The Completion of Missing Values by Neural Nets for Data Mining

A. Ultsch, S. Rolf

Department of Mathematics and Computer Science,
Philipps-University Marburg, D-35032 Marburg, Germany

Abstract: Missing values are a problem occuring in the analysis of most application datasets. Though the deletion of observations with missing values is common practice, the loss of information in doing so is regrettable. This is even more the case for high-dimensional data, where the structural information in many components of an observation is lost because of few missing values. A new method is suggested for missing value imputation. The basis of the presented approach is an emergent self-organizing neural network that adjusts to the inherent structure of the input data. The technique has been successfully tested on a real-world example of financial data and turned out to be superior to the other methods under consideration.

1 Introduction to the Problem

Data Mining and Knowledge Discovery deal with the analysis of large multivariate datasets with respect to their inherent structure. For this task usually classes of elementary states are sought in the data. The Data Mining tool "Neuro Data Mine" (Ultsch (1999)) has proven to be a valueable tool for Knowledge Discovery in the last years.

A major problem to most Data Mining methods is the problem of missing values within many real-world datasets. 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.

Since missing values are a problem which occurs in many real-world data sets, the question of its handling is of high relevance to the Data Mining community (e.g. Lakshmirnarayan et al. (1996); Timm, Klawonn (1998); Ragel (1998)). 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 (Ultsch (1999)) for the imputation of missing values in comparison to commonly used imputation techniques.

2 Methods for Missing Value Imputation

2.1 Standard-Methods

Bankhofer (1995) distinguishes three major types of imputation techniques for missing data for cases in which the pattern of missingness is not systematic and the scale of the data is metric:

Simple techniques:
Imputation of each missing value by a chosen measure of central tendency as e.g. mean or median. If the underlying distribution of the data is known, the missing values can also be imputed by generating random numbers from that distribution.

Imputation within classes:
If classes of similar objects are known, this information can be used for the imputation of missing values. Within its imputation-class a reference observation is chosen for each incomplete observation The missing values are then replaced by the corresponding values of the reference. The simplest case of imputation within classes is the Nearest Neighbor Imputation (NNI), where for each incomplete observation the most similar observation with respect to all available components is chosen to be the reference vector.

Multivariate techniques:
Multivariate techniques build a multivariate model of the data and estimate the model parameters. On the basis of this model the missing values are estimated.

2.2 Neural Net Methods

The proposed Neural Net methods are based on emergent self-organizing maps (ESOMs) which have proven to be a powerful tool for structure analysis in high-dimensional data sets. ESOMs are Neural Nets with a large number of neurons arranged on a grid. The number of neurons usually exceeds the number of observations in the dataset. 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:

IESOM – Imputing emergent self-organizing map
The IESOM 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.

UDnet-Imputation
The UDnet (Ultsch (2000)) is a Neural Net which is able to handle data with skewed distributions and different scales. It adapts itself to the distribution of each variable. After its learning phase a ud-network produces an output that can be interpreted as a 'relative distance' which is close to 1 if two values are rather different and close to 0 if they are very similar. 'Similarity' in a ud-net depends on the underlying distribution of the variables under consideration. It could be demonstrated in (Ultsch(2000)) that this learned distance is reasonably robust against standard types of non-linear transformations.

Since ESOMs are sensitive to skewed distributions as well as to variables on different scales, 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 Ultsch (2000).

3 Comparison of Methods

To assess the quality of the different methods, a dataset was derived form Morningstar^{TM}, a Chicago-based company that provides data on the majority of stocks and funds listed on US exchange (Deboeck, Ultsch (1999)). Morningstar^{TM} publishes monthly and quarterly fundamental and technical information on over 7700 stocks listed on the NYSE, AMEX and NASDAQ exchanges. From these 7700 observations, 330 complete observations have been chosen. The 19 variables in the data set have been transformed into a standard-normal distribution using log and sqrt- transformations.

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, therefore the missing value mechanism is MCAR (missing completely at random) (Little, Rubin (1987)).

Since one subject of the experiment is the performance in dependence on the number of incomplete cases, four different numbers of missing val-

ues (100, 300, 500 and 700) have been inserted. No restrictions on the number of missing values per observation have been introduced. To get more reliable results, 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) (Schwab (1991)) will be used:

$$mae = \frac{1}{20} \sum_{d=1}^{20} \frac{1}{k_d} \sum_{d=1}^{20} \sum_{i=1}^{n} \sum_{j=1}^{m} |x_{ij} - y_{dij}| \tag{1}$$

with k_d no. of missing values in dataset d, n no. of observations, m no. of components, x_{ij} jth component of ith observation in the original dataset, y_{dij} corresponding value in the dth imputed dataset (either the value itself or the substitute for a missing value).

Though the mean absolute error will be considered as the main measurement for quality, some attention should also be given to the distortion of the underlying standard-normal distributions. Therefore the following two additional measures are introduced:

The mean absolute error in estimating the distribution's mean measures the bias in the estimation of the mean introduced by the substitution of the missing values. Since all variables in the 20 datasets follow a standard-normal distribution, the deviation from the true mean is averaged over all datasets and variables.

To measure the deviation from the true variance, the mean error for the estimation of the variance is reported. This error is splitted into the cases of over- and underestimation since – in contrary to the mean value – the direction of the bias is of some interest for the variance.

As could be expected, 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 therefore shows the improvement on Mean Imputation's maes reached by the other three methods under consideration.

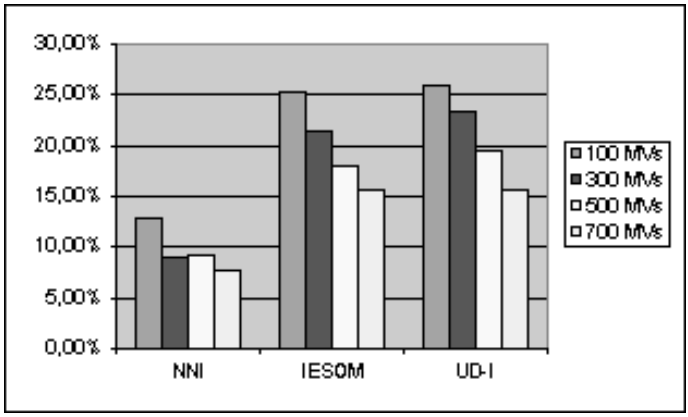

Figure 1: Mean Absolute Error Improvement in Comparison to Mean Imputation

It can be seen that the Nearest Neighbor Imputation (NNI) improves the results from 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.

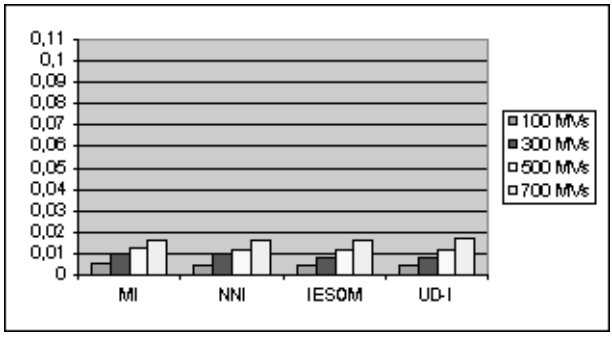

Figure 2: Mean Absolute Error in Estimation of Distribution's Mean

Considering the error in estimating the distribution's mean, it can be seen from Figure 2 that 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.

Figure 3 shows the outcomes for the methods concerning variance estimation splitted into under- and overestimation. The MI generates average variances between 0.98 and 0.88. The NNI overestimates the variance 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

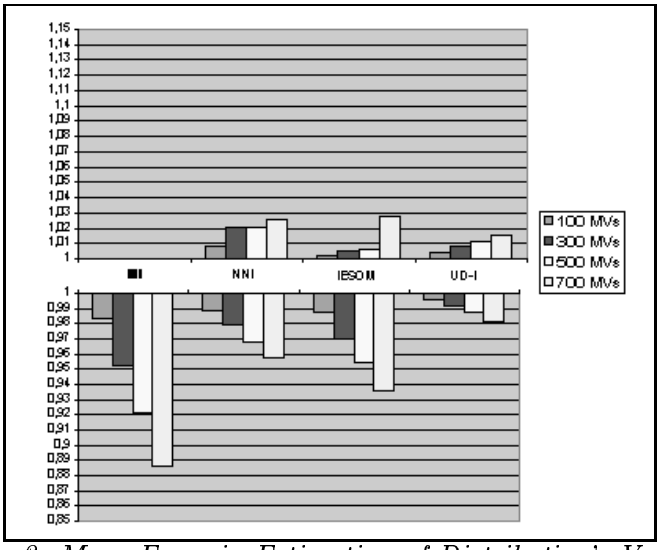

Figure 3: Mean Error in Estimation of Distribution's Variance

and 1.03 when overestimating and between 0.99 and 0.94 when underestimation. The UD-I underestimates variances on average 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

Within the experiment the Mean Imputation method turned out to generate the worst results in terms of Mean Absolute Error in substituting the missing values. This result 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%. It also turned out to be the least sensitive to the number of missing values. This observation can be explained by the fact that the Nearest Neighbor Imputation uses all available information in the data by allowing for completed observations to be used as a reference in the further imputation process.

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.

Though the Mean Absolute Error is considered to be the main quality-criterion, some attention should also be given to the distortion of the underlying distributions caused by the imputations. Inspection of the estimated means and variances leads to the following results:

All methods generated nearly the same distortion of the mean value by means of the imputation. However, the measured values between 0.004 and 0.017 do not give rise to concern. As a comparison: A Gauss-test (Hartung(1991), p. 178) with hypothesis "Mean equals zero" would only reject the hypothesis for values larger than 0.1081 at an α-level of 0.05.

In contrary to the mean value, the methods performed quite differently for the variance: The Mean Imputation showed worst results for all numbers of missing values. The Nearest Neighbor Imputation generated clearly better estimates, showing just slight differences in over- and underestimation. The IESOM showed larger distortions in underestimating the variance, still being clearly better than the Mean Imputation. Best performance was shown by the UD-Imputation, average distortions of no more than 0.02 were computed. A two-sided test on "Variance equals 1" (Hartung(1991), p. 179) would not reject the hypothesis in an intervall of [0.8530, 1.1585] at 0.05 α-level. Only the Mean Imputation for 700 missing values comes close to that border with an average estimated variance of 0.89.

6 Conclusion

In this paper two methods for the imputation of missing values based on Emergent Self-Organizing Maps (ESOM) are proposed: IESOM and UD-Imputation. An experiment has been performed on real data which showed that both methods are superior to commonly used methods as Mean Imputation and 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

BANKHOFER, U. (1995): Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse, Verlag Josef Eul, Bergisch Gladbach, Köln.

DEBOECK, G. and ULTSCH, A. (1999): Picking Stocks with Emergent Self-organizing Value Maps, to appear in Proc. PASE 2000.

HARTUNG, J. (1991): Statistik: Lehr- und Handbuch der angewandten Statistik. 8. Auflage, München, Wien, Oldenbourg, 1991.

LITTLE, R. J. A. and RUBIN, D. B. (1987): Statistical Analysis with Missing Data, Wiley, New York.

LAKSHMINARAYAN, K. and HARP, S.A. and GOLDMAN, R. and SAMAD, T. (1996): Imputation of missing data using machine learning techniques, in: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 140–145.

RAGEL, A. (1998): Preprocessing of Missing Values Using Robust Association Rules, in: Zytkow, J.M. (Ed.): Principles of data mining and knowledge discovery, 2nd European Symposium, Lecture Notes in Artificial Intelligence 1510, pp. 414–422.

SCHWAB, G. (1991): Fehlende Werte in der angewandten Statistik, Dt. Univ.-Verl., Wiesbaden.

TIMM, H. and KLAWONN, F. (1998): Classification of Data with Missing Values, in: Proceedings of the 6th European Conference in Intelligent Techniques and Soft Computing, pp. 639–643.

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, pp. 33–46.

ULTSCH, A. (2000): A Neural Network Learning Relative Distances, in: Amari, S., Giles, C., Gori, M., Piuri, V. (Eds.): Neural Computing: New Challenges and Perspectives for the New Millennium, 2000 IEEE International Joint Conference on Neural Networks (IJCNN 2000), Vol. V, pp. 553–558, Como, 2000.