Chapter #

Machine Learning Based Missing Value Imputation Method for Clinical Datasets

M. Mostafizur Rahman¹ and Darryl N. Davis²

¹M. M. Rahman is a PhD student in the Department of Computer Science, University of Hull, UK, and faculty member department of computer science, Eastern University Dhaka, Bangladesh. (e-mail: mmrbappy@gmail.com, M.M.Rahman@2009.hull.ac.uk)

²Darryl N. Davis is the Director of Research, Department of Computer Science, University of Hull, UK (e-mail: D.N.Davis@hull.ac.uk).

Abstract: Missing value imputation is one of the biggest tasks of data pre-processing when performing data mining. Most medical datasets are usually incomplete. Simply removing the incomplete cases from the original datasets can bring more problems than solutions. A suitable method for missing value imputation can help to produce good quality datasets for better analysing clinical trials. In this paper we explore the use of a machine learning technique as a missing value imputation method for incomplete cardiovascular data. Mean/mode imputation, fuzzy unordered rule induction algorithm imputation, decision tree imputation and other machine learning algorithms are used as missing value imputation and the final datasets are classified using decision tree, fuzzy unordered rule induction, KNN and K-Mean clustering. The experiment shows that final classifier performance is improved when the fuzzy unordered rule induction algorithm is used to predict missing attribute values for K-Mean clustering and in most cases, the machine learning techniques were found to perform better than the standard mean imputation technique.

Keywords: Cardiovascular, FURIA, Fuzzy Rules, J48, K-Mean, Missing Value.

1. INTRODUCTION

Many researchers have identified several important and challenging issues [1-3] for clinical decision support. In “Grand challenges for decision support” Sittig [1] et al. set out 10 critical problems for “designing, developing, presenting, implementing, evaluating, and maintaining all types
of clinical decision support capabilities for clinicians, patients and consumers”. However Sittig et al.’s identification does cover little about data preprocessing. Sometimes, improved data quality is itself the goal of the analysis, usually to improve processes in a production database [4] and designing of decision support.

Two types of databases are available in medical domain [5]. The first is the dataset acquired by medical experts, which are collected for a special research topic where data collection is triggered by the generated hypothesis of a clinical trial. The other type is a huge dataset retrieved from hospital information systems. These data are stored in a database automatically without any specific research purpose. These data records are often used for further analysis and building clinical decision support system. These types of datasets are very complex where the numbers of records are very huge, with a large number of attributes for each record; many missing values and typically the datasets are mostly imbalanced with regard to their class label. In this paper we will be addressing the issue of missing value in clinical (cardiovascular) datasets.

Many real-life data sets are incomplete. The problem with missing attribute values is a very important issue in Data Mining. In medical data mining the problem with the missing values has become a challenging issue. In many clinical trials, the medical report pro-forma allow some attributes to be left blank, because they are inappropriate for some class of illness or the person providing the information feels that it is not appropriate to record the values for some attributes [6].

Typically there are two types of missing data [7]; one is called missing completely at random or MCAR. Data is MCAR when the response indicator variables R are independent of the data variables X and the latent variables Z. The MCAR condition can be succinctly expressed by the relation 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 frequently written as P(R = r|X = x, Z = z, µ) = P(R = r|X^o = x^o, ) for all x^o, z and µ [8, 9].

In general, methods to handle missing values belong either to sequential methods like leastwise deletion, assigning most common values, arithmetic mean for the numeric attribute etc. or parallel methods where rule induction algorithm are used to predict missing attribute values [10]. There are reasons for which sequential like leastwise deletion is considered to be a good method [7], but several works [6, 7, 11] have shown that the application of this method on the original data can corrupt the interpretation of the data and mislead the subsequent analysis through the introduction of bias.
While several techniques for missing value imputation are employed by researchers, most of the techniques are single imputation approaches [12]. The most traditional missing value imputation techniques are deleting case records, mean value imputation, maximum likelihood and other statistical methods [12]. In recent years, research has explored the use of machine learning techniques as a method for missing values imputation in several clinical and other incomplete datasets [13]. Machine learning algorithm such as multilayer perception (MLP), self-organising maps (SOM), decision tree (DT) and k-nearest neighbours (KNN) have been used as missing value imputation methods in different domains [11, 14-21]. Machine learning methods like MLP, SOM, KNN and decisions tree have been found to perform better than the traditional statistical methods [11, 22].

In this paper we examine the use of Machine Learning techniques as a missing values imputation method for real life incomplete cardiovascular datasets. Where, we have used classifier to predict the value for a missing field and impute the predicted value to make the dataset complete. In order to compare the performance we have used four classifiers, Decision Tree [10], KNN [32], SVM [35] and FURIA [23] to predict the missing values. The datasets are later classified using Decision Tree, KNN, FURIA and K-Means Clustering; the results are compared with commonly used mean-mode imputation methods.

2. OVERVIEW OF FURIA

Fuzzy Unordered Rule Induction Algorithm (FURIA) is a fuzzy rule-based classification method, which is a modification and extension of the state-of-the-art rule learner RIPPER. Fuzzy rules are obtained through replacing intervals by fuzzy intervals with trapezoidal membership functions [23]:

\[
I^*(v) = \begin{cases} 
1 & \phi^L \leq v \leq \phi^U \\
\frac{v - \phi^L}{\phi^U - \phi^L} & \phi^L \leq v \leq \phi^C \\
\frac{\phi^U - v}{\phi^U - \phi^C} & \phi^C \leq v \leq \phi^U \\
0 & \text{else}
\end{cases}
\]

(1)

Where $\phi^C, L$ and $\phi^C, U$ are the lower and upper bound of the membership of the fuzzy sets. For an instance $x = (x_1, \ldots, x_n)$ the degree of the fuzzy membership can be found using the formula [23]:

\[
\]
\[ \mu_F(x) = \prod_{i=1}^{k} \mu_i^{\text{F}}(x_i) \]  

(2)

For fuzzification of a single antecedent only relevant training data is \( D_T^k \) considered and data are partitioned into two subsets and rule purity is used to measure the quality of the fuzzification [23]:

\[ D_T^k = \{ x = (x_1, \ldots, x_k) \in D_T^k | \mu_i^{\text{F}}(x_i) > 0 \text{ for all } j \neq i \} \subseteq D_T \]  

(3)

\[ \text{Pur} = \frac{p_i}{p_i + n_i} \]  

(4)

Where

\[ p_i = \sum_{x \in D_T^k} \mu_{A_i}(A) \]

\[ n_i = \sum_{x \in D_T^k} \mu_{A_i}(A) \]

The fuzzy rules \( r_1^{(j)}, \ldots, r_k^{(j)} \) have been learned for the class \( \lambda_j \), the support of this class is defined by [23]:

\[ s_j(x) = \sum_{i=1}^{k} \mu_{A_i}(A) \cdot \text{CF}(r_i^{(j)}) \]  

(5)

Where, the certainty factor of the rule is defined as

\[ \text{CF}(r_i^{(j)}) = \frac{2 \left| D_T^{(j)} \right| + \sum_{x \in D_T^{(j)}} \mu_{A_i}(A)}{2 + \sum_{x \in D_T^{(j)}} \mu_{A_i}(A)} \]  

(6)

The use of the algorithm in of data mining can be found in [23-25].

3. DECISION TREE

The decision tree classifier is one of the most widely used supervised learning methods. A decision tree is expressed 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. [10]. Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions [26].
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 [27]. The major advantage to the use of decision trees is the class-focused visualization of data. This visualization is useful in that it allows users to readily understand the overall structure of data in terms of which attribute mostly affects the class (the root node is always the most significant attribute to the class). Typically the goal is to find the optimal decision tree by minimizing the generalization error [28]. The algorithms introduced by Quinlan [29, 30] has proved to be an effective and popular method for finding a decision tree to express information contained implicitly in a data set. WEKA [31] makes use of an implementation of C4.5 algorithm called J48 which has been used for all of our experiments.

4. K-NEAREST NEIGHBOUR ALGORITHM (KNN)

K-nearest Neighbor Algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space (defined using for example the Similarity measure). K-NN is a type of instance-based learning [32] or lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms where an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbour.

5. K-MEANS CLUSTERING

K-means is one of the simplest unsupervised learning algorithms proposed by Macqueen in 1967, which has been used by many researchers to solve some well-known clustering problems [10]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters). The algorithm first randomly initializes the clusters center. The next step is to calculate the distance (discussed in the above section) between an object and the centroid of each cluster then take each point belonging to a given data set and associate it to the nearest centre and re-calculate the cluster centres. The process is repeated with the aim
of minimizing an objective function known as squared error function given by:

$$ J(v) = \sum_{i=1}^{C} \sum_{j=1}^{c_i} (||x_i - v_j||)^2 $$  \hspace{1cm} (8)

Where, $||x_i - v_j||$ is the Euclidean distance between $x_i$ and $v_j$, $c_i$ is the number of data points in $i^{th}$ cluster and $c$ is the number of cluster centers.

6. CARDIOVASCULAR DATA

We have used two datasets from Hull and Dundee clinical sites. The Hull site data includes 98 attributes and 498 cases of cardiovascular patients and the Dundee site data includes 57 attributes, and 341 cases from cardiovascular patients. After combining the data from the two sites, 26 matched attributes are left.

Missing values: After combining the data and removing redundant attributes we found that out of 26 attributes 18 attributes have a missing value frequency from 1% to 30% and out of 832 records 613 records have 4% to 56% missing values in their attributes.

From these two data sets, we prepared a combined dataset having 26 attributes with 823 records. Out of 823 records 605 records have missing values and 218 records do not have any missing values. Among all the records 120 patients are alive and 703 patients are dead. For this experiment according to clinical risk prediction model (CM1) [33], patients with status “Alive” are considered to be “Low Risk” and patients with status “Dead” are considered to be “High Risk”.

7. MEAN AND MODE IMPUTATION

This is one of the most frequently used methods. It consists of replacing the unknown value for a given attribute by the mean ($\bar{x}$) (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute [21].

$$ \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i $$ \hspace{1cm} (9)

It replaces all missing records with a single and unique value $\bar{x}$, which is the mean value of that attribute.
8. PROPOSED MISSING VALUE IMPUTATION PROCESS

The original data set is first portioned into groups. The records having missing values in their attributes are in one group (the complete data set) and the records without any missing values are placed in a separate group. The classifier is trained with the complete data sets, and later the incomplete data is given to the model 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 missing value imputed datasets are combined to make the finalised data. The final dataset is then fed to the selected classifier for classification (as shown in Figure 1).

![Figure 1 Missing Value Imputation Process](image_url)
9. RESULTS

We have experimented with a number of machine learning algorithms as missing value imputation mechanisms; such as FURIA, decision tree [34], and SVM [35]. The performance is compared with the most commonly used missing imputation statistical method mean-mode. The results are also compared with the previously published results of the same experimental dataset with mean-mode imputation for K-Mix clustering [36].

Table 1 Different missing imputation methods with k-mean clustering

<table>
<thead>
<tr>
<th>Missing Imputation Methods</th>
<th>Confusion Matrix</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk</td>
<td>Classified High Risk</td>
<td>Classified Low Risk</td>
<td>ACC</td>
<td>SEN</td>
<td>SPEC</td>
</tr>
<tr>
<td>Decision tree (J48)</td>
<td>High</td>
<td>36</td>
<td>84</td>
<td>0.64</td>
<td>0.30</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>212</td>
<td>491</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FURIA</td>
<td>High</td>
<td>52</td>
<td>68</td>
<td>0.58</td>
<td>0.43</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>281</td>
<td>422</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>High</td>
<td>36</td>
<td>84</td>
<td>0.62</td>
<td>0.30</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>229</td>
<td>474</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean and Mode</td>
<td>High</td>
<td>35</td>
<td>85</td>
<td>0.63</td>
<td>0.29</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>219</td>
<td>484</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Comparison results with k-mix clustering

| Classifier with Different Missing Imputation Methods | Confusion Matrix | | | | |
|-------------------------------------------------------|-----------------|---|---|---|
| K-Mix (With Mean Mode imputation)                     | Risk            | Classified High Risk | Classified Low Risk | SEN | SPEC |
| High                                                  | 35 | 21 | 0.25 | 0.89 |
| Low                                                   | 107 | 177 |     |     |
| K-Mean With Fuzzy Unordered Rule Induction Algorithm used as missing value imputation method | High | 52 | 68 | 0.43 | 0.60 |
| Low                                                   | 281 | 422 |     |     |

From the table 1 one can see that for K-mean clustering, decision tree imputation method shows accuracy of 64% (slightly better than the other methods) but the sensitivity is 30% which is almost as poor as the mean/mode imputation. SVM and mean/mode mutation show very similar performance with accuracy of 62% to 63% and sensitivity of 29% to 32%. On the other hand, fuzzy unordered rule induction algorithm as a missing value imputation method shows sensitivity of 43% with accuracy of 58%.
Table 2 shows the comparison results of previously published results of K-Mix [37] clustering algorithm with mean mode imputation and simple K-mean clustering with FURIA missing value imputation. The result shows that the K-mean with FURIA as missing value imputation has higher sensitivity (43%) than the K-mix with conventional mean/mode imputation method (0.25%).

The datasets prepared by different imputation methods are also classified using well known classifier decision tree (J48), KNN and also with FURIA. The classification outcomes are presented in Table 3 to Table 5. Table 6 presents the highest sensitivity value found of all the datasets prepared by different imputation methods and missing value imputation using FURIA shows the sensitivity 43.3% which is the highest among all the machine learning methods and statistical method explored in this paper.

Table 3 Different Missing Imputation Methods with J48 Classification

<table>
<thead>
<tr>
<th>Missing Imputation Methods</th>
<th>Confusion Matrix</th>
<th>Actual Risk</th>
<th>Classified Risk</th>
<th>ACC</th>
<th>SEN</th>
<th>SPEC</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision tree (J48)</td>
<td></td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>27</td>
<td>93</td>
<td>0.80</td>
<td>0.23</td>
<td>0.90</td>
<td>0.27</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>72</td>
<td>631</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-NN</td>
<td></td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>20</td>
<td>100</td>
<td>0.80</td>
<td>0.17</td>
<td>0.90</td>
<td>0.23</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>68</td>
<td>635</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FURIA</td>
<td></td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>24</td>
<td>96</td>
<td>0.80</td>
<td>0.20</td>
<td>0.90</td>
<td>0.25</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>72</td>
<td>631</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>18</td>
<td>102</td>
<td>0.78</td>
<td>0.15</td>
<td>0.89</td>
<td>0.19</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>79</td>
<td>624</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
<td>107</td>
<td>0.80</td>
<td>0.11</td>
<td>0.92</td>
<td>0.19</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>56</td>
<td>647</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For clinical data analysis it is important to evaluate the classifier based on how well the classifier is performing to predict the “High Risk” patients. As indicated earlier the dataset shows an imbalance on patient’s status. Only 120 records, out of 832 records, are of “High Risk” (14.3% of the total records). A classifier may give very high accuracy if it can correctly classify the “Low Risk” patients but is of limited use if it does not correctly classify the “High Risk” patients. For our analysis we gave more importance to Sensitivity and Specificity then Accuracy to compare the classification outcome.
If we analyse the ROC [38] space for all the imputation methods classified with three classifiers mentioned earlier and one clustering algorithm plotted in figure 2, we will find that most the machine learning methods are above
the random line and most of the cases better than the statistical mean/mode imputation.

Table 6 Highest sensitivity value found with each of the imputation method

<table>
<thead>
<tr>
<th>Missing Imputation Methods</th>
<th>Highest Sensitivity</th>
<th>With the Accuracy</th>
<th>The Classifier Used to Classify</th>
</tr>
</thead>
<tbody>
<tr>
<td>FURIA</td>
<td>43.3%</td>
<td>58%</td>
<td>K-Mean</td>
</tr>
<tr>
<td>K-NN</td>
<td>42.5%</td>
<td>51%</td>
<td>K-Mean</td>
</tr>
<tr>
<td>J48</td>
<td>40%</td>
<td>63%</td>
<td>FURIA</td>
</tr>
<tr>
<td>SVM</td>
<td>30%</td>
<td>62%</td>
<td>K-Mean</td>
</tr>
<tr>
<td>Mean</td>
<td>29%</td>
<td>63%</td>
<td>K-Mean</td>
</tr>
</tbody>
</table>

Figure 2 The ROC space and plots of the different imputation methods classified with J48, FURIA, KNN and K-Means.

If we evaluate the missing imputation based on the sensitivity than we can see the FURIA missing value imputation outperformed all the other machine learning and traditional mean/mode approaches to missing value imputation methods that we have examined in this work.
10. THE COMPLEXITY OF THE PROPOSED METHOD

The complexity of the proposed method is related with the complexity of the classifier is used for the missing value imputation. If we use FURIA, than the fuzzy unordered rule induction algorithm can be analysed by considering the complexity of the rule fuzzification procedure, rule stretching and re-evaluating the rules. For \(|D_T|\) training data and \(n\) numbers of attribute the complexity of the fuzzification procedure is \(O(|D_T|n^2)\) [23], with \(|RSI|\) numbers of rules and \(|D_T|\) training data the complexity of rule stretching is \(O(|D_T|n^2)\) [23], and rule \(r\) with antecedent set \(A(r)\) the complexity for the rule re-evaluating is \(O(|A(r)|)\). For the experimental data of 823 records with 23 attributes on an average it took 0.69 second to build the model for each attribute of missing values.

11. CONCLUSION

Missing attribute values are common in real life datasets, which causes many problems in pattern recognition and classification. Researchers are working towards a suitable missing value imputation solution which can show adequate improvement in the classification performance. Medical data are usually found to be incomplete as in many cases on medical reports some attributes can be left blank, because they are inappropriate for some class of illness or the person providing the information feels that it is not appropriate to record the values. In this work we examined the performance of machine learning techniques as missing value imputation. The results are compared with traditional mean/mode imputation. Experimental results show that all the machine learning methods which we explored outperformed the statistical method (Mean/Mode), based on sensitivity and some cases accuracy.

The process of missing imputation with our proposed method can be computationally expansive for large numbers of attribute having missing values in their attributes. However, we know that data cleaning is part of data pre-processing task of data mining which is not a real time task and neither a continuous process. Missing value imputation is a onetime task. With this extra effort we can obtain a good quality data for better classification and decision support.

We can conclude that machine learning techniques may be the best approach to imputing missing values for better classification outcome.
REFERENCES


