Techniques to Deal with Missing Data

Jadran Sessa
Department of Electrical Engineering
and Computer Science
Masdar Institute of Science
and Technology, Abu Dhabi, UAE
Email: jsessa@masdar.ac.ae

Dabeeruddin Syed
Department of Electrical Engineering
and Computer Science
Masdar Institute of Science
and Technology, Abu Dhabi, UAE
Email: syeddabeeruddin56@gmail.com

***Abstract*—Data is available to us in humongous amounts in the real world, but none of it is of practical use if not converted to useful information. However, the knowledge discovery is hindered because the real data is often incomplete and noisy. Nowadays, the problem of recovering missing data has found most important place in the field of data mining. Filling the missing data is a significant task, as it is paramount to use all available data for the given datasets are generally very small. In this paper, we deal with the real data with many missing values. Furthermore, we deal with the given data in three phases. The first phase considers the concept of feature selection, while the second phase iteratively considers filling in the missing values using probabilistic approach, keeping in mind the fact that features can be either nominal or numerical. Finally, the third phase deals with correcting the missing values that have been filled in. In our work, we have compared two imputation methods for dealing with the missing data, namely k-NN imputation method and mean and median imputation method. As a result, we have found that both of the imputation methods are efficient and yield more or less the same accuracy.**

Keywords*—Data mining, Missing data, Missing values, Probabilistic approach, k-NN imputation, Mean and Median imputation.

I. INTRODUCTION

For competitive advantage, all of the enterprises or organisations use knowledge discovery techniques to browse through the raw data available. This is done to analyze and understand the patterns and draw useful information from the raw data and here is where the data mining and the machine learning subjects come into picture.

In real life, the data is noisy and incomplete. Hence, data preparation is often the most critical part of data mining, and even the slightest fine-tuning can significantly increase the accuracy of prediction and classification. It is stated that data preparation requires as much as 80% of the engineering effort[14].

A vast number of real-life datasets contain missing values owing to either accidental or purposeful omission or due to the fact that they were not available at the time of insertion. The attributes with missing values are sometimes entitled as lost and denoted by sign "?" and sometimes named as "do not care" and marked as "*", where the participants are hesitant to provide the relevant information.

The dataset we worked on, only contained the first category of missing values, i.e. lost values. Process of replacing missing values is a daunting task, which has almost always to be faced in data mining.

One of the approaches to deal with the missing data is to ignore the instances which contain the missing data [11]. The disadvantage here is that precious data records are lost. The other approach is feature selection, i.e. deletion of the features that contain the missing values; however, the point to be noted is that the feature can be deleted only if it is non-significant.

Finally, the most reliable approach to deal with the missing data is the imputation of missing values and the method chosen depends on the amount of the data missing and the type of variable, i.e. for missing categorical data, mode is used instead of the mean or median. No matter what imputation method is chosen, there will be still some bias associated with the filling of the missing data and the bias can be reduced by using multiple imputation [9].

In our work, we have used smaller datasets. Owing to the size of datasets, each record is considered to be very valuable for the knowledge discovery. Therefore, it was not good idea to delete any of the instances containing missing values. When it comes to the feature selection, almost all of features were of the equal relevance, and hence they had to be preserved. Moreover, given features were of the two types, either nominal or numerical. We have used two different imputation methods, namely k-NN imputation and mean and median imputation approach. For verifying the accuracy of the preprocessed dataset, 10-fold cross validation with four different classification algorithms (Naïve Bayes, weighted k-NN, unweighted k-NN and j48 decision trees) was used, and obtained results have revealed that both methods performed well.

II. BACKGROUND WORK

A. Classification algorithms

1) Naïve Bayes Algorithm: The simplest classification algorithm is Naïve Bayes algorithm that outperforms the most sophisticated algorithms despite its simplicity [3]. Naïve Bayes algorithm is based on the concept of maximum aposteriori and the Bayes theorem of probability.

2) k-NN Algorithm: k-Nearest Neighbours algorithm considers the classification of a new object based on the k neighbours which are closest to the new point [5]. To obtain the correct prediction of classes, the value of k is to be carefully selected as high k values provides poor resolution, while low values have a tendency of poor noise resistance.

978-1-5090-5306-3/16/$31.00 ©2016 IEEE

B. *Data Pre-processing*

Data pre-processing is one of the most crucial steps in the process of data mining as this is what defines the quality of the classification model.

Data pre-processing includes some or all of the following steps: [8]

1) Data Cleaning
2) Data Integration
3) Data Transformation
4) Data Reduction

III. Dealing with the Missing Data

A. *Types of Missing Data*

The most crucial part of the data pre-processing is to deal with the missing data, which may be prevalent due to one of the following reasons:

1) Missingness completely at random: It can be said that data is missing completely at random if the probability for all missing values is equal.

2) Missingness at random: More likely, the probability that values are missing depends on available information. In this case, process can be modeled as logistic regression with outcome 1 for available and 0 for missing data. Moreover, missing data can be treated as NA, i.e. excluded with the condition that all missing variables are controlled by regression.

3) Missingness that depends on unobserved predictors: Data is no longer considered to be randomly missing if it is dependent on information that have not been recorded.

B. *Approaches to deal with missing values*

There are several approaches that can be used to fill in the missing data, depending on the type and sheer amount of missing data. Hence, the methods for dealing with missing values can be summarized as following:

1) Case deletion: This method involves deletion of all instances containing missing values for at least one feature. However, there are deviations of this method that consider the extent of the missing data where instances or features with high level of missing data are discarded. In case deletion, it is critical to determine the relevance of the attributes to the rest of data.

Likewise, it is important to consider the size of dataset before deletion of an instance with missing values. It is often not advised to delete relevant attributes, even if they contain a large amount of missing values, and same goes for the instances if the dataset is small. Hence, in most of the cases, the case deletion method is generally avoided, with the exception when dealing with the data missing completely at random [12] owing to the following:

1) Data is often valuable and limited, hence deleting records is wasteful.
2) Deletion of records may skew the data in cases where a particular field only applies for a specific subset of data.

2) Learning without handling missing data: This approach completely ignores and leaves the missing data in the given form. This method is deployed in the cases where the percentage of missing data is insignificant in comparison to the total size of the dataset [1].

3) Imputation: The most frequently used approach for dealing with the missing data is imputation, which substitutes missing values with meaningful replacements. There are numerous options for replacement values, among which the most popular ones include:

a) Mean imputation: This method involves replacing missing values of a given feature with the mean values of that feature for a given class. It is calculated as follows,

$$\bar{x_{ij}} = \sum_{i:x_ij \in C_k} \frac{x_{ij}}{n_k} \tag{1}$$

where n_k denotes the number of non-missing values in the j-th feature of k-th class.

Despite some of its drawbacks, such as overestimation of sample size or underestimation of variance, it often provides reasonably well results.

b) Median imputation: In the cases with skewed data distribution and outliers present, median imputation is used over mean imputation.

$$\hat{x_{ij}} = median_{i:x_{ij} \in c_k} x_{ij} \tag{2}$$

c) Mode imputation: The third measure of central tendency that can be used is mode. Mode imputation is appropriate method for missing nominal data and fairly competent method for numerical data.

d) k-NN imputation: k-NN imputation deploys the k-nearest neighbours algorithm to estimate the missing values and has the following advantages [2],

1) k-NN can predict the attributes by means of the most frequent value or the average among neighbours.
2) k-NN can be easily worked upon by using any attribute as class by just modifying which attributes will be used in the distance metric.

The only disadvantage of the k-NN imputation method is that the algorithm searches through the complete dataset, so it has limited scope when it comes to larger datasets.

There are various other less frequently used imputation methods, like Hot-deck Imputation [7], Regression Imputation [4], Multiple Imputation [9] etc.

IV. Related Work

Dealing with missing data has been a subject of research in data mining for the past few decades, and number of approaches have been proposed. Rubin [15] has warned about the mishaps of case deletion in his work and has inferred that when dealing with the real data, the statistician should consider the processes that cause the missing data and should properly felicitate the need for models for these processes.

As per the suggestions given in the work of Rubin et al., we have totally ignored the method of case deletion as the data is very valuable in our case and as deletion of rows may skew the data. Scheffer [17] thoroughly reviewed and compared a number of imputation methods for dealing with the missing data and illustrated how the mean and standard deviation are affected by different methods of imputation, given different missingness mechanisms. As per his work, we have worked on the mean imputation method for dealing with the missing data.

Quinlan et al. [16] suggested decision trees for filling the missing values, while Lakshminarayan et al. [10] handled missing values with the method of Auto clustering and also used variations of k-NN for filling in the missing values. Similarly, Olinsky et al. [13] compared the performance of five imputation methods in structural equation modelling, while Farhangfar et al. [6] examined the influence of imputation methods choice on the performance of different classifiers. Liu et al. [11] proposed Information Decomposition Imputation (IDIM) algorithm using fuzzy membership function, with the goal of solving the defined lower incomplete problem.

Acuna et al. [1] compared methods of mean imputation, median imputation, case deletion and k-NN imputation that deal with the missing values in a dataset, and found out that in datasets containing small amount of missing values there is no significant difference between case deletion and other imputation approaches. In their work, Batista et al. [2] artificially implanted missing data in three datasets and used cross validation in order to compare the efficiency of k-NN imputation method in comparison to the other algorithm based imputation methods, namely C4.5 and CN2.

V. Methodology

A. Data Collection

The dataset related to the chronic kidney disease was collected from the UCI Machine Learning Repository (archive.ics.uci.edu/ml/). Dataset consists of 25 features and 400 records with missing values denoted as '?'. The dataset taken in consideration was donated by Apollo Hospitals, India on July, 2015. The whole data is classified into two classes - chronic kidney disease (ckd) and no chronic kidney disease (notckd). Aim of this paper is to fill in the missing values and to propose and implement a classification system to solve the challenge of detecting chronic kidney diseases and ultimately, to report the accuracy of our system using 10-fold cross validation.

B. Data pre-processing

The dataset was preprocessed using Weka tool and it was observed that the data was spread across 25 features from which the last feature represents the class label. All features, except for the class label contained missing values. Considering the fact that the dataset is of only 400 records, it can be inferred that each record is precious. Thus, case deletion was not taken in consideration.

For the data preprocessing two imputation methods were implemented, namely k-NN imputation and mean, mode and median imputation.

a) Mean and Median Imputation: Even though there are a number of sophisticated and complex imputation methods for dealing with the missing data, we have decided to deploy mean, median or mode imputation for most of the given features. Those simple, yet very powerful imputation methods proved to be very effective. On the outset, the features could be noticed to be of two types i.e. numerical and nominal. For the numerical features, we have either used the concept of mean or that of median while keeping in consideration that the data is classified into two classes.

Consequently, for the numerical features, including age, blood pressure (bp), blood urea (bu), serum creatinine (sc), sodium (sod), packed cell volume (pcv) and white blood cell count (wbcc), median imputation was used. The reason behind the particular choice of median imputation over mean imputation was the number of outliers in the aforementioned features. For instance, the medical records include patients of different age groups, ranging from the toddlers of age one to the elder people of eighty years old. Similarly, other numerical records either contained noise or were skewed, so median was much better choice compared to the mean. On the other hand, for the numerical features such as red blood cells count (rbcc), blood glucose random (bgr), potassium (pot) and hemoglobin (hemo), no outliers were detected and due to the different array of values, mean imputation was appropriate method to use.

Whereas for the nominal features, the logical approach was to use the mode imputation. For the particular nominal feature of red blood cells, a large number of missing values was identified. Due to the presence of feature red blood cells count and non-matching or noisy values, missing values were marked as NA. Mode imputation, with respect to the belonging class was performed for filling the missing values for the rest of nominal values. Preprocessing part was performed with the aid of Python libraries: pandas, scikit-learn, statistics and numpy.

b) k-NN Imputation Method: We have used k-NN imputation with k=5. One of the main advantages of k-NN imputation method is that it can work equally well with both discrete and continuous attributes.Also, there is no need to create separate predictive model for each attribute in the dataset. On the other hand, performance can be sloppy when dealing with the larger datasets; however, in our case that was not an issue. Before proceeding to the imputation, we have decided to perform the feature selection, and consequently rbc (red blood cells) column was deleted due to the large number of missing values and more importantly, because it was redundant, since the feature called rbcc (red blood cells count) was already present. We have utilized the functionality of R programming language in order to perform the preprocessing using this particular method. Validation of accuracy for both methods, k-NN and mean and mode imputation was performed with the Java based WEKA tool.

VI. Experiment Results

The main goal of our analysis was to evaluate the overall performance of k-NN and mean-median imputation methods. The performance was compared using 10-fold cross validation with different classification algorithms, i.e. weighted 5-NN, j48 decision trees and Naïve Bayes algorithm.

TABLE 1
COMPARISON : 5-NN AND MEAN-MEDIAN IMPUTATION : USING WEIGHTED 5-NN : 10-FOLD CROSS VALIDATION

	k-NN	Mean-Median
Correctly Classified	397	400
Incorrectly Classified	3	0
Mean Absolute Error	0.0078	0.0003
Root Mean Sq. Error	0.0866	0.0004
Accuracy	99.25%	100%
Precision	100%	100%
Recall	98.8%	100%
F-Measure	99.4%	100%
ROC Area	1	1

TABLE 2
COMPARISON : 5-NN IMPUTATION AND MEAN-MEDIAN IMPUTATION : USING NAÏVE BAYES : 10-FOLD CROSS VALIDATION

	k-NN	Mean-Median
Correctly Classified	398	400
Incorrectly Classified	2	0
Mean Absolute Error	0.0041	0.001
Root Mean Sq. Error	0.0595	0.0193
Accuracy	99.5%	100%
Precision	100%	100%
Recall	99.2%	100%
F-Measure	99.6%	100%
ROC Area	1	1

TABLE 3
COMPARISON BETWEEN THE 5-NN IMPUTATION AND MEAN AND MEDIAN IMPUTATION METHODS USING J-48 DECISION TREES WITH 10-FOLD CROSS VALIDATION

	k-NN	Mean-Median
Correctly Classified	398	398
Incorrectly Classified	2	2
Mean Absolute Error	0.0055	0.0055
Root Mean Sq. Error	0.0702	0.0702
Accuracy	99.5%	99.5%
Precision	100%	100%
Recall	99.2%	99.2%
F-Measure	99.6%	99.6%
ROC Area	0.999	0.999

Finally, it can be clearly seen from the table 3 that both k-NN imputation and Mean and Median Imputation performed exactly the same when using j48 decision trees for 10-fold cross validation.

VII. CONCLUSION AND FUTURE WORK

In this paper, we analyzed the impact of missing data and ways of dealing with it on a small dataset. We analyzed the performance of mean-median and k-NN (k=5) imputation methods. The results of accuracy were verified using 10-fold cross validation with different classification algorithms. Both of the tested imputation methods produced promising results in all four cases. Despite the achieved performance for the given dataset, it cannot be generalized that the same results would be obtained for larger datasets. In other words, there are multiple ways of dealing with the missing data and every method has its own merits and demerits. For the smaller datasets, it is shown that both imputation methods are efficient. In the future work, the two methods should be further tested on different kinds of datasets of larger size, containing high number of missing values in order to verify if both methods can produce equally promising results.

REFERENCES

[1] E. Acuna and C. Rodriguez. The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications*, pages 639–647. Springer, 2004.

[2] G. Batista and M. Monard. K-nearest neighbour as imputation method. Technical report, Experimental Results. Tech. Report 186, ICMC-USP, 2002.

[3] K. Baumgartner, S. Ferrari, and G. Palermo. Constructing bayesian networks for criminal profiling from limited data. *Knowledge-Based Systems*, 21(7):563–572, 2008.

[4] J.-F. Beaumont. On regression imputation in the presence of nonignorable nonresponse. In *Proceeding of the Survey Research Methods Section, American Statistical Asssociation*, pages 580–585, 2000.

[5] P. S. David Hand, Heikki Mannila. Principles of data mining.

[6] A. Farhangfar, L. Kurgan, and J. Dy. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12):3692–3705, 2008.

[7] B. Ford. An overview of hot-deck procedures: Incomplete data in sample surveys, vol. 2, 1983.

[8] M. Han, Jiawei Kamber. *Data Mining : Concepts and Techniques (2nd Edition)*. Elsevier Science & Technology, March 2006.

[9] O. Harel and X.-H. Zhou. Multiple imputation: review of theory, implementation and software. *Statistics in medicine*, 26(16):3057–3077, 2007.

[10] K. Lakshminarayan, S. A. Harp, and T. Samad. Imputation of missing data in industrial databases. *Applied Intelligence*, 11(3):259–275, 1999.

[11] S. Liu and H. Dai. Examination of reliability of missing value recovery in data mining. In *Data Mining Workshop (ICDMW), 2014 IEEE International Conference on*, pages 306–313. IEEE, 2014.

[12] D. J. Mundfrom and A. Whitcomb. Imputing missing values: The effect on the accuracy of classification. 1998.

[13] A. Olinsky, S. Chen, and L. Harlow. The comparative efficacy of imputation methods for missing data in structural equation modeling. *European Journal of Operational Research*, 151(1):53–79, 2003.

[14] Y. Qin, S. Zhang, X. Zhu, J. Zhang, and C. Zhang. Pop algorithm: Kernel-based imputation to treat missing values in knowledge discovery from databases. *Expert systems with applications*, 36(2):2794–2804, 2009.

[15] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[16] S. L. Salzberg. C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993. *Machine Learning*, 16(3):235–240, 1994.

[17] J. Scheffer. Dealing with missing data. 2002.