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

Aliroza Farhangfar, Lukasz Kurgan and Witold Pedrycz
University of Alberta, 2nd Floor ECERF, Edmonton, Alberta, T6G 2V4

Abstract:
One of the important issues faced by researchers utilizing industrial and research databases is inconsistency of data, usually in terms of missing or erroneous values. There are many reasons for such inconsistencies, like modeled data entry procedures, incorrect measurements, equipment errors, etc. While some of the algorithms can deal directly from incomplete data, a large portion of them require complete data. Therefore, different strategies like deletion of observations with missing values, imputation using missing values, or both, are necessary to deal with such inconsistencies.

In this study, novel approaches for handling missing values are introduced. The approaches are a mixture of 15 synthetic and natural datasets, which were used to evaluate the performance of the algorithms. The proposed approach for handling missing values is to use a combined mean pre-imputation, confidence intervals and boosting. The approach is based on the assumption that having a complete training dataset would produce better results.

In this study, Naïve-Bayes algorithm is selected. Naïve-Bayes is a classification technique based on the simplest way of generating temporary values in the training dataset. It includes a mixture of 15 synthetic and natural datasets, which are used to evaluate the performance of the algorithms.

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. One group consists of univariate algorithms that do not target the missing values. These algorithms are supervised algorithms that are evaluated on different datasets, and which are used to evaluate the performance of the algorithms.

Novel Imputation

In this study, a method to filter out the least probable candidates for imputing the missing values is introduced. In order to design such a filter, the values that have high probability, and which are close to the observed data, are identified. This is done by using the combination of the mentioned methods.

Mean Pre-imputation

Based on an assumption that having a complete training dataset would produce better results, the proposed method pre-imputate missing values with temporary values. The values are used during the imputation procedure to be filtered, 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 deal with the complexity of the natural model since it is also linear.

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 have high probability, and which are close to the observed data, are identified. This is done by using the combination of the mentioned methods.

Results

In this study, a method to filter out the least probable candidates for imputing the missing values is introduced. In order to design such a filter, the values that have high probability, and which are close to the observed data, are identified. This is done by using the combination of the mentioned methods.

Experiments and Results

The experimental setup performed a comprehensive set of 15 datasets selected from the UCI Machine Learning Repository. The characteristics of these datasets are given in Table 1. The selected datasets originally contain missing values. The missing values are introduced artificially, using the Missing Completely at Random (MCAR) model. In the MCAR distribution of an example having a missing value for an attribute does not depend on either the observed data or the attribute itself. The missing values were introduced in 6 different quantities, i.e. 5%, 10%, 20%, 30%, 40% and 50% of data respectively.

None of the current methods pre-imputate missing values with temporary values. The values are used during the imputation procedure to be filtered, 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 deal with the complexity of the natural model since it is also linear. However, the proposed approach for handling missing values is not based on linear complexity, and it is used in an extensive evaluation study compared to the approach that directly applies Naïve-Bayes based approach. The experimental setup introduced a mixture of 15 synthetic and natural datasets, which were used to evaluate the performance of the algorithms.

The proposed approach for handling missing values is introduced. It includes a mixture of 15 synthetic and natural datasets, which are used to evaluate the performance of the algorithms. The proposed approach for handling missing values is to use a combined mean pre-imputation, confidence intervals and boosting. The approach is based on the assumption that having a complete training dataset would produce better results.

As shown in Figure 1, using mean pre-imputation results in a reduction of error rate up to 4% when compared to the Naïve Bayes algorithms. Figure 2 shows the effect of using confidence intervals with the Naïve-Bayes algorithms. Using the intervals results in a reduction of error rate up to 7% for large amount of missing values.

As shown in Figure 3, using mean pre-imputation results in a reduction of error rate up to 4% when compared to the Naïve Bayes algorithms. Figure 4 shows the effect of using confidence intervals with the Naïve-Bayes algorithms. Using the intervals results in a reduction of error rate up to 7% for large amount of missing values.

As shown in Figure 3, using mean pre-imputation results in a reduction of error rate up to 4% when compared to the Naïve Bayes algorithms. Figure 4 shows the effect of using confidence intervals with the Naïve-Bayes algorithms. Using the intervals results in a reduction of error rate up to 7% for large amount of missing values.

Conclusion

Most of the real world databases have the characteristics of containing missing values. The proposed method is a new approach towards imputation of missing values to databases. The proposed approach is based on the assumption that having a complete training dataset would produce better results compared to the approach that directly applies Naïve-Bayes based approach.

As shown in Figure 1, using mean pre-imputation results in a reduction of error rate up to 4% when compared to the Naïve Bayes algorithms. Figure 2 shows the effect of using confidence intervals with the Naïve-Bayes algorithms. Using the intervals results in a reduction of error rate up to 7% for large amount of missing values.

Average improvement in accuracy of imputations using boosting strategy with the Naïve-Bayes algorithms shown in Figure 4, and again shows consistent improvement of up to 7% reduction in error rate.

Finally, Figure 7 shows the improvement in accuracy of imputations using the proposed methods that include mean pre-imputation, confidence intervals and boosting. As it is evident on the graph, using the combination of the mentioned methodologies reduces the error rate of imputation up to 9% for large amount of missing values.

Mean pre-imputation was used to reducing the error rate of imputation up to 4% when compared to the Naïve Bayes algorithms. Figure 2 shows the effect of using confidence intervals with the Naïve-Bayes algorithms. Using the intervals results in a reduction of error rate up to 7% for large amount of missing values.