolerance means the patient is in a pre-diabetic state of dysglycemia that is associated with insulin resistance and increased risk of cardiovascular pathology [26]. Value Diet Rx pill indicates the patient takes Diet Rx pills. Duration is the duration of surgery in hours.

Age represents the age of the patient. Attribute Angina pectoris indicates if a particular angina pectoris is present. The value is set as none if there is no angina pectoris, other possible values are stable, controlled, uncontrolled. Attribute Arrhythmia indicates if a large and heterogeneous group of conditions in which there is abnormal electrical activity in the heart exists [27] the possible values for this attribute are none, a-fib $\geq$ 90, other, where a-fib $\geq$ 90 means atrial fibrillation is present for greater than 90 days .

ECG describes electrocardiography, i.e. a transthoracic (across the thorax or chest) interpretation of the electrical activity of the heart over a period of time. Several categorical values are used: normal, q waves, st-t waves, a-fib 60-90, a-fib ≤ 90, five ectopic, other abnormal rhythm, other. Value normal means there are no abnormalities in electrocardiography. Value q wave's means Q wave abnormalities are present. Value st-t waves means ST-T wave [28] abnormalities are present. Values a-fib 60 to 90 and a-fib ≥ 90 are related to atrial fibrillation [29]. Value five ectopic means the patient has five or more ectopic heart beats per minute. Value other abnormal rhythm means some other abnormal rhythm. Value other represents all other abnormalities.

Attribute	Data Type	Description	Missing
Age	Numerical	Value range: 38-93; mean: 67.98; standard deviation: 7.94.	0%
Angina pectoris	Categorical	Values: none, stable, controlled, uncontrolled; respective frequencies: 564, 110, 144, 1; 4	1.31%
Arrhythmia	Categorical	Values: none, a-fib ≥ 90, other; respective frequencies: 784, 34, 5.	0.83%
ASA grade	Categorical	Values: one, two, three, four, respective frequencies: 4,597, 180, 8; 34	4.53%
Aspirin	Categorical	Values: yes, no ; respective frequencies: 634, 24; 165	19.79%
Blood loss	Numerical	Value range: 0-2000; mean: 280.91; standard deviation: 195.86.	29.68%
Coronary artery bypass surgery	Categorical	Values: yes, no; respective frequencies: 52, 771.	0.95%
Carotid status	Categorical	Values: normal, asymptomatic carotid disease, transient ischaemic attack, vertebral basilar ischemia, nonhemispheric ischemia, postoperative, atrial fibrillation, respective frequencies: 3, 97, 298, 2, 27, 2, 126, 180, 69,cardiovascular arrest, cardiovascular arrest between 6 and 12 months, cardiovascular arrest within 6 months; 17; 2	0.24%
Congestive cardiac failure	Categorical	Values: none, less than 6 months, 6-12 months, more than 12 months; respective frequencies: 796, 10, 1, 16.	0.95%
Diabetes	Categorical	Values: none, impaired glucose tolerance, Diet Rx pill, type one, type two; respective frequencies: 723, 6, 19, 11, 55.	0.12%
Duration	Numeric	Value range: 0.7-100; mean: 1.69, standard deviation: 3.46.	8.81%
ECG	Categorical	Values: normal, q waves, st-t waves, afib 60-90, a-fib ≥ 90, five ectopic, other abnormal rhythm, other; respective frequencies: 564, 74, 35, 16, 7, 2, 16, 84; 25	3.81%
Hypertension	Categorical	Values: yes, no; respective frequencies: 441, 381; 1	0.72%
Myocardial infarct	Categorical	Values: none, within one month, 1 to 6 months, 6 to 12 months, more than 12 months; respective frequencies: 638, 2, 10, 154, 9; 10	2.15%
Patch	Categorical	Values: none, arm vein, leg vein, other vein, dacron, ptfe, stent, other; respective frequencies: 67, 3, 4, 61, 183, 167, 1, 83; 252	31%
Renal failure	Categorical	Values: yes, no; respective frequencies: 10, 813.	0.72%
Respiratory problem	Categorical	Values: none, mild COAD, moderate COAD, severe COAD; respective frequencies: 703, 92, 18, 2; 8	1.79%
Sex	Categorical	Values: female, male; respective frequencies: 331, 492.	0%
Shunt	Categorical	Values: yes, no; respective frequencies: 493, 316;	2%
Side	Categorical	Values: left, right; respective frequencies: 441, 382.	0%
Smoking	Categorical	Values: none, stopped, no more than 20 a day, more than 20 a day, cigars or pipes, cigars and pipes; respective frequencies: 141, 408, 34, 191, 7, 4; 38	5.96%
Warfarin	Categorical	Values: yes, no; respective frequencies: 25, 794; 4	0.60%
Risk	Categorical	Values: low, high; respective frequencies: 703, 120.	0%

Table 1: Description of the cardiovascular dataset showing missing value percentages for each attribute.

Hypertension indicates if a high blood pressure is present. Myocardial infarct indicates if heart attack has occurred or when it occurred. Patch indicates which material is used for by-pass patching in the patient's surgery. The values arm vein/leg vein/other vein indicate different patient body part sources used; while the values dacron and ptfe express the use of synthetic material, either Dacron or polytetrafluoroethylene. Value stent means a stent is inserted into the patient's body. Value none shows there has not been any bypass patching for the patient. Value other means something else is used. Renal failure indicates if renal insufficiency is present.

Respiratory problem indicates problems with breathing, possible values are mild COAD (chronic obstructive airway disease), moderate COAD and severe COAD. Sex represents the gender of the patient. Shunt indicates if a shunt is present. Attribute Side holds the side of surgery. Smoking relates to smoking habits of the patient. Attribute Warfarin indicates if the patient takes warfarin.

Class attribute Risk is used to classify instances into two possible class categorical values high and low risks. The values of class attribute are generated according to the following heuristic model [30]: an instance (cardiovascular patient) is classified into "high" if the patient's death or severe cardiovascular event (e.g. stroke, myocardial relapse or cardiovascular arrest) appears within 30 days after an operation.

Classifier evaluation

K-Fold cross validation is used to minimize the bias associated with random sampling of training and test data samples in comparing predictive accuracy of two or more methods [31]. Here the whole data set is randomly split into 'k' (in our case k=10) mutually exclusive subsets of approximately equal size. Classification model is trained and tested k times. The classification performance is evaluated by accuracy (ACC); sensitivity (Sen); specificity (Spec) rates, and the positive predicted value (PPV) and negative predicted value (NPV), based on values residing in a confusion matrix (see **Table 2**).

Assume that the cardiovascular classifier output set includes two typically risk prediction classes as: "High risk", and "Low risk". Each pattern xi (i=1, 2..n) is allocated into one element from the set (P, N) (positive or negative) of the risk prediction classes. Hence, each input pattern might be mapped into one of four possible outcomes: true positive true high risk (TP) when the outcome is correctly predicted as High risk; true negative true low risk (TN) when the outcome is correctly predicted as Low risk; false negative-false Low risk (FN) when the outcome is incorrectly predicted as Low risk, when it is High risk (positive); or false positivefalse high risk (FP) when the outcome is incorrectly predicted as High risk, when it is Low risk (negative). The set of (P, N) and the predicted risk set can be built as a confusion matrix.

		Predicted classes	
		High risk	Low risk
Expected/Actual	High risk	TP	FN
Classes	Low risk	FP	TN

Table 2: Confusion matrix.

The accuracy of a classifier is calculated by:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \qquad (7)$$

The sensitivity is the rate of number correctly predicted "High risk" over the total number of correctly predicted "High risk" and incorrectly predicted "Low risk". It is given by:

$$Sen = \frac{TP}{TP + FN} \qquad (8)$$

The specificity rate is the rate of correctly predicted "Low risk" over the total number of expected/actual "Low risk". It is given by:

$$Spec = \frac{TN}{TN + FP} \qquad (9)$$

Higher accuracy does not always reflect a good classification outcome. For clinical data analysis it is important to evaluate the classifier based on how well the classifier predicts the "High Risk" patients. In many cases it has been found that the classification outcome is showing good accuracy as it can predict well the low risk patients (majority class) but failed to predict high risk patients (the minority class). For completeness, we also show positive predictive value (PPV) and negative predictive value (NPV), where

$$PPV = \frac{TP}{TP + FP} \qquad (10)$$

$$NPV = \frac{TN}{TN + FN} \qquad (11)$$

Classification Algorithms

Decision tree

Decision trees are algorithms that automatically construct a decision tree from a given data sets. The algorithm generates an optimal decision minimizing the generalization error. A decision tree is articulated as a recursive partition of the instance space. It consists of a directed tree with a "root" node with no incoming edges and all the other nodes have exactly one incoming edge [5]. Decision trees models are mostly used in data mining to examine the data and generate decision rules describing that data. The induced tree and its associated rules are used to make predictions [32]. Ross Quinlan introduced a decision tree algorithm known as Iterative Dichotomiser (ID 3) in 1979. C4.5, as a successor of ID3, is the most widely-used decision tree algorithm. The major advantage to the use of decision trees is human readable and the class-focused visualization of data. This visualization is useful in that it allows users to easily understand the overall structure of data and the decision rules.

K-nearest neighbor algorithm (K-NN)

K-nearest Neighbor (K-NN) method has been becoming interesting topic in data science and proven to be one of the most powerful algorithms for classification. K-NN is a technique for classifying objects based on closest training examples in the feature space. K-NN is a type of lazy learning or instance-based learning [33], where the function is only approximated locally and all computation is deferred until classification.

$$Similarity(x, y) = -\sqrt{\sum_{i=1}^{n} f(x_i, y_i)} \qquad (13)$$

The k-NN is one of the simplest machine learning algorithms where an object is classified by a majority vote of its neighbor's, where the object being allocated to the class most common amongst its "k" nearest neighbor's (k is a positive integer, typically small).

Experiments

The data as described in section 4 was prepared using the procedure outlined in section 3. This is compared to previously published results [34, 35]. Missing values were replaced using the standard Mean/Mode imputation as the basis for comparison. Five classifiers, decision tree (J48), K-NN, Fuzzy Unordered Rule Induction Algorithm (FURIA), SVM and Ripple-down rules (Ridor) [36] were used for predicting missing values. Alternative datasets were prepared by using all the

classifiers and later classified using Decision Tree, K-NN, Neural Networks, Fuzzy Unordered Rule Induction Algorithm (FURIA) and K-Mean clustering.

Classification outcome using standard imputation methods

This experiment was designed to compare classification outcomes and establish a baseline classification for the data. For this, Decision Tree, Ripple-down rules (Ridor), K-NN, FURIA and Neural Network (Support Vector Machine and Multi-Layer Perceptron) classifiers were used. For this experiment the missing values were replaced using the standard Mean/Mode missing imputation technique. No class label balancing technique, see [37] or any other data pre-processing were used. The purpose of these experiments was to set a baseline classification outcome for the data set discussed in section 4.1. The results are presented in the **Table 3** and later compared with the results from other experiments.

Most of the classifiers are showing reasonable accuracy for this data (72% to 80%) but with very poor sensitivity (11% to 23%). Consider the sensitivity rate; the classification outcome of the imbalanced data is very poor because the classifiers give the same attention to the majority class (Low Risk) and the minority class (High Risk). When the imbalance level is huge, it is hard to build a good classifier using conventional learning algorithms. They aim to optimize the overall accuracy without considering the relative distribution of each class. This class imbalance problem is been addressed in our previous research [37]. For all the classifiers used in this experiment the results show that it is hardly possible to achieve an acceptable prediction rate for high-risk patients as they are a minority set in the case of this data. The highest value of sensitivity (23%) is found with the classifier FURIA, which is still very poor.

Classifiers	ACC (%)	SEN (%)	SPEC (%)	PPV (%)	NPV(%)
Decision Tree (J48)	80	11	92	19	86
Ripple-down rules (Ridor)	78	13	89	18	86
SVM	78	15	89	19	86
K-NN	77	21	87	21	87
FURIA	72	23	80	16	86
MLP	78.13	16.67	88.62	20	86.17

Table 3: Baseline classification using mean-mode imputation.

Classification outcome of the dataset prepared using machine learning based imputation methods

We have exhaustively tested the combinations of machine learning imputation and subsequent classification. Rather than present all these results, we will show the results from several combinations (highlighting the best and worst) and then provide a summary table and figure.

Table 4 presents the Decision Tree (J48) classification outcome of the datasets prepared by different missing value imputation methods. It can be observed that the Decision Tree (J48) classified accuracy of all the datasets of different missing values imputation methods are almost closed to each other (78% to 80%) and there is a big gap of sensitivity among all the imputation methods. The highest sensit ivity (23%) was found with the use of Decision Tree (J48) as imputation method.

Table 5 presents the K-NN classification outcome of all the datasets prepared by different missing value imputation methods. The K-NN classified accuracy of all the datasets of the different missing values imputation methods are from 71% to 81% and the highest sensitivity (24%) was found with the use of K-NN as imputation method, and the lowest was by Decision Tree (J48) (20%). The use of K-NN as missing imputation outperformed all the other methods. K-NN has the highest sensitivity (24%), specificity (91%) and accuracy (81%) among all the methods. The statistical method of missing values imputation (mean-mode) has slightly better sensitivity and accuracy then Decision Tree (J48) and SVM as missing imputation methods.

Missing Imputation Methods	ACC (%)	SEN (%)	SPEC (%)	PPV (%)	NPV (%)
Decision Tree (J48)	80	23	90	27	87
K-NN	80	17	90	23	86
FURIA	80	20	90	25	87
SVM	78	15	89	19	86
Ripple-down (Ridor)	78	13	89	18	86
Mean and Mode	80	11	92	19	86

Table 4: Different missing imputation methods with decision tree classification.

Missing Imputation Methods	ACC (%)	SEN (%)	SPEC (%)	PPV (%)	NPV (%)
Decision Tree (J48)	71	20	80	15	85
K-NN	81	24	91	32	88
FURIA	79	21	89	24	87
SVM	71	20	80	15	85
Ripple-down (Ridor)	80	21	90	26	87
Mean and Mode	77	21	87	21	87

Table 5: Different missing imputation methods with K-NN classification.

Table 6 presents the FURIA classification outcome of all the datasets prepared by different missing value imputation methods. First column of the table is the classifier used for training the model with the complete datasets and later used for predicting the missing field of the incomplete dataset. The last row of the table is the classification outcome of the dataset prepared by the standard Mean/Mode missing value imputation method. Again, different machine learning algorithms were applied on the dataset to predict the missing values. The classification results in **Table 6** show that the use of Decision Tree (J48) has high sensitivity (40%). The use of Decision Tree (J48) as missing imputation outperformed all the other methods. Decision Tree (J48) has the highest sensitivity (40%). Although SVM has the high specificity (83%), it shows very poor sensitivity (18%) compared to all the other imputation methods. Fuzzy Unordered Rule Induction Algorithm and K-NN have the same sensitivity of 30%. For Fuzzy Rule

Induction Algorithm (FURIA) the Decision Tree (J48) imputatio n method perform best for predicting the high risk patients.

Missing Imputation Methods	ACC (%)	SEN (%)	SPEC (%)	PPV (%)	NPV (%)
Decision Tree (J48)	63	40	67	17	87
K-NN	67	30	73	16	86
FURIA	67	30	73	16	86
SVM	74	18	83	16	86
Ripple-down (Ridor)	74	20	83	17	86
Mean and Mode	72	23	80	16	86

Table 6: Different missing imputation methods with FURIA classification.

Figure 2 shows the ROC of different combination of the non-stratified machine learning algorithms used for imputing missing values and classifying the final complete data. A random classificatio n line was also drawn to see how much better the classification outcomes are over random. From the figure it can be seen that apart from the combination B and F all the combinations where machine learning algorithm were used, the classification performances are better than random classifier. The combination A (FURIA-K-Means), where FURIA was used to predict and impute the missing values and K-Mean was used to classify the final complete data has got the highest sensitivity.

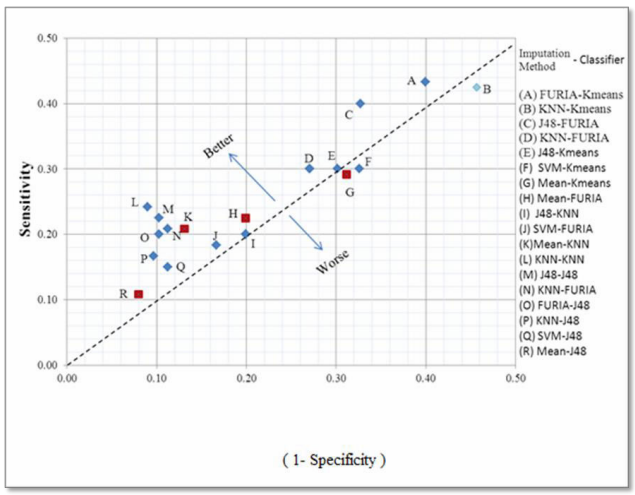

Figure 2: Sensitivity versus (1-Specificity) for All Imputation Methods. The data points A to R can be interpreted via the key with lists (Imputation Method-Classifier) pairings.

If we measure the perpendicular distance of the points from the random classification line the combination L and M are found to have the highest (best) distance from the random line. Some of the classification outcomes of classifiers where Mean/Mode was used to impute the missing vale also show better than random results. However most of them are very low compared to all the combinations where machine learning was used for missing value imputation. Out of the classifications where Mean/Mode was used as missing value imputation the combination K (Mean/ModeK-NN) found to be best.

Table 7 presents the highest sensitivity found from the classifiers used as missing value imputation. First column of the table is the name of the classifier used for missing value imputation and last column is the name of the classifier use to classify the final complete datasets. From the **Table 7** we can conclude that if the research aim is to achieve high sensitivity for unsupervised learning it is recommended to use FURIA as missing value imputation method and for supervised learning decision tree as missing value imputation method.

The results show that with the data prepared using mean mode as missing value we can get maximum 29% sensitivity with 63% accuracy for the K-Means classification. On the other hand we can get 40%-43% sensitivity if we use machine learning methods to predict the missing value. It is observed that in most of the cases if the same classifier is used for predicting the missing value and final classifier the performances are better than the other cases. This is likely because the bias of the classifiers in imputing missing values later benefits that classifier on the complete data. However, this is not always the case. We can also see some other combination of the imputation-classifier classification-classifier can produce good results. Some combinations are able to produce better sensitivity while some are producing better specificity. The appropriate selection of the classifier is an issue for this approach to missing value imputation. It is expected that selection will depend on the data and interests of the research. Preparing the data using Machine Learning algorithm X and achieving best results on that prepared data using the same Machine Learning algorithm X is also to be expected.

Missing Imputation Methods	Highest Sensitivity	With the Accuracy	The Classifier Used
FURIA	43.30%	58%	K-Mean
K-NN	42.50%	51%	K-Mean
Decision Tree (J48)	40%	63%	FURIA
Ripple-down rules (Ridor)	32%	62%	K-Mean
SVM	30%	62%	K-Mean
Mean and Mode	29%	63%	K-Mean

Table 7: The Highest sensitivity values of different missing imputation methods without stratification.

Using Mean-Mode we are imputing the unique value for the entire missing field but it is obvious that missing values cannot be unique. It is a big challenge to find the right value for the missing field. The proposed method uses pattern recognition technique to predict the value for the missing field by learning the pattern from the complete dataset. The experiments show that this method is giving an improved way of finding the best possible value for the missing fields. Finally, we show the effect of stratification on this. The results **(Table 8)** are shown without K-NN as this had no effect when stratified, with the results given above not improved on.

Datasets are prepared using stratified machine learning based missing value imputation method discussed in section 3, and are then classified using Decision Tree, K-NN, FURIA and Neural Network. Standard mean/mode imputation and non-stratified machine learning

based missing value imputation method also been used for comparison.

Classifier	Mean/Mode		Machine learning			
			Non Stratified		Stratified	
	SEN (%)	SPEC (%)	SEN (%)	SPEC (%)	SEN (%)	SPEC (%)
Decision tree	11	92.13	20	90	39.17	90.61
K-NN	21	87.2	21	89	20.83	90.47
FURIA	22.5	80.09	30	72.97	62.5	67.57
Neural network (MLP)	16.67	88.62	20.83	87.97	25.83	91.32

Table 8: Experimental results for alternative missing value imputation methods and strategies.

The summary of the results, presented in **Table 8** and **Figure 3**, show that proposed stratified machine learning based missing imputation method outperform other methods discuss in this paper. Apart from K-NN classification (which is omitted from **Figure 3**) all the other classification performances have significantly improved using the proposed method for missing value imputation.

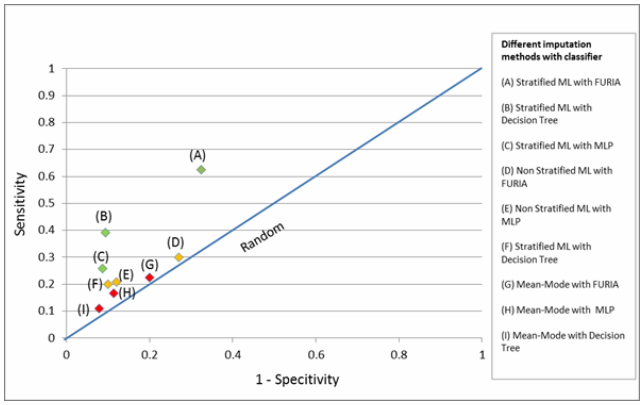

Figure 3: Summary of best results for sensitivity versus (1-Specificity) across all imputation methods. The data points A to I can be interpreted via the key (different imputation methods with classifier). (K-NN is omitted to no improvement on imputation via stratification).

Conclusion

Like many other real life data sets medical data are usually found to be incomplete, which causes many problems in analytics and knowledge discovery. This work proposed a missing value imputation framework using stratified machine learning techniques. The results are compared with non-stratified machine learning based missing value imputation and statistical (mean/mode) imputation. Experimental results show that the proposed stratified machine learning methods outperformed the statistical method (Mean/Mode) and other non-stratified machine learning methods.

The proposed method might be computationally expansive for a big datasets having large numbers attributes with missing fields. However, it is known that data cleaning is part of data pre-processing task and a one-off process. With this extra effort we can achieve a good quality data for better knowledge discovery and decision support.

In agreement with other recent research [38] and findings of this experiment we can infer that machine learning techniques may be the best approach to imputing missing values for better classification outcomes. However providing a generic answer for which is the best combination of machine learning algorithm for missing value imputation and final classification remains an open question. Unlike [38-43], we found that K-NN is not an optimal strategy to follow when using stratified imputation. The results shown here and in other work [35] suggest that the data domain and label used in the classification problem have a bearing on this question. We can confidently say that stratified machine learning imputation does improve final classification results in the datasets tested. Furthermore, the machine learning algorithm used for missing value imputation is not necessarily the best for final classification; so countering the argument that the method produces a data bias for the given classifier.

Acknowledgement

The work reported here was undertaken while the first author was a PhD student at the University of Hull. The second (and corresponding) author was the originator of the project and supervisor for the PhD. The PhD study was funded by a Department of Computer Science SEED PhD scholarship.

References

1. Chan P, Dunn OJ (1972) The treatment of missing values in discriminant analysis. Journal of the American Statistical Association 6: 473-477.
2. Almeida RJ, Kaymak U, Sousa JMC (2010) A new approach to dealing with missing values in datadriven fuzzy modelling. IEEE International Conference on Fuzzy Systems (FUZZ). IEEE, Barcelona, pp: 1-7.
3. Roderick JAL, Donald BR (2002) Statistical Analysis with Missing Data.
4. Marlin BM (2008) Missing Data Problems in Machine Learning. Graduate Department of Computer Science.
5. Maimon O, Rokach L (2010) Data Mining and Knowledge Discovery Handbook. Springer, London.
6. Jerez JM, Molina L, García-Laencina PJ, Alba E, Ribelles N, et al. (2010) Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artificial Intelligence in Medicine 50: 105-115.
7. Peugh JL, Enders CK (2004) Missing data in educational research: A review of reporting practices and suggestions for improvement, Review of Educational Research 74:525-556.
8. Esther-Lydia SRR, Pino-Mejias M, Lopez-Coello MD, Cubiles-de-la-Vega (2011) Missing value imputation on missing completely at random data using multilayer perceptrons, Neural Networks.
9. García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR (2010) Pattern classification with missing data: A review, Neural Computing and Applications 19: 263-282.
10. Meesad P, Hengpraprohm K (2008) Combination of K-NN-Based Feature Selection and K-NNBased Missing-Value Imputation of Microarray Data. 3rd International Conference on Innovative Computing Info rmation and Control, 2008. ICICIC '08, p: 341.
11. Lingras P, Zhong M, Sharma S (2008) Evolutionary Regression and Neural Imputations of Missing Values. Chapter in Soft Computing Applications in Industry, Volume 226 of the series Studies in Fuzziness and Soft Computing, pp: 151-163.

12. Setiawan NA, Venkatachalam P, Hani AFM (2008) Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory. International Conference on BioMedical Engineering and Informatics, BMEI 2008, pp: 360-310.
13. Wang L, Fu DM (2009) Estimation of Missing Values Using a Weighted K-Nearest Neighbors Algorithm. International Conference on Environmental Science and Information Application Technology, 2009, pp: 660-663.
14. Weiss SM, Indurkhya N (2000) Decision-rule solutions for data mining with missing values, In: IBERAMIA-SBIA, pp: 1-10.
15. Yun-fei Q, Xin-yan Z, Xue L, Liang-shan S (2010) Research on the missing attribute value data-oriented for decision tree. 2nd International Conference onSignal Processing Systems (ICSPS) 2010.
16. Heikki JH, Niska K, Tuppurainen J, Ruuskanen M, Kolehmainen (2004) Methods for imputation of missing values in air quality data sets, Atmospheric Environment 38: 1352-2310.
17. Bruha I (2004) Meta-Learner for Unknown Attribute Values Processing: Dealing with Inconsistency of MetaDatabases. J Intell Inf Syst 22: 71-87.
18. Gajawada S, Toshniwal D (2012) 'Missing Value Imputation Method Based on Clustering and Nearest Neighbours', International Journal of Future Computer and Communication 1: 206-208.
19. Batista G, Monard MC (2003) 'An analysis of four missing data treatment methods for supervised lear ning', Applied Artificial Intelligence 17: 519-533.
20. Hühn J, Hüllermeier E (2009) Fuzzy Unordered Rules Induction Algorithm, Data Mining and Knowledge Discovery 19: 293-319.
21. Sun Jh, Xu XL (2009) Large Rotating Machinery Fault Diagnosis and Knowledge Rules Acquiring Based on Improved RIPPER. Intelligent Computation Technology and Automation, 2009. ICICTA '09. Second International Conference, pp: 549-552.
22. Lotte F, Lecuyer A, Arnaldi B (2007) FuRIA: A Novel Feature Extraction Algorithm for Brain-Computer Interfaces using Inverse Models and Fuzzy Regions of Interest. 3rd International IEEE/EMBS Conference on Neural Engineering, CNE '07, pp: 175-178.
23. Lotte F, Lecuyer A, Arnaldi B (2009) FuRIA: An Inverse Solution Based Feature Extraction Algorithm Using Fuzzy Set Theory for Brain–Computer Interfaces, IEEE Transactions on Signal Processing 57: 3253-3263.
24. Nguyen TTT (2009) Predicting Cardiovascular Risks using Pattern Recognition and Data Mining, Ph.D Thesis, Computer Science, University of Hull.
25. Daabiss M (2011) American Society of Anaesthesiologists physical status classification. Indian Journal of Anaesthesia 55: 111–115.
26. Fuller J, Shipley M, Rose G, Jarrett RJ, Keen H (1980) Coronary-heart-disease risk and impaired glucose tolerance The Whitehall Study. The Lancet 315: 1373-1376.
27. Trial CAS, II Investigators (1992) Effect of the antiarrhythmic agent moricizine on survival after myocardial infarction. N Engl J Med 327: 227.
28. Zhang L, Timothy KW, Vincent GM, Lehmann MH, Fox J, et al. (2000) Spectrum of ST-T–wave patterns and repolarization parameters in congenital long-QT Syndrome ECG findings identify genotypes. Circulation 102: 2849-2855.
29. Mittal S, Pokushalov E, Romanov A, Ferrara M, Arshad A, et al. (2013) Longterm ECG monitoring using an implantable loop recorder for the detection of atrial fibrillation after cavotr icuspid isthmus ablation in patients with atrial flutter. Heart Rhythm 10: 1598-1604.
30. Davis DN, Nguyen TTT (2008) Generating and Verifying Risk Prediction Models Using Data Mining: A Case Study from Cardiovascular Medicine. In: Data Mining and Medical Knowledge Management: Cases and Applications.
31. University of Toronto, Toronto, Canada.
32. Barros RC, Basgalupp MP, De Carvalho AC, Freitas AA (2012) A survey of evolutionary algorithms for decision-tree induction. Systems, Man, and Cybernetics, Part C: Applications and Reviews. IEEE Transactions 42: 291-312.
33. Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Machine learning 6: 37-66.
34. Rahman MM, Davis DN (2014) Semi Supervised Under-Sampling: A Solution to the Class Imbalance Problem for Classification and Feature Selection, Chapter in: Transactions on Engineering Technologies (Special Volume of the World Congress on Engineering 2013), pp: 611-625.
35. Rahman MM (2014) Machine learning based data pre-processing for the purpose of medical data mining and decision support, PhD Thesis, Dept of Computer Science, University of Hull.
36. Gaines BR, Compton P (1995) 'Induction of Ripple-Down rules applied to modelling large databases', Journal of Intelligent information system 5: 221-228.
37. Rahman MM, Davis DN (2013) Addressing the Class Imbalance Problem in Medical Datasets, International Journal of Machine Learning and Computing, pp: 224-228.
38. Liu ZG, Pan Q, Dezert J, Martin A (2016) Adaptive imputation of missing values for incomplete pattern classification, Pattern Recognition 52: 85-95.
39. Han J, Kamber M (2001) Data Mining: concepts and techniques, San Francisco: Morgan Kaufmann Publishers.
40. Honghai F, Guoshun C, Cheng Y, Bingru Y, Yumei C (2005) 'A SVM Regression Based Approach to Filling in Missing Values'. In: Khosla R, Howlett R, Jain L (eds.), Knowledge-Based Intelligent Information and Engineering Systems, Springer Berlin Heidelberg, pp: 581-587.
41. Pooling Project Research Group (1978) Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: final report of the Pooling Project. Journal of Chronic Diseases 31: 201-306.
42. Quinlan JR (1993) C4.5: programs for machine learning, San Mateo: Morgan Kaufmann.
43. JL (1997) Analysis of Incomplete Multivariate Data. Chapman and Hall, London.

This article was originally published in a special issue, entitled: **"Industrial and Data Mining",** Edited by S1