Mining Class Association Rules on Dataset with Missing Data

Hoang-Lam Nguyen[1,2], Loan T. T. Nguyen[1,2](✉), and Adrianna Kozierkiewicz[3]

[1] School of Computer Science and Engineering, International University,
Ho Chi Minh City, Vietnam
ITITIU16038@student.hcmiu.edu.vn, nttloan@hcmiu.edu.vn
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Faculty of Computer Science and Management, Wroclaw University of Science and Technology, Wrocław, Poland
Adrianna.kozierkiewicz@pwr.edu.pl

Abstract. Many real-world datasets contain missing values, affecting the efficiency of many classification algorithms. However, this is an unavoidable error due to many reasons such as network problems, physical devices, etc. Some classification algorithms cannot work properly with incomplete dataset. Therefore, it is crucial to handle missing values. Imputation methods have been proven to be effective in handling missing data, thus, significantly improve classification accuracy. There are two types of imputation methods. Both have their pros and cons. Single imputation can lead to low accuracy while multiple imputation is time-consuming. One high-accuracy algorithm proposed in this paper is called Classification based on Association Rules (CARs). Classification based on CARs has been proven to yield higher accuracy compared to others. However, there is no investigation on how to mine CARs with incomplete datasets. The goal of this work is to develop an effective imputation method for mining CARs on incomplete datasets. To show the impact of each imputation method, two cases of imputation will be applied and compared in experiments.

Keywords: Missing value · Class association rules · Incomplete instance · Imputation method

1 Introduction

In the field of knowledge management, the task of machine learning and data mining often the same techniques to achieve their goal and thus, they share many common aspects [1]. Machine learning focuses on prediction using known properties obtained from training data, data mining focuses on the discovery of unknown properties in data (or often referred as knowledge discovery from databases). Data mining uses several machine learning algorithms and also, machine learning uses many data mining techniques, mostly in its preprocessing steps to improve the learner accuracy. Classification is one of the most important tasks in the field of machine learning and data mining. Two main processes of classification are training and application. In the training process, a

N. T. Nguyen et al. (Eds.): ACIIDS 2021, LNAI 12672, pp. 104–116, 2021.
https://doi.org/10.1007/978-3-030-73280-6_9

classifier is built and will be used later in the application (or test) process. In reality, there are several applications for classification, such as face [1] or fingerprint recognition [2], movie rating [3], healthcare [4], etc. Among classification algorithms, CAR results in higher accuracy compared with others. However, the main issue with CAR is that CAR can only be used effectively on complete datasets.

Unfortunately, numerous real-world datasets contain missing value. The majority of datasets in the UCI machine learning repository are incomplete. Industrial datasets might contain missing values as a result of a machine malfunction during the data collection process. In social surveys, data collection is often insufficient because respondents might refuse to answer personal questions. In the field of medical, the data can be missing since not all patients did all the given tests. Researchers cannot always collect data due to undesired conditions (for example, unsatisfactory weather conditions).

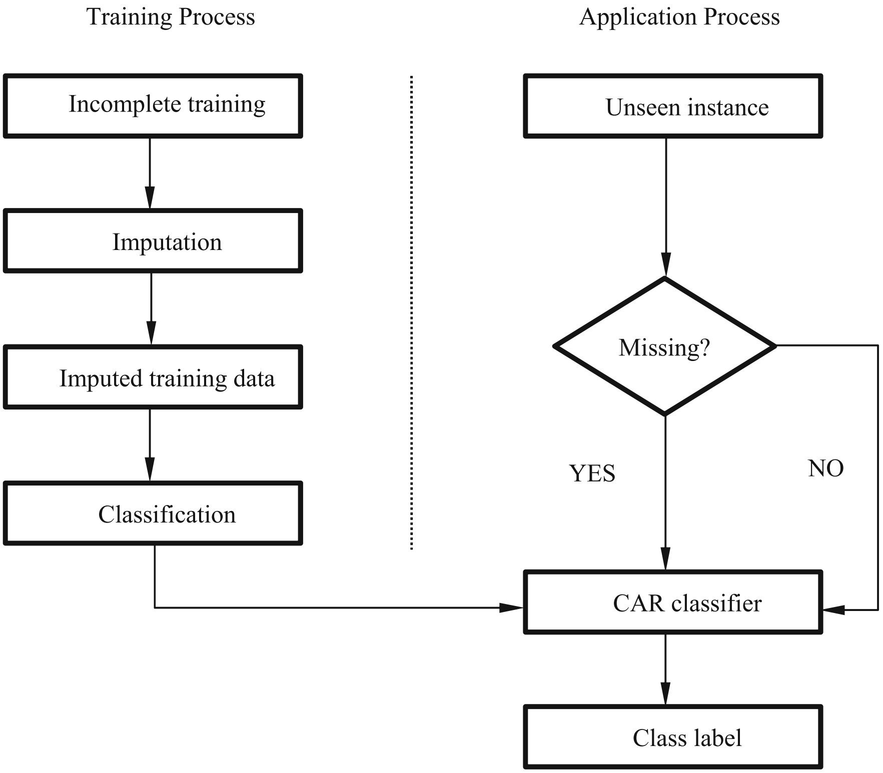

Fig. 1. Missing data handling method

One of the most common approaches to handle incomplete dataset is using an imputation method to fill the missing values. For example, mean imputation has been used commonly to process incomplete dataset [5]. The idea of mean imputation is that it replaces each missing value with the average of existing values in the features. For that reason, single imputation can be applied to almost any dataset. It (such as mean imputation) is very efficient but the accuracy is inadequate and it produced biased results [5]. In contrast, multiple imputation has been proven to be more accurate than single imputation but the major drawback is its high computational cost [6]. There are no certain ways to determine a suitable imputation method for CARs to enhance its accuracy. One of the possible methods to examine it is discussed in this article. In general, all imputation methods help generate a complete dataset which can be used to build a CAR model.

This work aims to examine and develop an effective method for classification based on association rules with incomplete dataset. Imputation methods are used to fill missing values on training datasets. Each imputation method is compared with the others for CAR classification. The main contributions of this paper are as follow:

Effectively and efficiently apply imputations with incomplete data for CARs.

Providing a comparison of imputation methods in terms of accuracy to determine what an appropriate imputation method for each of the dataset.

The rest of this paper is organized as follows. Section 2 presents a literature review of the related works. Section 3 lists important definitions, and points out how to build the CAR model with incomplete dataset. Experimental studies are conducted and presented in Sect. 4. Finally, conclusions and future work are discussed in Sect. 5.

2 Related Work

This section discusses related works and approaches to handle missing value in the incomplete datasets.

2.1 Type of Missing Data

According to Rubin [7] there are 3 types of missing data based on the mechanisms of missingness.

- Missing completely at random (MCAR) [7] is defined as when the probability of missing data on a variable is unrelated to any other measured variable and is unrelated to the variable value itself.
- Missing at random (MAR) [7] is when the missing data is related to the observed data. For example, a child does not attend the exam because he/she has a problem. We can predict this situation because data on the child's health has been previously collected.
- Another type of missing data is Missing not at random (MNAR) [7]. Missing not at random is when the missing data is related to the unobserved data. The missingness is related to the event or the factor even when the researcher does not take any measurements on it. For example, a person cannot attend a drug test because the person took the drug test the night before. MNAR is called "non-ignorable", it is necessary to use a suitable imputation method to find out what likely the value is and why it is missing.

The reason we need to thoroughly examine the mechanism is that some imputation method can only be applied to a suitable dataset. One example is that Multiple Imputation assumes the data are at least MAR. One record in the dataset can have missing data in many features and they may not have the same mechanism. Therefore, it is fundamental to investigate the mechanism of each feature before selecting a suitable approach.

Table 1. The *example of mean imputation*

Name	Age
A	17
B	22
C	23
D	?
E	25
F	27
G	?
H	32
I	32
J	35
K	?

Although the simplest way is to delete the missing value, this approach is not rational as it can result in an enormous loss of missing value and the consequence might be a decrease in classification accuracy. For that reason, the imputation methods are the most common way to handle missing values. Imputation method transforms original data with missing values into complete data before training a model. Then when it detects a new incomplete instance, it can be classified directly. The advantage of this method is that it is applicable to any classification algorithm. It can also deal with a large number of missing values.

Figure 1 shows the way to classify incomplete datasets, imputation methods are used to preprocess data. This step generates suitable values for each missing value. After that, Imputed training data can be used to build a model. After this process, new instances can be classified directly whether they are missing or not by using the CAR model built beforehand.

2.2 Imputation Method

There are two traditional imputation methods, single imputation and multiple imputation.

Single imputation means each missing value is filled with one value. Mean imputation is the most popular way to process incomplete data [5]. It fills all the missing values with the mean of the columns. Sometimes if the data are categorical, mode and median imputation should be considered. For example, if Male is 1 and Female is 0, mean imputation cannot be used as it will yield a meaningless number like 0.5. Hence, in some instance's mode and median imputation can be considered as a better option.

Mean/mode/median imputation should not be used in MNAR. These methods can handle good MCAR and MAR. Consider the dataset presented in Table 1.

For mean imputation, the mean value will be generated by using the following calculation:

$$Value = \frac{17 + 22 + 23 + 25 + 27 + 32 + 32 + 35}{11} = 19$$

The obtained dataset after performing imputation is given in Table 2 with the imputation values highlighted in bold text.

Since 32 appears twice and others appear only once. Mode imputation will replace all missing values with 32. With median imputation, it picks the element in the middle position. For example, in this dataset, the data series is numeric. The length of the series is equal to 8, the median index would be 4, and thus the median value equals 27. Median imputation will then replace all missing values with 27.

One disadvantage of mean/mode/median imputation is that it can lead to distortions in the histogram and underestimated variance because the method generates the same value for all the missing variables [5].

Table 2. After filling imputation method

Name	Age
A	17
B	22
C	23
D	**19**
E	25
F	27
G	**19**
H	32
I	32
J	35
K	**19**

K-nearest neighbors (KNN) imputation has been proved to be one of the most powerful single imputation [8]. The idea of KNN is that it searches for K-nearest neighbors to fill missing values. After that it will look for one value for the missing value by computing the average. The problem with KNN is its time-consuming nature compared to other single imputation methods (such as mean, mode, median imputation). This is because identifying *K* neighbors is required before the calculation process. The Euclidean metric is often used by KNN to determine the neighborhood.

Depending on the dataset, each single imputation method will affect the classification algorithm differently (in this paper, it is CAR). Thus, selecting a suitable imputation method is heavily based on the given dataset. Single imputation has an advantage in terms of running time over Multiple imputation. Besides KNN has been shown that it outperforms others single imputation methods.

Multiple imputations were introduced by Rubin [9]. Multiple imputation generates a set of values for one missing value as opposed to a single imputation, which only calculates one value for each missing value. Although it requires more time to calculate one value, multiple imputation produces more accurate results than single imputation [10–13].

The advantages of multiple imputation include:

- Reducing bias. Bias refers to the error that affects the analysis.
- Increasing precision, meaning how close two or more measurements to each other.
- Resistance to outliers.

Multiple imputations use Chain Equation (MICE). MICE has been used in many classification algorithms. The main idea of MICE is that it uses a regression method in order to estimate missing value. First, each missing value will be replaced by a random value in the same feature. Next, each incomplete feature is regressed on the other features to compute a better estimate for the feature. This process is repeated several times until the whole incomplete feature is imputed. Then the whole procedure is again repeated several times to provide imputed datasets. Finally, the result is calculated by the average of the imputed datasets previously.

Many studies show that MICE outperforms single imputation. MICE is a powerful imputation method. However, in reality, MICE requires long execution time in the process of estimating the missing values [14]. Therefore, further investigations are required for an effective and efficient use of this method.

2.3 Mining Class Association Rules

In 1998, Liu and partners proposed the CBA method [15] (Classification based on association) for mining class association rule. CBA includes 2 main stage:

- The stage to generate the rule – CBA-RG algorithm.
- The stage to build a classifier.

In 2001, Liu et al. proposed the CMAR algorithm CMAR (classification based on multiple association rules) [16]. This method is based on the FP–tree structure to compress the data and use projection on the tree to find association rules. In 2004, Thabtah et al. proposed the MMAC (multi-class, multi-label associative classification) [17]. In 2008, Vo and Le proposed ECR–CARM (equivalent class rule – class association rule mining) [18]. The authors have proposed the ECR tree structure, based on this tree, they presented the ECR-CARM algorithm to mine CARs in only one dataset scan. The object identifiers were used to quickly determine the support of the itemset. However, the biggest disadvantage of ECR-CARM is time-consuming for generating-and-test candidates because all itemsets are grouped into one node in the tree. When the two nodes I_i and I_j are joined to generate a new node, each element of I_i will be checked with each element of I_j to determine if their prefix is the same or not. In 2012, Nguyen et al. proposed a new method for pruning redundant rules based on lattice [19].

3 Model for Mining Class Association Rules with Incomplete Datasets

This section presents in detail on how to apply the CAR model to incomplete datasets. Furthermore, the process of applying imputation methods to improve the performance of CAR is also discussed. It describes the details of the training process and the application process.

3.1 Definition

Let $D = \{(X_i, c_i)|(i = 1, \dots m)\}$ be a dataset. X_i represents an input instance with its associated class label c_i, and m is the number of instances in the dataset. The subset of features is denoted by $F = \{F_1, ...F_n\}$. An instance X_i is represented by a vector of n values $(x_{i1}, x_{i2}, ..., x_{in})$ where an x_{ij} is a value of j feature or "?". It means the value is unknown (or is called the missing value).

An instance X_i is called an incomplete instance if and only if it contains at least one missing value. A dataset is called an incomplete dataset if it has at least one incomplete instance. A feature (F_j) is defined as an incomplete feature if it contains one incomplete instance, X_i with a missing value x_{ij}, The dataset shown in Table 3 contains 5 incomplete instances. It has 4 incomplete features.

3.2 Method

The main idea is to use imputation during the training progress, not in the application progress. The goal of applying imputation is to generate a complete dataset to improve the accuracy of the classifiers. Good imputation methods such as multiple imputation are computationally expensive. However, in the training process, there is no time limit in any application. Therefore, the use of multiple imputation in this case is acceptable.

3.3 Training Process

The training process has 2 main purposes. The first purpose is to build complete datasets. It first splits a dataset into m folds (depend on the user). It takes $m-1$ fold for the training process and the remaining fold is a test set. The process starts by using an imputation method to estimate missing values on a training dataset. It first begins with a single imputation method (KNN and mean/mode/median). The imputation will be used to generate an imputed dataset. After having complete datasets, CAR will be applied on complete dataset in order to build a classifier. A test set is used to evaluate the competency of the classifier without having any imputation on that. The whole process is repeated with a new imputation method in order to find the best methods which can lead the construction of a good classifier.

Table 3. Sample dataset

	F_1	F_2	F_3	F_4	F_5
X_1	1	?	12	18	23
X_2	2	7	13	?	?
X_3	3	8	?	19	25
X_4	4	9	?	20	26
X_5	5	?	16	21	27
X_6	6	11	17	22	28

Implementation Steps

- **Step 1:** Divide the dataset into m folds (usually m is 10). Take $m - 1$ for a training set and the last one is used as the test set.
- **Step 2:** Use imputation method on training set only. Imputation methods include single imputation (mean and *KNN* imputation)
- **Step 3:** Use a test set to evaluate the model.
- **Step 4:** Repeat all steps with different imputation methods in order to find the models with highest accuracy.

3.4 Application Process

An application process is to classify new instances using the learnt classifier without having any imputation on this. An input in an application process is an instance with some missing attributes. The algorithm will output the most suitable class label for that instance.

4 Experimental Studies

The algorithms used in the experiments were coded on a personal computer with Weka 3.8.4 using Windows 10, Intel® Core® i5 9600K (6 CPUs @ 3.7 GHz) and 16 GB RAM. The experiments were conducted on the datasets collected from UCI Machine Learning Repository. The characteristics of the experimental datasets are shown in the Table 4. The first column presents a name of the dataset. The second columns show the attribute. The third columns show the class. The fourth columns show the number of instances of each datasets. And the final column contain % missing value that the dataset has.

There are different features in the experimental datasets. Some datasets have many attributes with several instances while others have average and large one (*mushroom* dataset). Missing values can have on multiple attributes. However, *mushroom* dataset only has one attribute with missing value. The datasets also have varying types of features including real, integer and nominal. The choice of datasets is intended to reflect incomplete problems of varying difficulty, size, dimensionality, and feature types.

The experimental used two imputation method. KNN based imputation and Mean/mode/median imputation. The implement of KNN imputation choose the number of neighbors k with yield the best accuracy for the model to compare with mean/mode/median imputation. Both of the imputation methods were performed by Weka 3.8.4. Ten-fold cross validation was used to separate each dataset into different training and test sets.

We performed experiments to evaluate the effectiveness of the imputation methods. In Table 5, the first column shows the datasets, and the second column shows classification algorithm with KNN imputation. The third column shows classification algorithm with mean imputation. From Figs. 2, 3, 4, 5, 6, 7 and 8 show the details of each fold of each dataset. The blue line shows the accuracy of CAR algorithm with mean imputation. And the red line shows the accuracy of CAR algorithm with KNN imputation.

In further investigation, House vote and *mammographic masses* dataset contains some folds that mean imputation yield better result than KNN. In *house vote*, the difference seems insignificant. However, the difference between mean and KNN imputation for each fold in *mammographic_masses* can be up to 10%. Overall, In Fig. 2, the KNN imputation can increase the accuracy up to 2.8%.

Table 4. The characteristic of the experimental datasets.

Dataset	Attribute	Classes	#instances	%missing value
House vote	16	2	435	66.2
CRX	15	2	690	9.7
mammographic_masses	5	2	961	16.9
chronic_kidney_disease	24	2	400	74
Hepatitis	19	2	155	72.3
Mushroom	22	3	8124	30.5

Overall, the results, as presented from Figs. 2, 3, 4, 5, 6, 7 and 8, Table 5, have shown that KNN seems to produce higher accuracy than mean imputation. Even though some results indicate that the accuracy of both methods are the same. However, KNN performs better in most tested situations.

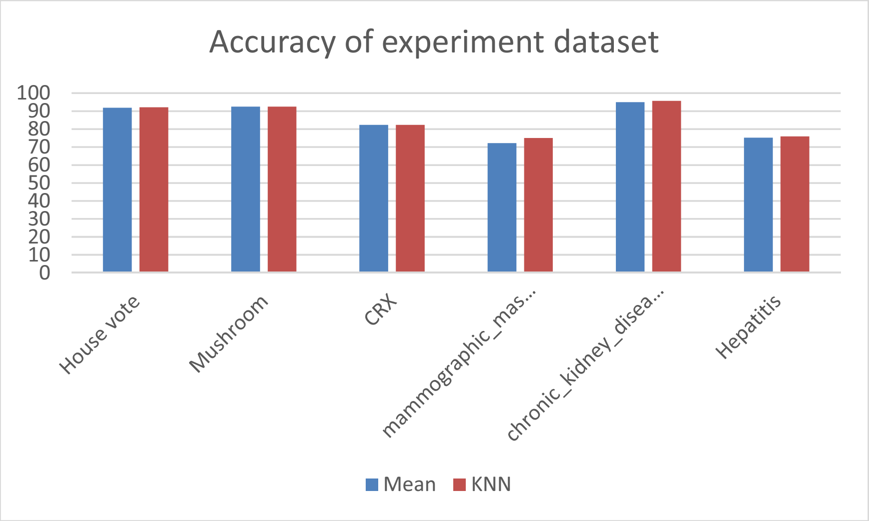

Fig. 2. Comparison of experimental dataset

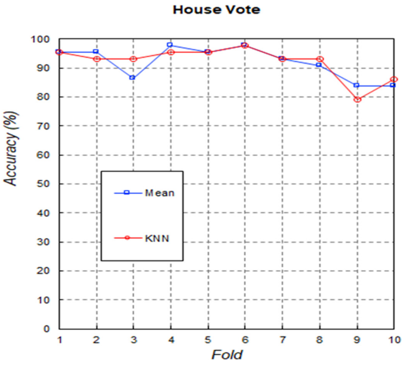

Fig. 3. Accuracy on the *House Vote*

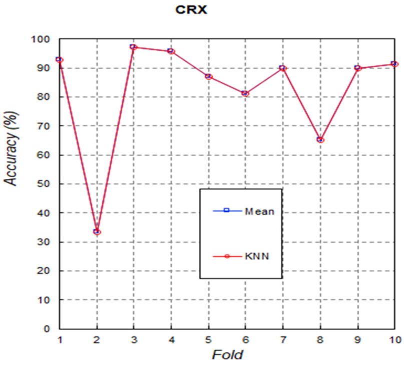

Fig. 4. Accuracy on the *CRX*

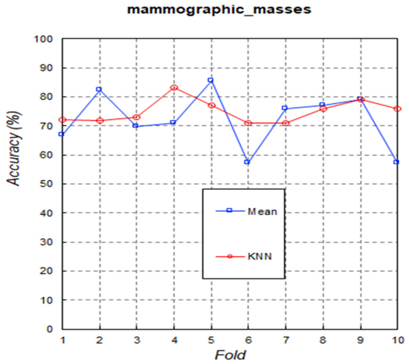

Fig. 5. Accuracy on the *mammographic_masses*

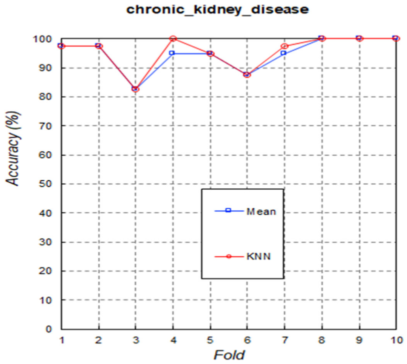

Fig. 6. Accuracy on the *chronic_kidney_diseas*

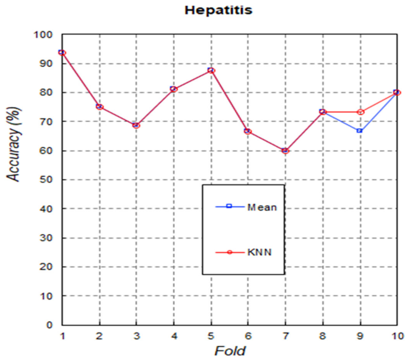

Fig. 7. Accuracy on the *hepatitis*

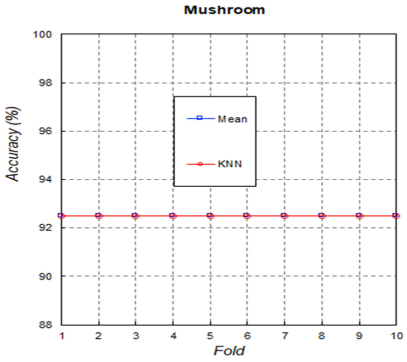

Fig. 8. Accuracy on the *mushroom*

Table 5. Result of the experimental datasets.

Dataset	KNN	Mean
House vote	92.16	91.93
CRX	82.32	82.32
mammographic_masses	75.03	72.22
chronic_kidney_disease	95.75	95
Hepatitis	75.96	75.29
Mushroom	92.49	92.49

5 Conclusion and Future Work

This paper proposed a method for mining incomplete datasets with CAR by using single imputation. In the training process, the imputation method is used to generate a complete training dataset. The experiments show the comparison between KNN and

Mean/Mode/Median imputation. The use of the KNN imputation gives a better result than Mean/Mode/Median imputation. Even though the percentage of missing value is different on each dataset, the CAR with the use KNN imputation yield higher accuracy. In addition, in some datasets, the computational time between KNN and Mean/Mode/Median imputation has no difference.

Missing values are a common issue in many datasets. In addition, mining dataset with CAR has been developed in recent years. However, there has not been much work on handling missing data in other CAR algorithms. In the future, further research with different CAR methods will be conducted on incomplete dataset.

Acknowledgements. This research is funded by International University, VNU-HCM under grant number SV2019-IT-03.

References

1. Nguyen, N.T.: Advanced Methods for Inconsistent Knowledge Management. Springer, London (2008). https://doi.org/10.1007/978-1-84628-889-0
2. Zong, W., Huang, G.B.: Face recognition based on extreme learning machine. Neurocomputing **74**(16), 2541–2551 (2011)
3. Adiraju, R.V., Masanipalli, K.K, Reddy, T.D., Pedapalli, R., Chundru, S., Panigrahy, A.K.: An extensive survey on finger and palm vein recognition system. Mater. Today Proc. (2020). https://doi.org/10.1016/j.matpr.2020.08.742
4. Wei, S., Zheng, X., Chen, D., Chen, C.: A hybrid approach for movie recommendation via tags and ratings. Electron. Commer. Res. Appl. **18**, 83–94 (2016)
5. Ahmad, M.A., Teredesai, A., Eckert, C.: Interpretable machine learning in healthcare. In: Proceedings of 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018, p. 447 (2018)
6. Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M.: Review: a gentle introduction to imputation of missing values. J. Clin. Epidemiol. **59**(10), 1087–1091 (2006)
7. Darmawan, I.G.N.: NORM software review: handling missing values with multiple imputation methods. Eval. J. Australas. **2**(1), 51–57 (2002)
8. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 3rd edn. Wiley (2019)
9. Jadhav, A., Pramod, D., Ramanathan, K.: Comparison of performance of data imputation methods for numeric dataset. Appl. Artif. Intell. **33**(10), 913–933 (2019)
10. Rubin, D.B.: An overview of multiple imputation. In: Proceedings of the Survey Research Methods Section, pp. 79–84. American Statistical Association (1988)
11. Gómez-Carracedo, M.P., Andrade, J.M., López-Mahía, P., Muniategui, S., Prada, D.: A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemom. Intell. Lab. Syst. **134**, 23–33 (2014)
12. Nguyen, N.T.: Consensus systems for conflict solving in distributed systems. Inf. Sci. **147**(1–4), 91–122 (2002)
13. Nguyen, N.T.: Using consensus methods for solving conflicts of data in distributed systems. In: Hlaváč, V., Jeffery, K.G., Wiedermann, J. (eds.) SOFSEM 2000. LNCS, vol. 1963, pp. 411–419. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-44411-4_30
14. Musil, C.M., Warner, C.B., Yobas, P.K., Jones, S.L.: A comparison of imputation techniques for handling missing data. West. J. Nurs. Res. **24**(7), 815–829 (2002)
15. Liu, B., Hsu, W., Ma, Y., Ma, B.: Integrating classification and association rule mining. In: Knowledge Discovery and Data Mining, pp. 80–86 (1998)

16. Li, W., Han, J., Pei, J.: CMAR: accurate and efficient classification based on multiple class-association rules. In: Proceedings of IEEE International Conference on Data Mining, ICDM, pp. 369–376 (2001)
17. Thabtah, F.A., Cowling, P., Peng, Y.: MMAC: a new multi-class, multi-label associative classification approach. In: Proceedings of Fourth IEEE International Conference on Data Mining, ICDM 2004, pp. 217–224 (2004)
18. Vo, B., Le, B.: A novel classification algorithm based on association rules mining. In: Richards, D., Kang, B.-H. (eds.) PKAW 2008. LNCS (LNAI), vol. 5465, pp. 61–75. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01715-5_6
19. Nguyen, L.T.T., Vo, B., Hong, T.P., Thanh, H.C.: Classification based on association rules: a lattice-based approach. Expert Syst. Appl. **39**(13), 11357–11366 (2012)