Processing Missing Values with Self-Organized Maps

David Sommer, Tobias Grimm, Martin Golz
University of Applied Sciences Schmalkalden
Department of Computer Science
D-98574 Schmalkalden, Germany
Phone: +49-3683-688-4107, Fax: +49-3683-688-4499
email:{d.sommer, m.golz}@fh-sm.de

ABSTRACT: This paper introduces modifications of Self-Organizing Maps allowing imputation and classification of data containing missing values. The robustness of the proposed modifications is shown using experimental results of a standard data set. A comparison to modified Fuzzy cluster methods [Timm et al., 2002] is presented. Both methods performed better with available case analysis compared to complete case analysis. Further modifications of the SOM using k-nearest neighbor calculations result in lower classification errors and lower variances of classification errors.

KEYWORDS: Missing values, self-organizing map, fuzzy c-means, classification, imputation

INTRODUCTION

Missing values are a common problem in many data mining applications. They represent loss of information and are in general not restorable. Depending on their origin a restoration becomes possible if further knowledge is provided or if their appearance is the result of a random process. Some exemplary causes of their appearance are

- errors in sampling, in transmission, in storing or in pre-processing,
- high costs to get the missing information,
- lapses in experimental design
- if dealing with questionnaires insufficient knowledge, or low motivation, inattentiveness, refusal to answer may have played a role.

We define missing values for cases where an attribute value is missing but it is possible to get this attribute in principle. On the other hand empty values are present if it is not possible to get this attribute, because the real object has no counterpart to it [Pyle, 1999]. Therefore an imputation of empty values makes no sense, but is desirable for missing values. Classification tasks of data containing missing and empty values are reasonable in both cases.

The mechanisms of generating missing values are important for further analysis. For a given data set with 'm' samples and 'n' attributes let $x_i = (x_{i1}, x_{i2},\ldots, x_{in})$, $i = 1, \ldots, m$, denote the attribute values, which are values of a continuous random variable $X = (X_1, X_2,\ldots, X_n)$. Let $f_X(x_i; \theta)$ be their probability density function, which equals the joint distribution of the observed part X_{obs} of X and the missing part X_{mis} of X, and is expressed by $f_X(x_i; \theta) = f_{X_{obs}, X_{mis}}(x_{obs}, x_{mis}; \theta)$.

Let $M = (M_1, M_2,\ldots, M_n)$ be a binary indicator variable for all variables X_j ($j = 1,\ldots, n$), indicating observed values with $M_{ij} = 1$ and missing values with $M_{ij} = 0$.

The conditional probability density function of M given X depends generally on a set of parameters θ and on an unknown noise parameter ξ : $f_{M \mid X}(m \mid x; \xi, \theta)$. We call this general case MNAR (missing not at random) [Little & Rubin, 1987], because M and X_{mis} may be dependent variables and ξ may depend on θ. Hence the mechanisms of generating missing values are not ignorable, and therefore some authors call this case NI (non-ignorable).

If on the other hand θ will not provide any information on ξ and vice-versa, and if M and X_{mis} are independent variables and only M and X_{obs} may be dependent, the MAR condition (missing at random) is fulfilled with $f_{M \mid X}(m \mid x; \xi, \theta) = f_{M \mid X_{obs}}(m \mid x_{obs}; \xi)$.

If furthermore M and X_{obs} are independent variables, the MCAR case (missing completely at random) is fulfilled, i.e. $f_{M \mid X}(m \mid x; \xi, \theta) = f_M(m; \xi)$. The mechanisms of generating missing values are ignorable, and therefore MAR and MCAR are also called ignorable.

In many applications there is some suggestion of MNAR processes, e.g. interviewees with higher income show a higher probability of refusal to answer. But mostly the ignorable case is assumed because of a lack of deeper insight into the process statistics.

Many methods of cluster analysis and many classification methods only allow the complete case analysis, where every sample containing missing values has to be eliminated beforehand. Alternatively, if as a consequence of a MAR process all missing values are distributed in a few attributes, the elimination of attributes containing missing values may be a better choice. Complete case analysis is a rationale if the number of missing values is very small. For example, a data set with n = 15 attributes is given and the probability of non-missing values is p = 0.9 . Assume that the missing values were generated by a MCAR process where $Y = \sum_{i=1}^{n} M_i$ is a binomial distributed variable, because the M_i are Bernoulli processes. Then the probability of complete samples, i.e. the number of available attributes is k = 15, is given by $P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{15}{15} 0.9^{15} (1-0.9)^0 = 0.9^{15} = 0.206$. In this case four of five samples have to be eliminated beforehand. As a consequence a large amount of valuable information is discarded. Therefore the application of methods using all available cases becomes necessary.
This paper introduces modifications of Self-Organizing Maps [Kohonen, 1982] allowing imputation and classification of data containing missing values. The robustness of the proposed modifications is shown using experimental results of a standard data set. A comparison to modified Fuzzy Cluster methods [Timm et al., 2002] is presented.

MODIFICATIONS TO SELF-ORGANIZING MAPS

Artificial Neural Networks are known as robust methods. They are capable to deal with uncertain inputs and also with missing inputs [Ishibuchi et al., 1994] [Tresp & Ahmad, 1995] [Gupta & Lam, 1996]. They used Error Backpropagation Networks, whereas [Kaski & Kohonen, 1996] applied the Self-Organizing Map (SOM) to the incomplete world development data set [World Bank, 1992]. This data set consists of 126 samples (countries) and 39 attributes. As a result these authors generated a so called “poverty map”, but a quantitative consideration on the influence of missing values is left.

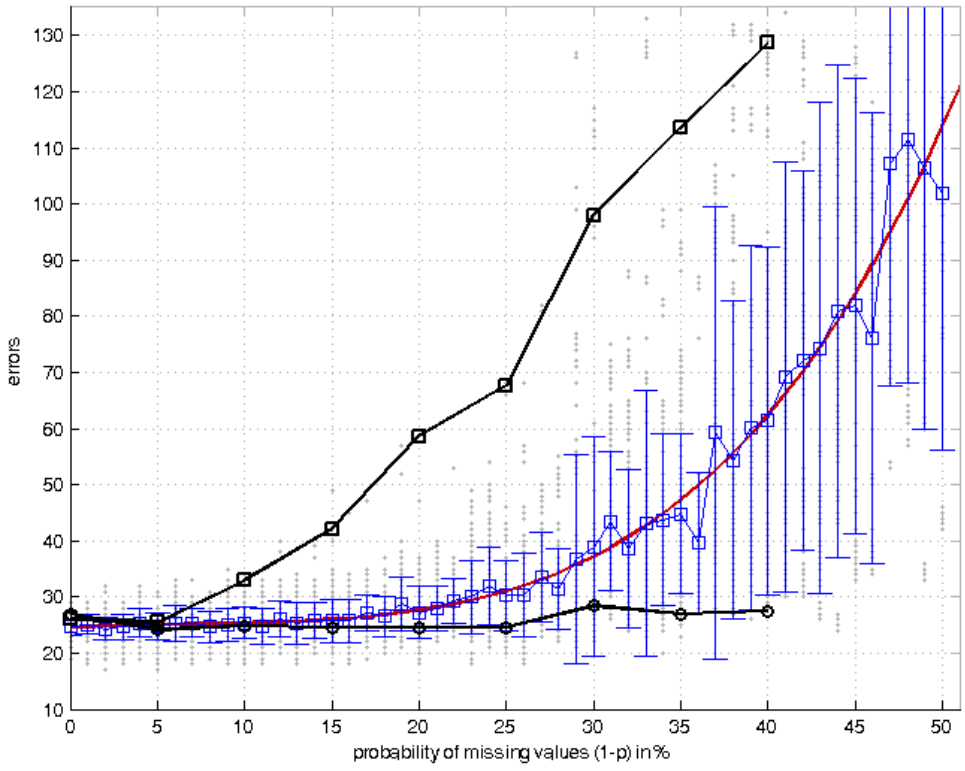

Figure 1: Complete case analysis. Number of misclassified samples versus probability of missing values. Wisconsin breast cancer data set; MCAR missing values. Light gray dots: SOM-algorithm, complete case; squares with error bars: SOM, mean ± standard deviation; thick line: SOM, trend polynom; thick line with squares: fuzzy c-means, complete case; thick line with circles: fuzzy c-means, available case.

SOM is an unsupervised method based on estimation of distances in $\Re^n$. It is straightforward to show that axioms of a metric space, like the positive definiteness and the Schwarz inequality, cannot hold for elements containing missing values. Thus working with a distance-based method is only possible under the assumption that in practical cases the degradation due to missing values is low.
The distance $\|x - w_j\|$ between the input vector x of the actual iteration and the prototype vectors w_j has to be modified by $\|\mathrm{diag}(m)\ (x - w_j)\|$, where diag(m) is a matrix in diagonal form containing the elements of the actual indicator values m in the main diagonal. The learning rule has also to be modified by introducing the indicator values: $\Delta w_j(t) = \eta(t)\ h_{cj}(t)\ m\ (x - w_j)$. Both modifications result in calculations of available cases. This concept is related to the pairwise deletion method known from literature [Little & Rubin, 1987].
Because of metric violations caused by $\|\mathrm{diag}(m)\ (x - w_j)\|$ the determination of the closest prototype vector $\mathbf{w}_c$ as winner of the competition during iterative training of the SOM is erroneous with an unknown probability. In such a case the false winner and all prototype vectors in his topological neighborhood are adapted. Among other influences this can lead to errors in topological ordering. With a high probability the correct prototype vector should be under the first k-nearest prototypes. Therefore we investigated experimentally if the classification errors could be decreased by adapting the first k nearest prototypes. At the end of SOM training we perform an imputation of missing values by the mean of the related attribute values of the k nearest prototypes. This improves the following calibration step and results in lower numbers of errors.

EXPERIMENTS

The modified SOM algorithm was tested with the breast cancer dataset [Mangasarian & Wolberg, 1990], which has 458 samples in the class 'benign' and 241 samples in the class 'malign'. In one of the 9 attributes 16 values are missing.
At first a complete case analysis was performed. The following steps were repeated 200 times for every parameter setting: (1) generate additional missing values completely at random by deleting attribute values; (2) execute SOM training; (3) execute SOM calibration; (4) count classification errors during the recall phase of the SOM.

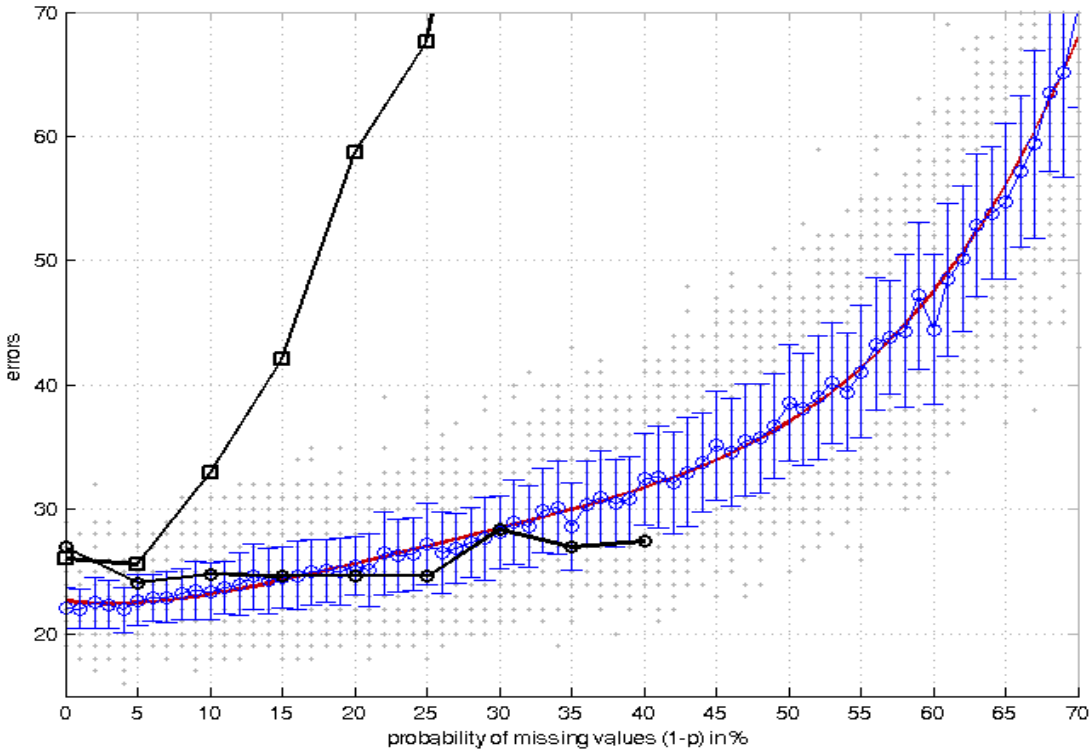

Figure 2: Available case analysis. Number of misclassified samples versus probability of missing values. Wisconsin breast cancer data set; MCAR missing values. Light gray dots: SOM-algorithm, available case Circles with error bars: SOM, mean ± standard deviation; Thick line: SOM, trend polynom; Thick line with squares: fuzzy c-means, complete case; Thick line with circles: fuzzy c-means, available case

Figure 1 shows the classification errors versus (1 - p), the probability of missing values (light gray points). The mean values over all trials with a fixed probability of missing values and their standard deviations (squares with error bars) as well as a trend polynom were added. For comparison the results for complete case analysis (squares without error bars) and the results for available case analysis (circles without error bars) both with the fuzzy c-means algorithm are shown [Timm et al, 2002]. Furthermore, for comparisons with the fuzzy algorithms SOMs with few prototypes were used. Expected better results with increased number of prototypes are shown with an example in Figure 3 (dash-dot line).

The SOM consists of 4 x 4 neurons and shows a surprisingly low mean error rate until $(1 - p) \leq 0.25$ and is in the same range as the fuzzy c-means algorithm for the available case. Disadvantageously the standard deviation is constantly growing with (1 - p). The large standard deviations are mainly caused by the random MCAR deletion process and the subsequent deletion of incomplete cases. Repeated runs of SOM training and calibration result in much lower standard deviations.

For the available case analysis using SOM the same steps as for the complete case analysis were performed. The only differences were the modification of distance calculation and of the learning rule as explained above. The mean values and the standard deviations (circles with error bars) as well as the trend polynom are shown (Figure 2). The two results of fuzzy c-means algorithms [Timm et al, 2002] are added without changes.

The number of errors is lowered by performing available case analysis with SOM. Again 4x4 neurons were used. For $(1 - p) \leq 0.15$ SOM performs better than fuzzy c-means. With increasing $(1 – p)$ the number of errors is increasing, but in contrast to the complete case analysis the standard deviations are distinctly lower. Notwithstanding the modified fuzzy c-means algorithm seems to be slightly better than SOM for $(1 - p) > 0.15$.

A further slight improvement can be obtained using the k-nearest neighbor modification of the SOM (Figure 3). With increasing (1 - p) the number of errors is increasing slowlier and the performance is comparable to the fuzzy c-means algorithm in the range $0.15 \leq (1 – p) \leq 0.40$. The standard deviations of modified SOM show no clear difference with respect to the unmodified SOM.

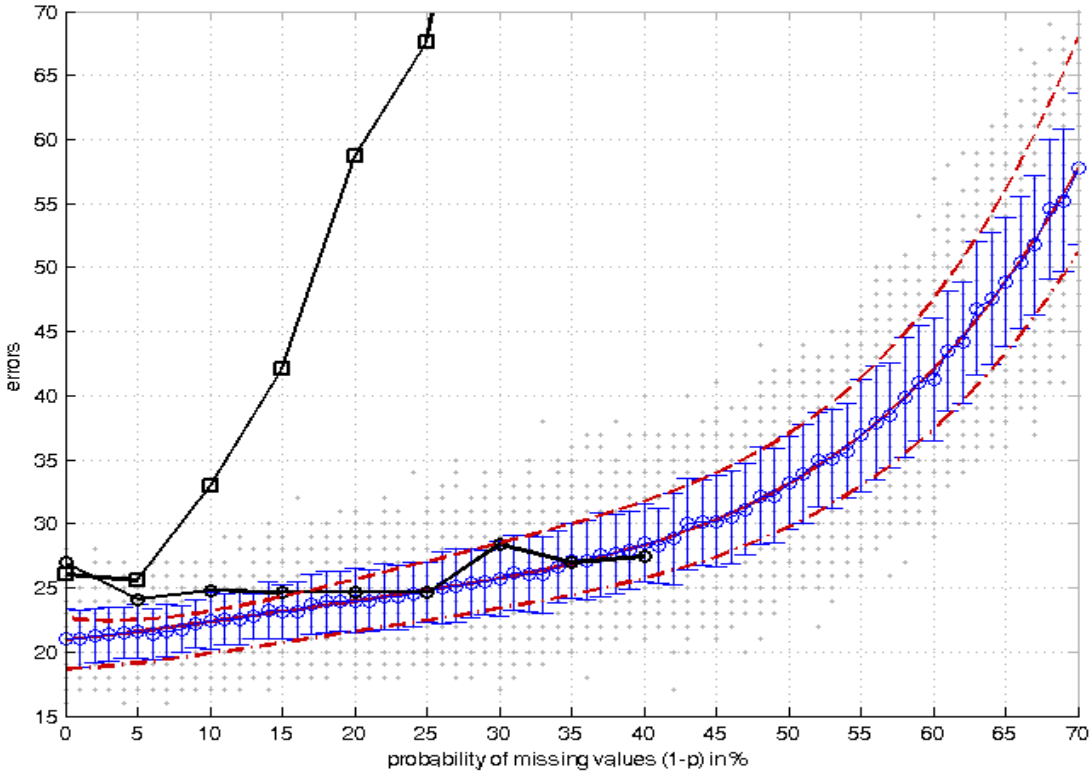

Figure 3: Modified available case analysis. Number of misclassified samples versus probability of missing values. Wisconsin breast cancer data set; MCAR missing values. Light gray dots: modified SOM-algorithm, available case Circles with error bars: SOM, mean ± standard deviation; Thick line: modified SOM, trend polynom; Dashed line: unmodified SOM, trend polynom (Figure 2); Dash-dot line: modified SOM with 8x8 neurons, trend polynom; Thick line with squares: fuzzy c-means, complete case; Thick line with circles: fuzzy c-means, available case

CONCLUSIONS

In this paper we investigated the ability of Self-Organizing Maps to deal with missing values. Caused by missing values violations of metric spaces can occur with an unknown probability. Though distance based methods, like the Self-Organizing Map and the fuzzy c-means algorithm, should therefore be problematic. As it was shown experimentally for one data set, both methods can be utilized successfully to the available case analysis and result in lower classification errors than in complete case analysis and result in lower variances of misclassification. Probabilities of missing values can reach 0.4 without noticeable performance degradation. The performances of both methods differ slightly. SOM performed better at low probabilities of missing values, whereas the fuzzy c-means algorithm tends to be better at higher probabilities of missing values.

REFERENCES

Gupta, A. and Lam, M.S.; Estimating missing values using neural networks. J Operational Research Society, 47, 229-238.; 1996

Ishibuchi, H., Miyazaki, A. and Tanaka, H.; Neural-Network-based Diagnosis Systems for Incomplete Data with Missing Inputs. IEEE World Congress on Computational Intelligence (pp. 3457-3460). Orlando, Florida.; 1994

Kaski, S., Honkela, T., Lagus, C. and Kohonen, T.; Creating an Order in Digital Libraries with Self-Organizing Maps. WCNN96: World Congress an Neural Networks, 814-817.; 1996

Kohonen, T.; Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics, 43, 59-69.; 1982

Little, R. and Rubin, D.; Statistical Analysis with Missing Data. New York: John Wiley & Sons.; 1987

Mangasarian, O. and Wolberg, W.; Cancer diagnosis via linear programming, SIAM News, Volume 23, Number 5, pp 1 & 18.; 1990

Pyle, D.; Data Preparation for Data Mining. San Francisco: Morgan Kaufmann Publ. Inc.; 1999

Timm, H., Döring, C. and Kruse, R.; Fuzzy Cluster Analysis of Partially Missing Data. Proc. Europ. Symp. Intell. Technol. (EUNITE 2002) (pp. 426-431). Albufeira, Portugal.; 2002

Tresp, V., Neuneier, R. and Ahmad, S.; Efficient Methods for Dealing with Missing Data in Supervised Learning. In G. Tesauro, D. S. Touretzky and T. K. Leen (Ed.), (pp. 689- 696). Cambridge, MA:MIT Press.; 1995

World Bank.; World Development Report 1992. New York, Oxford University Press.; 1992