Imputation of Missing Values by Neural Networks for Data Mining Applications

Alfred Ultsch, Susanne Rolf
Department of Computer Science
Philipps-University Marburg
Hans-Meerwein-Strasse
35039 Marburg
ultsch@informatik.uni-marburg.de, rolf@...

June 14, 2000

Abstract

Especially in the case of data mining, the problem of missing values becomes very difficult because usually there are no preliminar hypotheses about the components' distributions or their relationships, which makes the use of traditional techniques that rely on model-building impossible. Therefore in the framework of data mining a method for imputation of missing values has to be identified that does not rely on certain distributions or elaborate modelling. On the basis of an emergent self-organizing map (ESOM) ([4]), two methods for the imputation of missing values are proposed. Both methods use the ESOM's ability to adjust to the inherent structure of highdimensional data.

1 Introduction to the Problem

Though it is common practice to simply ignore those observations that contain missing values in one or more components, the loss of information in doing so is regrettable. This is especially the case for high-dimensional data, where the information in all other components is left out because of one missing value in a single component. Moreover, if the share of missing values is too high, it is very likely that structural features can no longer be captured by the analyzing algorithms.

Two main options for the handling of missing data can be distinguished: On one hand methods are developed that are able to impute missing values such that all following methods are provided with a complete data matrix. The second option is to adapt the used Data Mining techniques such that they can handle incomplete data. The main objective of this paper is to study the capacity

of emergent self-organizing maps ([4]) for the imputation of missing values in comparison to commonly used imputation techniques.

2 Neural Net Methods for Missing Value Imputation

The proposed Neural Net methods are based on emergent self-organizing maps (ESOMs). ESOMs are Neural Nets with a large number of neurons arranged on a grid. In the course of a training phase the map adjusts to the high-dimensional structure of the data and represents it on the two-dimensional grid. Starting from this concept, two methods for missing value imputation are proposed:

The **IESOM** (Imputing emergent self-organizing map) method trains an ESOM with all complete observations. Once the IESOM is trained, the most similar neuron (the "best match") is sought. All missing components are then replaced by the corresponding values of the neuron. IESOM uses the fact that the neurons of an ESOM interpolate between the values presented during the training phase. Therefore not only the values in the data set are substituted for missing values, but also meaningful values between them.

The UDnet ([5]) is a Neural Net which is able to handle data with skewed distributions and different scales. Since ESOMs are sensitive to both problems, the combination of the two methods offers an approach to structure analysis that needs no preprocessing. For the **UDnet-Imputation**, the data is presented to the ESOMs through the UDnet. In the course of the training, the UDnet learns the different distributions and scales in the dataset and is therefore able to interpret the raw data for the ESOM. The ESOM then adapts to the data-inherent structure as usual. For a more precise definition of the UDnet-method see [5].

3 Comparison of Methods

To assess the quality of the different methods, a dataset was derived form Morningstar™, a Chicago-based company that provides data on the majority of stocks and funds listed on US exchange ([2]). Out of the 7700 available observations, 330 complete observations have been chosen. The 19 variables in the data set have been transformed into a standard-normal distribution.

For the experiment, missing values have been inserted into the dataset. The values to be deleted have been chosen randomly over all observations and variables. Four different numbers of missing values (100, 300, 500 and 700) have been inserted. No restrictions on the number of missing values per observation have been introduced. 20 data sets have been generated for each number of missing values, such that there are 80 data sets in the experiment. For 100 missing values an average of 26.21% of the observations are incomplete. This value increases to 60.74% for 300 missing values, 79.36% for 500 missing values and reaches 89.33% for 700 missing values.

4 Results

The main aim of the experiment was to assess the quality of the imputation techniques in terms of the error made in substituting missing values. To measure the error, the mean absolute error (*mae*) ([3]) will be used. Though the mean absolute error will be considered as the main measurement for quality, some attention will also be given to the distortion of the underlying standard-normal distributions.

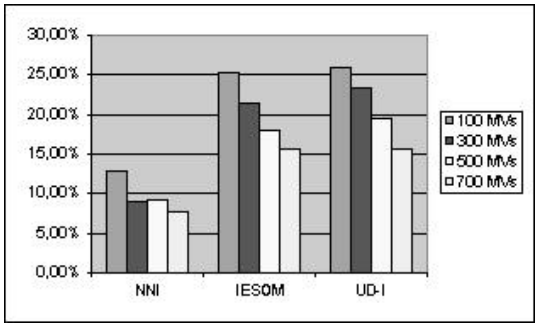

Figure 1: Mean Absolute Error Improvement in Comparison to Mean Imputation

The Mean Imputation (MI) shows the worst result for the *mae* with errors around 0.75 for all numbers of missing values. This is no surprise, because it is the only method which does not use relationships between the variables. Nevertheless the MI is a widely used technique for missing value handling and is therefore used as a reference for this study. Figure 1 shows the improvement on Mean Imputation's *mae*s reached by the other three methods under consideration. The Nearest Neighbor Imputation (NNI) improves the results of MI by about 8% to 13%. The IESOM's *mae*s are between 16% and 25% better than the references. Approximately the same holds for the UDnet-Imputation (UD-I), where improvements between 16% and 26% are attained.

Considering the error in estimating the distribution's mean, the methods show nearly the same performance in this respect. All methods generate errors at about 0.005 for 100 missing values in the dataset. The error increases with the number of missing values to up to about 0.016 for 700 missing values.

The MI generates average variances between 0.98 and 0.88. The NNI overestimates in the range from 1.01 to 1.026 and underestimates between 0.99 and 0.96. The IESOM computes average variances between 1.002 and 1.03 when overestimating and between 0.99 and 0.94 when underestimating. The UD-I underestimates between 0.996 and 0.98 and overestimates between 1.004 and 1.02. All methods show a greater bias when the number of missing values increases.

5 Discussion and Conclusion

Within the experiment the Mean Imputation method turned out to generate the worst results. This was expected because it is the only method under consideration which does not exploit the data-inherent structure. Nevertheless Mean Imputation is still widely used and therefore used as a reference in this study.

The Nearest Neighbor Imputation could already improve the results attained by Mean Imputation by up to 13%. The presented methods based on Neural Nets, IESOM and UD-Imputation, gave better results than Mean Imputation or Nearest Neighbor Imputation. Both methods lowered the values of the Mean Imputation up to 26%. Since both methods rely only on information available in complete observations, their performance decreases with the number of missing values in the dataset. Nevertheless, even for the worst case of 700 missing values the results of the Neural Net methods are still better than the Nearest Neighbor Imputation.

Since the Neural Net methods need no prior information about the data, no interaction with the analyzer and no explicit model building, they can be considered to be well suitable for Data Mining applications.

Unlike Nearest Neighbor Imputation, the Neural Net methods do not use the whole information available in the data. Therefore a Neural Net capable of handling missing values would be expected to perform even better. This is subject to further investigation.

References

[1] Bankhofer, U. (1995): *Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse* Verlag Josef Eul, Bergisch Gladbach, Köln.

[2] Deboeck, G. and Ultsch, A. (1999): Picking Stocks with Emergent Self-organizing Value Maps. In: Novac, M.: *Neural Network World*, Vol. 10, Nr. 1–2, Prague, 1999, pp. 203 – 216.

[3] Schwab, G. (1991): *Fehlende Werte in der angewandten Statistik*, Dt. Univ.-Verl., Wiesbaden.

[4] Ultsch, A. (1999): Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series. In: Oja, E., Kaski, S. (eds.): *Kohonen Maps*, Elsevier, 33–46.

[5] Ultsch, A. (2000): A Neural Network to Compare Highdimensional Data with Skewed and Unknown Distributions. In: Proceedings 24. Jahrestagung der Gesellschaft für Klassifikation.