A Comparative Study of Missing Value Imputation with Multiclass Classification for Clinical Heart Failure Data

Y. Zhang, C. Kambhampati, D. N. Davis
DRIS,Department of Computer Science,
University of Hull,
Hull, United Kingdom

K. Goode, J. G. F. Cleland
Department of Cardiology,
HYMS, University of Hull,
Hull, United Kingdom

***Abstract*— Clinical data often contains missing values. Imputation is one of the best known schemes to overcome the drawbacks associated with missing values in data mining tasks. In this work, we compared several imputation methods and analyzed their performance when applied to different classification algorithms. A clinical heart failure data set was used in these experiments. The results showed that there is no universal imputation method that performs best for all classifiers. Some imputation-classification combinations are recommended for the processing of clinical heart failure data.**

Keywords - Missing value; Imputation; Classification; Clinical data; Heart failure

I. INTRODUCTION

Real-life data sets often contain Missing Values (MVs). There are many reasons for the existence of MVs, especially when manual data entry procedures are in place. For clinical data collected as part of a clinical trial, the medical report pro-forma allows certain attributes to be left blank. This may be, because they are not considered appropriate for the class of illness being treated or because the patient may not wish certain information to be recorded about them (e.g. metrics of anxiety or depression).

MVs make the performance of data analysis difficult. The presence of MVs usually requires a pre-processing step before moving on to the knowledge extraction process. Inappropriate handling of the MVs can result in misleading or inappropriate conclusions being drawn. The following three types of problems are usually associated with MVs during data mining: loss of efficiency; complication in data handling and analysis; and bias resulting from differences between missing and complete data [1]. In the particular case of classification, incomplete data in either the training set or test set or in both sets affect the prediction accuracy. The seriousness of this problem depends in part on the proportion of MVs.

The research reported in this paper aimed to analyse the use of a set of imputation methods on various classifiers. As a database, a clinical Heart Failure (HF) dataset was used. The results will help us to explain how imputation may be a useful tool to overcome the negative impact of MVs and to choose the most suitable imputation-classification combination for clinical HF data.

The rest of the paper is organized as follows. In section 2, a brief introduction to the clinical HF dataset is presented. Imputation methods and classification models used in the experiments are described in section 3 and 4 separately. Section 5 contains the details of the experiments. Conclusions are drawn in section 6.

II. CLINICAL HEART FAILURE DATA

Data used in this research is obtained from the Hull-Lifelab clinical database (University of Hull, UK). Hull-LifeLab provides both longitudinal and horizontal data on (probably) the largest epidemiologically representative cohort of patients with suspected or confirmed heart failure in the world. Since 2000, more than 7,500 patients have been evaluated. The Hull-Lifelab dataset is composed of more than 100 tables and thousands of physiologic and symptomatic variables. Some tables in Hull-Lifelab contain tens of thousands of records, e.g., there are more than 50,000 records in the blood table. As a large clinical dataset, Hull-Lifelab presents significant challenges for data mining such as; noisy data, missing values, diverse clinical features, varied value scales and high dimensionality.

In order to develop a predictive model for the management of patients with HF, missing value imputation and feature selection are two key pre-processing steps. A study on feature selection for clinical HF dataset has already been performed in our research group [2]; the current paper focuses on missing value imputation. As has been pointed out by medical experts, no matter which feature selection approach is applied to Lifelab, there are four significant biomarkers (creatinine, sodium, urea, and NT-proBNP) which should always be included in the analysis. In this comparative study, we chose these four biomarkers as input variables. The survival period of a patient after a blood test was selected to be the predicted

978-1-4673-0024-7/10/$26.00 ©2012 IEEE

class attribute C. There are six possible categorical values $\{c1, c2, c3, c4, c5, c6\}$ for this class attribute; death within 6 months; 12 months; 18 months; 24 months' 36 months; or after 36 months.

III. MISSING VALUE IMPUTATION METHODS

Due to the computational overheads of multiple imputation schema, and the assumptions they make regarding data distribution and MV randomness (i.e. we should know the underlying distributions of the complete data and the proportion of MVs prior to their applications), we will use single imputation methods in this study. From another point of view, Gheyas and Smith [3] indicated that the single imputation methods were able to show better prediction capabilities than multiple imputation ones for a wide range of data sets due to their lower over-fitting responses.

Seven different imputation methods were used in this comparative study: Imputation with Most Common value (MCI); Imputation with Concept Most Common value (CMCI); Imputation with K-Nearest Neighbor (KNNI); K-means Clustering Imputation (KMI); Imputation with Fuzzy K-means Clustering (FKMI); Expectation-Maximization Imputation (EMI); and Support Vector Machine Imputation (SVMI). MCI, KNNI and EMI are commonly used imputation methods. CMCI is an extension of MCI. FKMI and SVMI are recommended as best imputation methods by Luengo et al [4]. KMI is selected as a contrast to FKMI. A brief description of the imputation methods is given below;

- MCI uses a relatively simple algorithm: For nominal attributes, the MV is replaced with the most common value of the attribute. For numerical attributes, the MV is replaced with the average value of the attribute [5].
- CMCI is similar to MCI, but replaces the MV by the mode if nominal or the mean value if numerical, but considers only the instances with the same class as the reference instance [5].
- KNNI is an instance-based algorithm. Every time a MV is found in a current instance, KNNI computes the K nearest neighbours and a value from them is imputed. For nominal values, the most common value among all neighbours is taken, and for numerical values, the average value is used [6]. A proximity measure between instances should be defined. The Euclidean distance is most commonly used in the literature.
- KMI measures the intra-cluster dissimilarity by the addition of distances among the objects and the centroid of the cluster to which they are assigned. A cluster centroid represents the mean value of the objects in the cluster. Once the clusters have converged, the last process is to fill in all the non-reference attributes for each incomplete object based on the cluster information. Data objects that belong to the same cluster are taken to be nearest neighbors of each other, and KMI applies a nearest neighbor algorithm to replace MVs, in a similar way to KNNI [7].
- FKMI differs from KMI in the degree to which a data object belongs to a certain cluster. In FKMI, a data object cannot be assigned to a concrete cluster represented by a cluster centroid (as is done in the basic K-mean clustering algorithm), because each data object belongs to all K clusters with different membership degrees. FKMI replaces non-reference attributes for each incomplete data object based on the information about membership degrees and the values of cluster centroids [7].
- EMI iteratively computes the expected values for missing observations by repeatedly updating maximum-likelihood parameter estimates and imputing updated expected values until convergence is achieved [8].
- SVMI uses an SVM (Support Vector Machines) regression-based algorithm to fill in MVs. It sets the decision attributes (output or class) as the condition attributes (input attributes) and the condition attributes as the decision attributes, then SVM regression can be used to predict the missing condition attribute values [9].

IV. CLASSIFICATION MODELS

The effect of the employment of various imputation methods on the clinical HF data was analysed using different classifiers. Four classification methods were applied for building classification models: Naive Bayes (NB); K-Nearest Neighbour (KNN); Decision Tree (J48); and Neural Networks using Multi-Layer Perceptron (MLP). Each classifier is based on a different principle: NB is one of the most successful learning algorithms for text categorization; KNN is a nonparametric classifier; J48 is a rule induction classifier; MLP belongs to the group of artificial neural network based classifiers, and artificial neural networks was the most commonly used analytical tool in medicine [10]. A brief description of each of these classifiers is given below;

- Naive Bayes utilizes a probabilistic method for classification by multiplying the individual probabilities of every attribute-class pair [11]. This simple algorithm assumes independence among the attributes.
- K-Nearest Neighbour uses an integer parameter, K. Given an input x, the algorithms finds the K closest training data points to x, and predicts the class of x based on the class of the K points [12].
- Decision Tree builds a binary classification tree. Each node corresponds to a binary predicate on one attribute. Each leaf is labelled by a class. To predict the class label of an input, a path to a leaf from the root is found depending on the value of the predicate at each node that is visited. In our experiments, we

used the J4.8 version of the decision tree algorithm, which is implemented in WEKA [13].

- The Multi-Layer Perceptron consists of input layer (attributes), output layer (classes) and hidden layer(s) that are interconnected through various neurons. The back propagation algorithm tends to optimize the weights of these connections through training instances of the dataset [14].

V. Experiment Results

The main purpose of these experiments was to compare the performance of different missing value imputation methods on the clinical HF data. Four classifiers used for verification purpose are implemented in WEKA and default parameters (as shown in table 1) for each classifier were applied. Performance of a particular imputation-classification combination was measured using Precision and Recall:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \quad (2)$$

Where TP is True Positive, FP is False Positive, and FN presents False Negative.

A group of 20,003 blood test instances corresponding to patients present in the mortality tables were derived from the Hull-Lifelab dataset. Four input attributes (creatine, sodium, urea, NT-proBNP) and one class attribute were included in each instance. The missing value percentage of each input attribute is listed in table 2. The distribution of the class attribute is illustrated in figure 1.

Results of the experiments are given in table 3, where the best classifier for each imputation method is highlighted in bold font. When examining the results for each class value, the predictions for death at 36m and 6m were more accurate than the ones for 18m and 24m. This is probably due to the uneven frequency of the class attribute caused by case-wise delete in patients that did not achieve at least 12, 18, 24 or 36m of follow-up respectively.

When observing the results for each imputation method, we found that KNN classifier perform best for MCI, KMI and FKMI; the decision tree classifier was the best for CMCI, KNNI, and SVMI; the MLP classifier works best for EMI. However, Naïve Bayes classifier was not the optimal classifier for any of the imputation methods. This may reveal that the attribute independence assumption (of NB) was not true for our clinical HF data.

When comparing all the classifiers together; the CMCI method obtained the best average precision and recall. The EMI and SVMI methods had similar average precision and recall, and they were not far from CMCI. Average precision and recall obtained by the other four imputation methods were much lower.

Considering the imputation-classification combination; the CMCI-Decision tree, SVMI-Decision tree, and EMI-MLP were the best combinations for the processing of the clinical HF data.

TABLE 1 CLASSIFIER PARAMETERS

Methods	Parameters
NB	Use supervised discretization = false
KNN	K = 1 Distance function = euclidean
Decision Tree	Prune = true Confidence factor = 0.25 minimum number of instances per leaf = 2
MLP	Number of epochs = 500 Learning rate = 0.3 Momentum of updating weights = 0.2

TABLE 2 MISSING VALUE OF EACH INPUT ATTRIBUTE

Feature	Creatine	Sodium	Urea	NT-proBNP
MV percentage	22%	6%	1%	78%

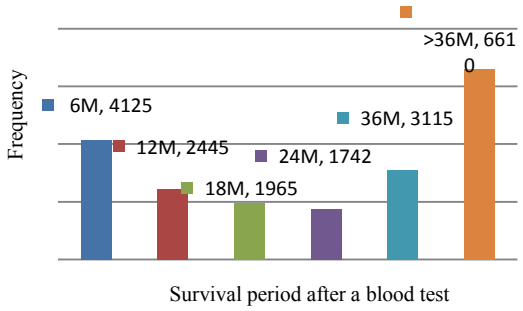

FIGURE 1 DISTRIBUTION OF THE CLASS ATTRIBUTE

VI. Conclusion

The systematic collection of health information about a population in electronic form has resulted in the availability of large clinical datasets. This data often contains unbalanced attributes in which one attribute is represented by large number of samples while others are presented by a few numbers, and a great deal of missing values. These present a unique set of problems for developing appropriate classification/prediction models.

This paper aimed to investigate performances of different missing value imputation methods when they are applied to pre-processing for multi-class classification. We attempted to understand and find the most suitable imputation approach for the development of predictive models for the management of patients with heart failure. Based on the experimental results, the following conclusions can be drawn:

TABLE 3 EXPERIMENTAL RESULTS

Imputation methods	Class	Naïve Bayes		KNN		Decision tree		MLP	
		Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Most Common Imputation	6M	0.379	0.272	**0.306**	**0.343**	0.343	0.41	0.379	0.398
	12M	0.094	0.004	**0.163**	**0.167**	0.183	0.129	0.212	0.02
	18M	0	0	**0.135**	**0.134**	0.112	0.066	0.5	0.001
	24M	0	0	**0.097**	**0.089**	0.09	0.045	0	0
	36M	0.182	0.001	**0.199**	**0.191**	0.187	0.122	0	0
	>36M	0.355	0.907	**0.402**	**0.385**	0.394	0.554	0.373	0.872
Concept Most Common Imputation	6M	0.621	0.673	0.863	0.883	**0.903**	**0.9**	0.879	0.917
	12M	0.111	0.063	0.585	0.593	**0.822**	**0.817**	0.759	0.678
	18M	0.125	0.001	0.512	0.474	**0.84**	**0.808**	0.736	0.642
	24M	0	0	0.416	0.369	**0.844**	**0.777**	0.87	0.628
	36M	0.041	0.003	0.554	0.565	**0.839**	**0.802**	0.844	0.781
	>36M	0.438	0.924	0.779	0.796	**0.841**	**0.889**	0.808	0.937
KNN Imputation	6M	0.379	0.363	0.311	0.35	**0.315**	**0.348**	0.374	0.407
	12M	0.125	0	0.158	0.164	**0.151**	**0.132**	0.147	0.002
	18M	0	0	0.131	0.13	**0.128**	**0.099**	0	0
	24M	0	0	0.107	0.1	**0.113**	**0.074**	0	0
	36M	0	0	0.186	0.175	**0.179**	**0.135**	0	0
	>36M	0.364	0.884	0.39	0.373	**0.384**	**0.482**	0.371	0.869
KMeans Imputation	6M	0.383	0.344	**0.307**	**0.343**	0.334	0.379	0.386	0.415
	12M	0	0	**0.16**	**0.166**	0.158	0.118	0.5	0.001
	18M	0	0	**0.131**	**0.131**	0.129	0.084	0	0
	24M	0	0	**0.109**	**0.103**	0.082	0.046	0	0
	36M	0	0	**0.193**	**0.182**	0.173	0.115	0	0
	>36M	0.36	0.888	**0.391**	**0.372**	0.39	0.541	0.37	0.87
Fuzzy KMeans Imputation	6M	0.344	0.269	**0.309**	**0.342**	0.329	0.371	0.375	0.411
	12M	0.141	0.006	**0.168**	**0.175**	0.162	0.127	0.222	0.01
	18M	0	0	**0.138**	**0.137**	0.128	0.087	0.5	0.001
	24M	0	0	**0.1**	**0.095**	0.093	0.058	0	0
	36M	0.316	0.002	**0.204**	**0.195**	0.195	0.143	0	0
	>36M	0.355	0.895	**0.394**	**0.375**	0.389	0.512	0.373	0.867
Expectation Maximization Imputation	6M	0.381	0.28	0.664	0.675	0.78	0.798	**0.76**	**0.845**
	12M	0.197	0.451	0.446	0.442	0.648	0.621	**0.828**	**0.671**
	18M	0.087	0.268	0.446	0.434	0.642	0.612	**0.772**	**0.702**
	24M	0	0	0.431	0.404	0.635	0.591	**0.736**	**0.649**
	36M	0.536	0.169	0.666	0.654	0.777	0.755	**0.848**	**0.786**
	>36M	0.603	0.549	0.808	0.824	0.851	0.881	**0.844**	**0.91**
Support Vector Machines Imputation	6M	0.425	0.339	0.71	0.717	**0.896**	**0.9**	0.778	0.857
	12M	0.101	0.007	0.407	0.408	**0.832**	**0.806**	0.62	0.573
	18M	0.5	0.001	0.435	0.415	**0.84**	**0.807**	0.643	0.651
	24M	0	0	0.308	0.296	**0.826**	**0.781**	0	0
	36M	0.208	0.004	0.473	0.475	**0.844**	**0.805**	0.58	0.798
	>36M	0.363	0.906	0.739	0.751	**0.845**	**0.892**	0.852	0.892

- Missing values are a major issue which affects the classification process. There is no universal imputation method that performs best for all classifier.
- Handling missing values using appropriate techniques is significant in data mining processes. It is vital to impute missing values with methods specific for a particular dataset. Three imputation-classification combinations: CMCI-Decision tree, SVMI-Decision tree, and EMI-MLP, are recommended for the processing of clinical HF data.

In this study, only the four biomarkers suggested by medical experts were used as input attributes. Finding out other significant attributes from the mountain of variables with suitable approaches and including them in the model may considerably improve the prediction accuracy. Furthermore, the results may be improved by tuning the parameters used in the classification methods

Future work will also be conducted for a deeper understanding of the results from a theoretical point of view.

REFERENCE

[1] H. Wang and S. Wang, "Mining incomplete survey data through classification," Knowl Inf Syst, vol. 24, pp. 221-233, 2010.

[2] N. Poolsawad, C. Kambhampati, and J. G. Cleland, "Feature Selection Approaches with Missing Values Handling for Data Mining - A Case Study of Heart Failure Dataset," in World Academy of Science, Engineering and Technology, 2011, pp. 828-837.

[3] I. Gheyas and L. Smith, "A neural network-based framework for the reconstruction of incomplete data sets. ," Neurocomputing, vol. 73, pp. 3039-3065, 2010.

[4] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," Knowledge and Information Systems, 2011.

[5] J. Grzymala-Busse, L. Goodwin, W. Grzymala-Busse, and X. Zheng, "Handling missing attribute values in preterm birth data sets," presented at the Proceedings of 10th international conference of rough sets and fuzzy sets and data mining and granular computing(RSFDGrC), 2005.

[6] G. Batista and M. Monard, " An analysis of four missing data treatment methods for supervised learning," Appl Artif Intell, vol. 17, pp. 519-533, 2003.

[7] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towardsmissing data imputation: a study of fuzzy k-means clustering method," presented at the Proceedings of 4th international conference of rough sets and current trends in computing (RSCTC), 2004.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, " Maximum likelihoodfrom incomplete data via the EM algorithm (with discussion)," Journal of the Royal Statistical Society (B) vol. 39, pp. 1-38, 1977.

[9] H. Feng, C. Guoshun, Y. Cheng, B. Yang, and Y. Chen, "A svm regression based approach to filling in missing values," in lecture notes in computer science. vol. 3683 R. Khosla, R. Howlett, and L. Jain, Eds., ed Berlin: Springer, 2005, pp. 581-587.

[10] A. Ramesh, C. Kambhampati, J. Monson, and P. Drew, "Artificial Intelligence in Medicine," Ann R Coll Surg Engl, vol. 86, pp. 334-8, 2004.

[11] I. Rish, "An empirical study of the naive Bayes classifier," presented at the IJCAI Workshop on Empirical Methods in Artifical Intelligence, 2001.

[12] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory vol. 13 pp. 21-27, 1967.

[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, 2009.

[14] S. Haykin, Neural networks: a comprehensive foundation, 2nd ed. London: Pearson Education, 1998.