Machine Learning-Based Missing Value Imputation Method for Clinical Datasets

M. Mostafizur Rahman and D. N. Davis

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

M. M. Rahman (✉) · D. N. Davis
Department of Computer Science, University of Hull, Hull, UK
e-mail: mmrbappy@gmail.com

D. N. Davis
e-mail: D.N.Davis@hull.ac.uk

M. M. Rahman
Department of Computer Science, Eastern University Dhaka, Dhaka, Bangladesh
e-mail: M.M.Rahman@2009.hull.ac.uk

G.-C. Yang et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 229, DOI: 10.1007/978-94-007-6190-2_19,

1 Introduction

Many researchers have identified several important and challenging issues [1–3] for clinical decision support. In "Grand challenges for decision support" Sittig et al. [1] setout ten 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 $P(R|X, Z, \mu) = P(R|\mu)$. The second category of missing data is called missing at random or MAR. The MAR condition is frequently written as $P(R = r|X = x, Z = z, \mu) = P(R = r|X^{\circ} = x^{\circ},)$ for all x^{μ}, 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 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]:

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

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 = $(x_1 \ldots\ldots x_n)$ the degree of the fuzzy membership can be found using the formula [23]:

$$\mu_{r^F}(x) = \prod\nolimits_{i=1\ldots k} i_i^F(x_i) \tag{2}$$

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 [23]:

$$D_T^i = \{x = (x_{1\ldots}x_k) \in D_T^i | I_j^F(x_j) > 0 \; for \; all \; j \neq i\} \subseteq D_T \tag{3}$$

$$\mathrm{Pur} = \frac{\mathrm{p_i}}{\mathrm{p_i} + \mathrm{n_i}} \tag{4}$$

where

$$\mathrm{p_i} \overset{\mathrm{def}}{=} \sum_{\mathrm{x} \in \mathrm{D^i_{T+}}} \mu_{\mathrm{A_i}}(\mathrm{A})$$

$$\mathrm{n_i} \overset{\mathrm{def}}{=} \sum_{\mathrm{x} \in \mathrm{D^i_{T-}}} \mu_{\mathrm{A_i}}(\mathrm{A})$$

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

$$s_j(x) \overset{\mathrm{df}}{=} \sum_{i=1\ldots k} \mu_{r_i^{(j)}}(x) \cdot CF\left(r_i^{(j)}\right) \tag{5}$$

where, the certainty factor of the rule is defined as

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

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

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.

$$\mathbf{Similarity}\,(\mathbf{x},\mathbf{y}) = -\sqrt{\sum\nolimits_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{f}\,(\mathbf{x_i},\mathbf{y_i})} \tag{7}$$

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 knows as squared error function given by:

$$\mathbf{J}\,(\mathbf{v}) = \sum\nolimits_{\mathbf{i}=1}^{\mathbf{C}} \sum\nolimits_{\mathbf{j}=1}^{\mathbf{C_i}} (||\mathbf{x_i} - \mathbf{v_j}||)^2 \tag{8}$$

where, $|\left|x_i - v_j\right||$ is the Euclidean distance between x_i and v_i, 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 consider to be “Low Risk” and patients with status “Dead” are consider 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].

$$\mathbf{\bar{x}} = \frac{\mathbf{1}}{\mathbf{n}} \cdot \sum\nolimits_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i} \tag{9}$$

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 in to 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 Fig. 1).

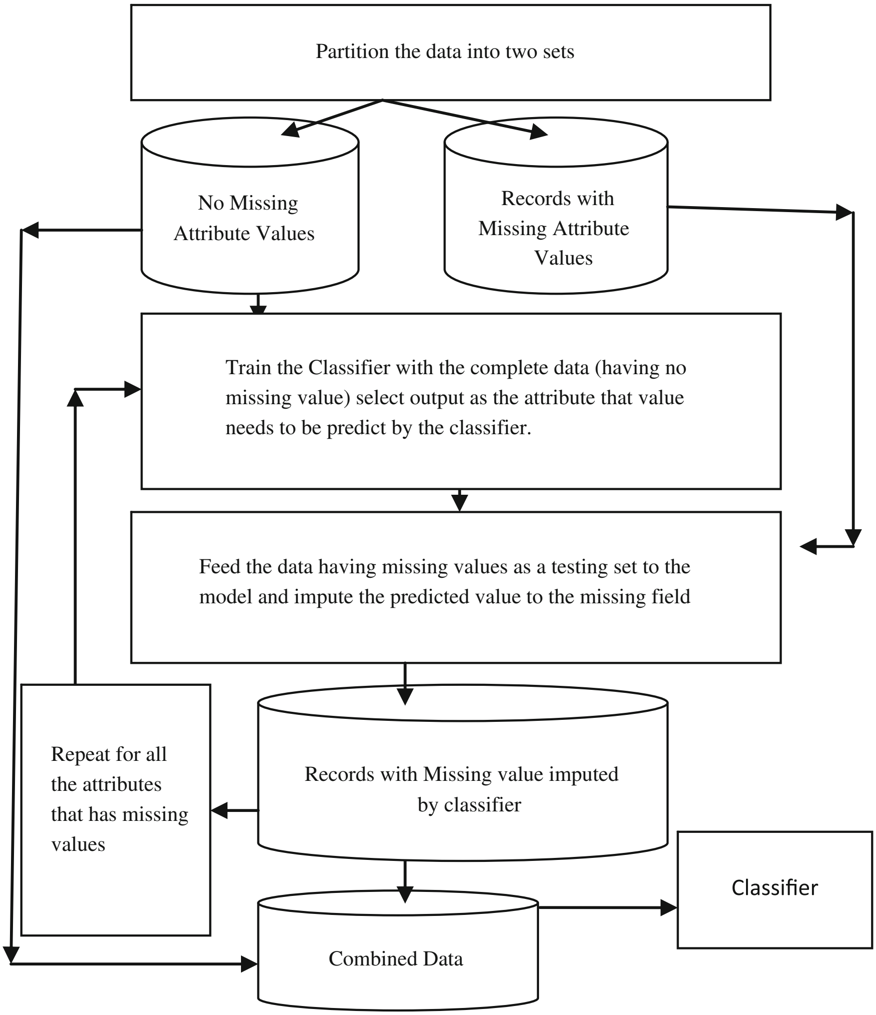
Partition the data into two sets
No Missing Attribute Values
Records with Missing Attribute Values
Train the Classifier with the complete data (having no missing value) select output as the attribute that value needs to be predict by the classifier.
Feed the data having missing values as a testing set to the model and impute the predicted value to the missing field
Repeat for all the attributes that has missing values
Records with Missing value imputed by classifier
Combined Data
Classifier

Fig. 1 Missing value imputation process

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

Missing imputation methods	Confusion matrix							
	Risk	Classified high risk	Classified low risk	ACC	SEN	SPEC	PPV	NPV
Decision tree (J48)	High	36	84	0.64	0.30	0.70	0.15	0.85
	Low	212	491					
FURIA	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					
Mean and mode	High	35	85	0.63	0.29	0.69	0.14	0.85
	Low	219	484					

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 62–63 % and sensitivity of 29–32 %. 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 Tables 3, 4, 5. Table 6 presents the highest sensitivity value found of all the datasets prepared by different imputation methods and miss-

Table 2 Comparison results with k-mix clustering

Classifier with different missing imputation methods	Confusion matrix				
	Risk	Classified high risk	Classified low risk	SEN	SPEC
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		

Table 3 Different missing imputation methods with J48 classification

Missing imputation methods	Confusion matrix							
	Actual risk ↓	Classified risk		ACC	SEN	SPEC	PPV	NPV
		High	Low					
Decision tree (J48)	High	27	93	0.80	0.23	0.90	0.27	0.87
	Low	72	631					
K-NN	High	20	100	0.80	0.17	0.90	0.23	0.86
	Low	68	635					
FURIA	High	24	96	0.80	0.20	0.90	0.25	0.87
	Low	72	631					
SVM	High	18	102	0.78	0.15	0.89	0.19	0.86
	Low	79	624					
Mean	High	13	107	0.80	0.11	0.92	0.19	0.86
	Low	56	647					

Table 4 Different missing imputation methods with K-NN classification

Missing imputation methods	Confusion matrix							
	Actual risk	Classified risk		ACC	SEN	SPEC	PPV	NPV
		High	Low					
Decision tree (J48)	High	24	96	0.71	0.20	0.80	0.15	0.85
	Low	140	563					
K-NN	High	29	91	0.81	0.24	0.91	0.32	0.88
	Low	63	640					
FURIA	High	25	95	0.79	0.21	0.89	0.24	0.87
	Low	79	624					
SVM	High	24	96	0.71	0.20	0.80	0.15	0.85
	Low	140	563					
Mean	High	25	95	0.77	0.21	0.87	0.21	0.87
	Low	92	611					

ing 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.

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 Fig. 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 5 Different missing imputation methods with Fuzzy Rule Induction Algorithm classification

Missing imputation methods	Confusion matrix							
	Actual risk	Classified risk		ACC	SEN	SPEC	PPV	NPV
		High	Low					
Decision tree (J48)	High	48	72	0.63	0.40	0.67	0.17	0.87
	Low	230	473					
K-NN	High	36	84	0.67	0.30	0.73	0.16	0.86
	Low	190	513					
FURIA	High	36	84	0.67	0.30	0.73	0.16	0.86
	Low	190	513					
SVM	High	22	98	0.74	0.18	0.83	0.16	0.86
	Low	117	586					
Mean	High	27	93	0.72	0.23	0.80	0.16	0.86
	Low	140	563					

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

Missing imputation methods	Highest sensitivity (%)	With the accuracy (%)	The classifier used to classify
FURIA	43.3	58	K-Mean
K-NN	42.5	51	K-Mean
J48	40	63	FURIA
SVM	30	62	K-Mean
Mean	29	63	K-Mean

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 $|RS|$ 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 s to build the model for each attribute of missing values.

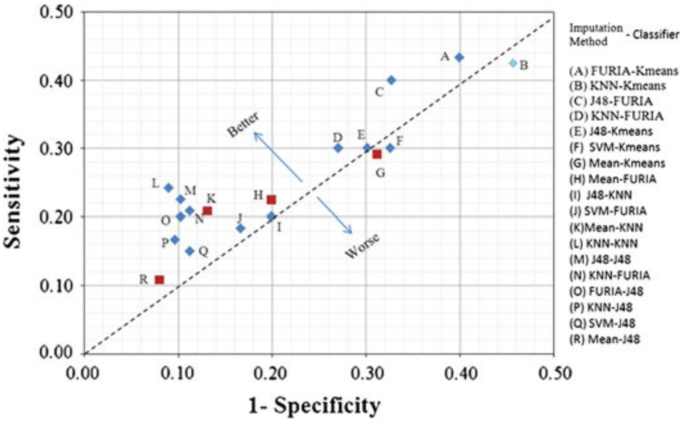

Fig. 2 The ROC space and plots of the different imputation methods classified with J48, FURIA, KNN and K-Means.

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

1. Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS et al (2008) Grand challenges in clinical decision support. J Biomed Inform 41:387–392
2. Fox J, Glasspool D, Patkar V, Austin M, Black L, South M et al (2010) Delivering clinical decision support services: there is nothing as practical as a good theory. J Biomed Inform 43:831–843
3. Bellazzi R, Zupan B (2008) Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 77:81–97
4. Dasu T, Johnson T (2003) Exploratory data mining and data cleaning. Wiley-Interscience, New York
5. Tsumoto S (2000) Problems with mining medical data. In: Computer software and applications conference, COMPSAC, pp 467–468
6. Almeida RJ, Kaymak U, Sousa JMC (2010) A new approach to dealing with missing values in data-driven fuzzy modelling. IEEE International Conference on Fuzzy Systems (FUZZ), Barcelona
7. Roderick JAL, Donald BR (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
8. Marlin BM (2008) Missing data problems in machine learning. Doctor of Philosophy, Graduate Department of Computer Science, University of Toronto, Toronto, Canada
9. Baraldi AN, Enders CK (2010) An introduction to modern missing data analyses. J Sch Psychol 48:5–37
10. Maimon O, Rokach L (2010) Data mining and knowledge discovery handbook. Springer, Berlin
11. Jerez JM, Molina I, Garcı'a-Laencina JP, Alba E, Nuria R, Miguel Mn et al (2010) Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 50:105–115
12. Peugh JL, Enders CK (2004) Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev Educ Res 74:525–556
13. Rahman MM, Davis DN (2012) Fuzzy unordered rules induction algorithm used as missing value imputation methods for K-Mean clustering on real cardiovascular data. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering (2012) London, UK, pp 391–394
14. Esther-Lydia S-RR, Pino-Mejias M, Lopez-Coello M-D, Cubiles-de-la-Vega (2011) Missing value imputation on missing completely at random data using multilayer perceptrons. Neural Networks 24:1
15. Weiss SM, Indurkhya N (2000) Decision-rule solutions for data mining with missing values. In: IBERAMIA-SBIA, pp 1–10
16. Pawan L, Ming Z, Satish S (2008) Evolutionary regression and neural imputations of missing values. Springer, London
17. Setiawan NA, Venkatachalam P, Hani AFM (2008) Missing attribute value prediction based on artificial neural network and rough set theory. In: Proceedings of the international conference on biomedical engineering and informatics, BMEI 2008, p 306–310
18. 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 on signal processing systems (ICSPS) 2010
19. Meesad P, Hengpraprohm K (2008) Combination of KNN-based feature selection and KNN based missing-value imputation of microarray data. In: Proceedings of the 3rd international conference on innovative computing information and control, ICICIC '08
20. Wang L, Fu D-M (2009) Estimation of missing values using a weighted K-nearest neighbors algorithm. In: Proceedings of the international conference on environmental science and information application technology, pp 660–663
21. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR, Verleysen M (2009) K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neuro Comput 72:1483–1493

22. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38:1352–2310
23. Hühn J, Hüllermeier E (2009) Fuzzy unordered rules induction algorithm. Data Min Knowl Disc 19:293–319
24. 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. In: Proceedings of the 3rd international IEEE/EMBS conference on neural engineering, CNE '07
25. Lotte F, Lecuyer A, Arnaldi B (2009) FURIA: An inverse solution based feature extraction algorithm using Fuzzy set theory for brain-computer interfaces. IEEE Trans Signal Process 57:3253–3263
26. Barros RC, Basgalupp MP, de Carvalho ACPLF, Freitas AA (2012) A survey of evolutionary algorithms for decision-tree induction. IEEE Trans Syst Man Cybern Part C Appl Rev 42:291–312
27. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF et al (Aug 2012) Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 36:2431–48
28. Maimon O, Rokach L (2010) Data mining and knowledge discovery handbook. Springer, Berlin
29. Quinlan JR (1985) Induction of decision trees. School of Computing Sciences, Broadway, N.S.W., Australia: New South Wales Institute of Technology
30. Quinlan JR (1993) C4.5: programs for machine learning. San Mateo: Morgan Kaufmann
31. Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P et al (2010) WEKA-Experiences with a Java open-source project. J Mach Learn Res 11:2533–2541
32. Aha DW, Kibler D, Albert MK (Jan 1991) Instance-based learning algorithms. Mach Learn 6:37–66
33. Davis DN, Nguyen TTT (2008) Generating and veriffying risk prediction models using data mining (A case study from cardiovascular medicine). Presented at the European society for cardiovascular surgery, 57th Annual congress of ESCVS, Barcelona Spain, 2008
34. Marsala C (2009) A fuzzy decision tree based approach to characterize medical data. In: Proceedings of the IEEE International Conference on Fuzzy Systems, 2009
35. Devendran V, Hemalatha T, Amitabh W (2008) Texture based scene categorization using artificial neural networks and support vector machines: a comparative study. ICGST-GVIP, vol 8. 2008
36. Nguyen TTT (2009) Predicting cardiovascular risks using pattern recognition and data mining. Ph.D., Department of Computer Science, The University of Hull, Hull, UK
37. Nguyen TTT, Davis DN (2007) A clustering algorithm for predicting cardioVascular risk. Presented at the international conference of data mining and knowledge engineering, London, 2007
38. Landgrebe TCW, Duin RPW (2008) Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. IEEE Trans Pattern Anal Mach Intell 30:810–822