Novel Method for Handling Missing Values in Databases based on Mean Pre-Imputation, Confidence Intervals and Boosting

Alireza Farhangfar, Lukasz Kurgan and Witold Pedrycz
University of Alberta, 2nd Floor ECERF, Edmonton, Alberta, T6G 2V4
Department of Electrical and Computer Engineering

Abstract:
One of the important issues faced by researchers utilizing industrial and research databases is incompleteness of data, usually in terms of missing or erroneous values. There are many reasons for such incompleteness, like manual data entry procedures, incorrect measurements, equipment errors, etc. While some of the algorithms can learn directly from incomplete data, a large portion of them requires complete data. Therefore, different strategies, like deletion of incomplete records, imputation (filling) of missing values through variety of statistical and machine learning (ML) procedures, are developed to fill in missing values in incomplete data.

This study introduces a new approach for missing data imputation by pre imputing the missing values with mean imputation and subsequently imputing missing values using Naïve-Bayes ML algorithm. The proposed method also applies two extensions to the basic Naïve-Bayes algorithm. First, **confidence intervals** are defined based on the frequency of the values for each attribute to filter out the least probable candidates for imputing missing values. In addition to intervals, **boosting** is also used to improve accuracy of imputation. In boosting multiple imputation iterations are performed. In each iteration predicted values that satisfy a predefined threshold for the probability computed by the Naïve-Bayes algorithm are imputed, while the remaining values are left missing. Therefore in each subsequent iteration already imputed value are used to improve imputation of the remaining missing values.

The proposed approach is characterized by linear complexity, and improvement in accuracy of imputation when compared to the approach that directly applies Naïve-Bayes based approach. To demonstrate the improvement in accuracy of imputation a comprehensive benchmark analysis is carried out. It includes a mixture of 15 synthetic and natural datasets, which accommodate for different types of missing data.

Methods for Dealing with Missing Values in Databases

In general two groups of algorithms used to preprocess databases that contain missing values can be distinguished. First group concerns unsupervised algorithms that do not use target class values. Second group are supervised algorithms that use target class values, and which are most commonly implemented by using supervised ML algorithms [5].
The unsupervised algorithms for handling missing data range from very simple methods like *Mean imputation* to statistical methods based on parameter estimation, such as *Expectation Maximization* based imputation. Several simple algorithms are described in [4].

Mean Imputation
In this method, mean of the values of an attribute that contains missing data is used to fill in the missing values. In case of a categorical attribute, the mode, which is the most frequent value, is used instead of mean [3]. The algorithm imputes missing values for each attribute separately.

Alternatively, the supervised algorithms usually use ML algorithms for imputation of missing values. Imputation is carried out by performing multiple classification tasks using a ML algorithm. Each classification task is performed in two steps. First, during the learning step the ML algorithm generates the model using learning data. The data model is used to classify examples into a set of predefined classes. Second, during the testing step, the generated model is used to impute missing data for the testing data, which was not used during learning. Figure 1 illustrates the above procedure. Several different kinds of ML algorithms, such as decision trees, probabilistic, and decision rule, can be used, but the underlying methodology remains the same.

Naïve-Bayes ML Algorithm
In this study Naïve-Bayes algorithm is selected. Naïve-Bayes is a classification technique based on computing a priori probabilities [2]. It analyzes relationship between each independent variable and the target class to derive a conditional probability for each relationship. When a new example is analyzed, a prediction is made by combining the effects of the independent variables on the target class. Naïve-Bayes requires only one pass through the training set to generate a classification model, which makes it linear and very efficient. It generates data model that consists of set of conditional probabilities, and works only with discrete data.

Mean Pre-imputation
Based on assumption that having a complete training dataset would produce a better model for the data, the proposed method pre-imputes missing values with temporary values. The values are used during the imputation procedure to be finally substituted by the imputed values. The simplest way of generating temporary values in the training dataset is to use mean pre-imputation. Mean pre-imputation does not add to the complexity of the entire method since it is also linear.

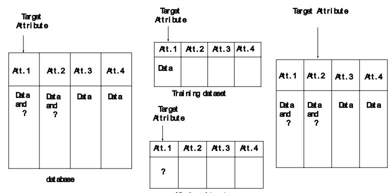

Figure 1.Supervised imputation process using a ML algorithm.

Confidence Intervals
Confidence intervals are used to filter out the least probable candidates for imputing the missing values. In order to design such a filter, the values that appear less frequently in each attribute will be filtered out. This is based on an assumption that low frequency values have small probability of being correctly imputed. In this study, average frequency of values for each class in each attribute is defined as the threshold to define the intervals, and different confidence intervals are computed for each attribute and each target class.

Boosting
In this study, strategy of boosting is used to improve the performance of Naïve-Bayes algorithm for imputation of missing values. For this purpose, a threshold is defined to select the values with high probabilities computed by the Naïve-Bayes algorithm. After applying the Naïve-Bayes algorithm to the dataset, the predicted values that have lower probability than the threshold are left missing and the process is repeated. In the next iteration the values that passed the threshold and were imputed are used to impute the remaining values improving the accuracy of the imputation. The procedure is repeated for a certain number of times, and in the last iteration the threshold is ignored and all remaining missing values are imputed.

Figure 2 shows the proposed procedure for missing data imputation.

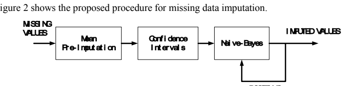

Figure 2.Procedure of missing data imputation using the proposed method

Experiments and Results

The experiments were performed using a comprehensive set of 15 datasets selected from the UCI ML repository [1]. The characteristics of these datasets are given in Table 1. The selected datasets originally did not contain missing values. The missing data were introduced artificially, using the Missing Completely at random (MCAR) model. In MCAR the distribution of an example having a missing value for an attribute does not depend on either the observed data or the missing data. The missing data was artificially generated to enable verification of the quality of imputation, which was preformed by comparing the imputed values with the original values. The missing values were introduced in 6 different quantities, i.e. 5%, 10%, 20%, 30%, 40% and 50% of data was randomly turned into missing values. This assures that entire spectrum, in terms of amount of missing values, is covered.

Table 1.Description of the datasets used in the experimentation

Name	# Examples	# Attributes	# Classes	% Boolean attributes
Soybean (small)	47	35	4	36
Postoperative Patient Data	87	9	3	11
Promoters	106	58	2	2
Monks1	432	6	2	43
Monks2	432	6	2	43
Monks3	432	6	2	43
Balance	625	5	3	0
Tic-tac-toe	958	9	2	11
CMC	1473	10	3	30
Car	1728	6	4	0
Splice	3190	62	3	0
Kr-vs-kp	3196	36	2	97
LED	6000	7	10	87
Nursery	12960	8	5	11
Kr-V-K	28056	7	17	0

The average reduction of error rates of missing data imputation for the 15 datasets using combination of mean pre-imputation and Naïve-Bayes algorithm is shown in Figure 4.

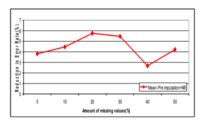

Figure 4.Average reduction in error rate of imputation using mean pre imputation and Naïve-Bayes algorithm

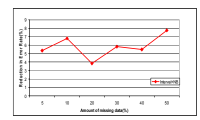

Figure 5.Average reduction in error rate of imputation using confidence intervals within the Naive-Bayes algorithm

As shown in the figure, using mean pre-imputation results in reduction of error rate up to 6% when compared with using the Naïve-Bayes algorithm without pre-imputation.
Figure 5 shows the effect of using confidence intervals with the Naïve-Bayes algorithm. Using the intervals results in improvements of up to 8% for large amount of missing values.
Average improvement in accuracy of imputation using boosting strategy with the Naïve-Bayes algorithm is shown in Figure 6, and again shows consistent improvement of up to 5% reduction in error rates.

Figure 6.Average reduction in error rate of imputation using boosting strategy on Naive-Bayes

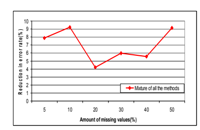

Figure 7.Average reduction in error rate using mean pre-imputation, confidence intervals and boosting the Naive-Bayes algorithm

Finally, Figure 7 shows the improvement in accuracy of imputation using the proposed methods that includes mean pre-imputation, confidence intervals and boosting. As it is evident on the graph, using the combination of the mentioned methods reduces the error rate of imputation up to 9% for large amount of missing values. However in some cases such as 20% and 30% missing values, mean pre-imputation can achieve a higher accuracy when compared to the combined method.

Conclusions

Most of the real world databases have the shortcoming of containing missing values. This paper proposes a new approach toward imputation of missing values in databases. The proposed method uses Naïve-Bayes machine learning algorithm as the basis of the imputation method and improves its accuracy by using a combination of mean pre-imputation, confidence intervals and boosting strategies. Experiments presented in this paper investigate improvement of accuracy of the proposed method versus the base Naïve-Bayes algorithm on a comprehensive range of benchmarking datasets. We show that each of the improvement strategies in separation consistently improves accuracy of imputation. The results of using combination of all strategies are also investigated. The combined strategies provide highest improvement in accuracy when compared to the base Naïve Bayes algorithms, and using each of the improvement strategies in separation. We note that execution of additional strategies does not worsen the asymptotic complexity of the imputation method, which is still linear.

References

1. C.L. Blake, and C.J. Merz, UCI Repository of Machine Learning Databases, [http://www.ics.uci.edu/~mlearn/MLRepository.html], Irvine, CA: U. of California, Department of Information and Computer Science, 1998
2. R.O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley, 1977
3. K. Lakshminarayan, S.A. Harp, and T. Samad, Imputation of Missing Data in Industrial Databases, *Applied Intelligence*, vol. 11, pp. 259 – 275, 1999
4. R.J. Little, and D.B. Rubin, *Statistical Analysis with Missing Data,* John Wiley and Sons,1987
5. T.M., Mitchell, *Machine Learning*, McGraw-Hill, 1997