

Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data

M. Mostafizur Rahman*, and Darryl N. Davis**

Department of Computer Science,
The University of Hull, Hull, United Kingdom
M.M.Rahman@2009.hull.ac.uk, 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 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 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.

Keywords: Missing Value, FURIA, K-Mix, Fuzzy Rules, K-Mean, Cardiovascular, Decision Tree.

I. INTRODUCTION

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 of some attributes [1].

Typically there are two types of missing data [2], 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 $P(R|X, Z, \mu) = P(R|\mu)$. The second category of missing data is called missing at random or MAR. The MAR condition is frequently written as $P(R = r|X = x, Z = z, \mu) = P(R = r|X^o = x^o, \mu)$ for all x^o , z and μ [3].

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 [4]. There are reasons for which sequential like leastwise deletion is considered to be a good method [2],

but several works [1, 2, 5] 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 [6]. The most traditional missing value imputation techniques are deleting case, mean value imputation, maximum likelihood and other statistical methods [6]. 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. Machine learning algorithm such as multilayer perception (MLP), self-organising maps (SOM), decision tree (DT) and k-nearest neighbours (KNN) were used as missing value imputation methods in different domains [5, 7-13]. Machine learning methods like MLP, SOM, KNN and decisions tree have been found to perform better than the traditional statistical methods [5, 14].

In this paper we examine the use of fuzzy unordered rules induction algorithm [15] as a missing values imputation method for real life incomplete cardiovascular datasets. The results are compared with decision tree, SVM, and mean-mode used as imputation methods. K-Mean clustering algorithm is used as the final classifier for each case.

II. 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 function [15].

$$I^F(v) \stackrel{\text{df}}{=} \begin{cases} 1 & \phi^{c,L} \leq v \leq \phi^{c,U} \\ \frac{v - \phi^{s,L}}{\phi^{c,L} - \phi^{s,L}} & \phi^{s,L} \leq v \leq \phi^{c,L} \\ \frac{\phi^{s,U} - v}{\phi^{s,U} - \phi^{c,U}} & \phi^{c,U} \leq v \leq \phi^{s,U} \\ 0 & \text{else} \end{cases} \quad (1)$$

Where $\phi^{c,L}$ and $\phi^{c,U}$ are the lower and upper bound of the membership of the fuzzy sets. For an instance $x = (x_1, \dots, x_n)$ the degree of the fuzzy membership can be found using the formula [15]:

$$\mu_{r^F}(x) = \prod_{i=1 \dots k} i_i^F(x_i) \quad (2)$$

For fuzzification of a single antecedent only relevant training data is D_T^i considered and data are partitioned into two subsets and rule purity is used to measure the quality of the fuzzification [15].

$$D_T^i = \{x = (x_1, \dots, x_k) \in D_T^i | I_j^F(x_j) > 0 \text{ for all } j \neq i\} \subseteq D_T \quad (3)$$

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

Where

$$p_i \stackrel{\text{def}}{=} \sum_{x \in D_T^+} \mu_{A_i}(A)$$

$$n_i \stackrel{\text{def}}{=} \sum_{x \in D_T^-} \mu_{A_i}(A)$$

The fuzzy rules $r_1^{(j)} \dots r_k^{(j)}$ have learned for the class λ_j , the support of this class is defined by [15]

$$s_j(x) = \sum_{i=1 \dots k} \mu_{r_i^{(j)}}(x) \cdot CF(r_i^{(j)}) \quad (5)$$

where the certainty factor of the rule is defined as

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

The use of the algorithm in different areas of data mining can be found in [15-17].

III. 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 the 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) [18], patients with status ‘‘Alive’’ are consider to be ‘‘Low Risk’’ and patients with status ‘‘Dead’’ are consider to be ‘‘High Risk’’.

IV. MISSING VALUE IMPUTATION PROCESS

The original data set is first portioned in to groups. 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. 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 complete data. The final dataset is then fed to the selected classifier for classification (as shown in Figure 1).

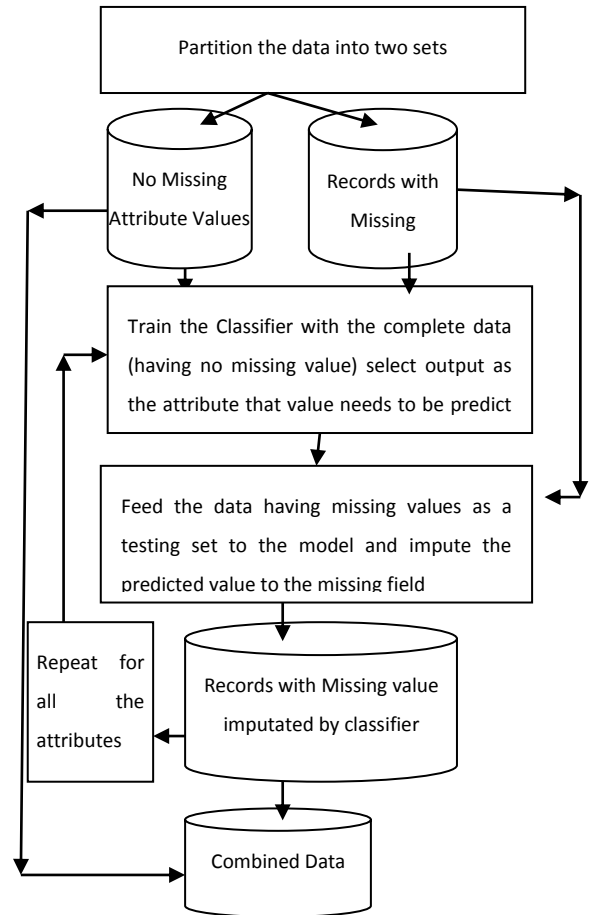


Figure 1. Missing Value Imputation process

V. RESULTS

We have experimented with a number of machine learning algorithms as missing value imputation mechanisms; such as FURIA, decision tree [19], SVM [20] and ripple-down rules [21]. 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 [22].

TABLE 1.
DIFFERENT MISSING IMPUTATION METHODS WITH K-MEAN CLUSTERING

Missing Methods	Imputation	Confusion Matrix			ACC	SEN	SPEC	PPV	NPV
		Risk	Classified High Risk	Classified Low Risk					
Decision tree (J48)	High	36	84	0.64	0.30	0.70	0.15	0.85	
	Low	212	491						
Fuzzy Unordered Rule Induction Algorithm	High	52	68	0.58	0.43	0.60	0.16	0.86	
	Low	281	422						
SVM	High	36	84	0.62	0.30	0.67	0.14	0.85	
	Low	229	474						
Ripple-down rules	High	38	82	0.62	0.32	0.67	0.14	0.85	
	Low	230	473						
Mean and Mode	High	35	85	0.63	0.29	0.69	0.14	0.85	
	Low	219	484						

TABLE 2.
COMPARISON RESULTS WITH K-MIX CLUSTERING [23]

Classifier with Different Missing Imputation Methods	Risk	Confusion Matrix		SEN	SPEC
		Classified High Risk	Classified Low Risk		
K-Mix (With Mean Mode imputation)	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 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, Ripple-down rules, 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 [23] 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%).

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 are of “High Risk” out of 832 records (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 than Accuracy to compare the classification outcome. 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.

The complexity of 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 $O(|D_T|n^2)$ [15], with $|RS|$ numbers of rules and $|D_T|$ training data the complexity of rule stretching is $O(|D_T|n^2)$ [15], and rule r with antecedent set $\mathcal{A}(r)$ the complexity for the rule re-evaluating is $O(|\mathcal{A}(r)|)$. 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. The process of missing imputation with FURIA can be a bit computationally expansive for large numbers of attribute having missing in their attribute values but can produce a high quality cleaned dataset.

VII. 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 for K-Mean clustering and the results are compared with traditional mean/mode imputation. Experimental results shows that the all the machine learning methods that we explored outperformed the statistical method (Mean/Mode), based on sensitivity and some cases accuracy, and out of all the machine learning technique that we explored the Fuzzy Unordered Rule Induction Algorithm found to be the preferred technique for missing value imputation.

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

REFERENCES

- [1] R. J. Almeida, U. Kaymak, and J. M. C. Sousa, "A new approach to dealing with missing values in data-driven fuzzy modelling," in IEEE International Conference on Fuzzy Systems (FUZZ), Barcelona, 2010, pp. 1 - 7.
- [2] J. A. L. Roderick, and B. R. Donald, *Statistical Analysis with Missing Data*, second edition ed.: Wiley, 2002.
- [3] B. M. Marlin, "Missing Data Problems in Machine Learning," Graduate Department of Computer Science, University of Toronto, Toronto, Canada, 2008.
- [4] O. Maimon, and L. Rokach, *Data Mining and Knowledge Discovery Handbook*, Second ed., London: Springer, 2010.
- [5] J. M. Jerez, I. Molina, J. P. Garcí'a-Laencina *et al.*, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105-115, 2010.
- [6] J. L. Peugh, and C. K. Enders, "Missing data in educational research: A review of reporting practices and suggestions for improvement," *Review of Educational Research*, vol. 74, pp. 525-556, 2004.
- [7] S.-R. R. Esther-Lydia, Pino-Mejias.Manuel, Lopez-Coello.Maria-Dolores, Cubiles-de-la-Vega., "Missing value imputation on missing completely at random data using multilayer perceptrons," *Neural Networks*, no. 1, 2011.
- [8] S. M. Weiss, and N. Indurkha, "Decision-rule solutions for data mining with missing values," *In: IBERAMIA-SBIA*, pp. 1-10, 2000.
- [9] L. Pawan, Z. Ming, and t. S. Sa, *Evolutionary Regression and Neural Imputations of Missing Values*, p.^pp. 151-163, London: Springer.
- [10] N. A. Setiawan, P. Venkatachalam, and A. F. M. Hani, "Missing Attribute Value Prediction Based on Artificial Neural Network and Rough Set Theory." pp. 360-310.
- [11] Q. Yun-fei, Z. Xin-yan, L. Xue *et al.*, "Research on the missing attribute value data-oriented for decision tree," in 2nd International Conference on Signal Processing Systems (ICSPS) 2010, 2010.
- [12] P. Meesad, and K. Hengprapromh, "Combination of KNN-Based Feature Selection and KNNBased Missing-Value Imputation of Microarray Data," in *3rd International Conference on Innovative Computing Information and Control, 2008. ICICIC '08.*, 2008, pp. 341.
- [13] L. Wang, and D.-M. Fu, "Estimation of Missing Values Using a Weighted K-Nearest Neighbors Algorithm." pp. 660-663.
- [14] H. Heikki Junninen, Niska.Kari, Tuppurainen.Juhani, Ruuskanen.Mikko, Kolehmainen., "Methods for imputation of missing values in air quality data sets," *Atmospheric Environment*, vol. 38, no. 18, pp. 1352-2310, 2004.
- [15] J. Hühn, and E. Hüllermeier, "Fuzzy Unordered Rules Induction Algorithm," *Data Mining and Knowledge Discovery*, vol. 19, no. 3, pp. 293-319, 2009.
- [16] F. Lotte, A. Lecuyer, and B. Arnaldi, "FuRIA: A Novel Feature Extraction Algorithm for Brain-Computer Interfaces using Inverse Models and Fuzzy Regions of Interest," in 3rd International IEEE/EMBS Conference on Neural Engineering, CNE '07 2007, pp. 175-178.
- [17] F. Lotte, A. Lecuyer, and B. Arnaldi, "FuRIA: An Inverse Solution Based Feature Extraction Algorithm Using Fuzzy Set Theory for Brain-Computer Interfaces," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 3253-3263, 2009.
- [18] D. N. Davis, and T. T. T. Nguyen, "Generating and Verifying Risk Prediction Models Using Data Mining (A Case Study from Cardiovascular Medicine)," in European Society for Cardiovascular Surgery 57th Annual Congress of ESCVS, Barcelona Spain, 2008.
- [19] C. Marsala, "A fuzzy decision tree based approach to characterize medical data," in IEEE International Conference on Fuzzy Systems, 2009, pp. 1332-1337.
- [20] V. Devendran, T. Hemalatha, and W. Amitabh, "Texture based Scene Categorization using Artificial Neural Networks and Support Vector Machines: A Comparative Study," *ICGST-GVIP*, vol. 8, no. 5, 2008.
- [21] R. Brian, Gaines., and C. Paul, "Induction of Ripple-Down rules applied to modelling large databases," *Journal of Intelligent information system*, vol. 5, no. 3, pp. 221-228, 1995.
- [22] T. T. T. Nguyen, "Predicting Cardiovascular Risks using Pattern Recognition and Data Mining," Department of Computer Science, The University of Hull, Hull, UK, 2009.
- [23] T. T. T. Nguyen, and D. N. Davis, "A Clustering Algorithm For Predicting CardioVascular Risk," in International Conference of Data Mining and Knowledge Engineering, London, 2007.