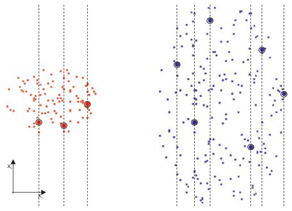
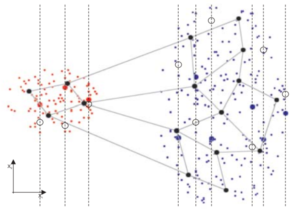
Processing Missing Values with Self-Organizing Maps

David Sommer, Tobias Grimm, Martin Golz
University of Applied Sciences Schmalkalden, Department of Computer Science, PF 182
D-98574 Schmalkalden, Germany
{d.sommer, m.golz}@fh-sm.de, http://www.sund.de, http://www.sund.de/golz

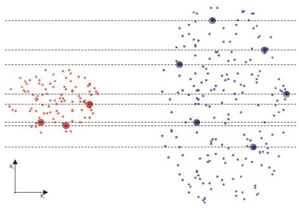
A common problem in data mining applications is the occurrence of missing values. Many pattern recognition algorithms cannot handle such a problem. As a consequence one has to eliminate all feature vectors containing missing values (Complete case analysis). Disadvantageously the information of the eliminated vectors can not be used and the performance is strongly decreased in case of high missingness.



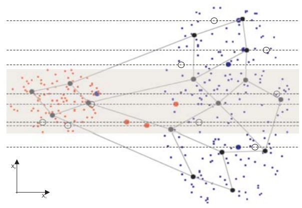
Example 1: two-dimensional artificial data; Attribute x_1 is discriminating, while x_2 is not. 9 missing values (black circles) were generated randomly only in attribute x_2.



Example 1: Calibrated Self-Organizing Map (4x4 neurons) trained on the complete case solves classification problem sufficiently.

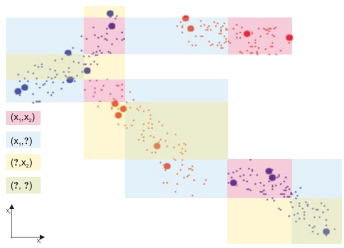


Example 2: the same artificial data set; 9 missing values (circles) were generated randomly in the discriminating attribute x_1.

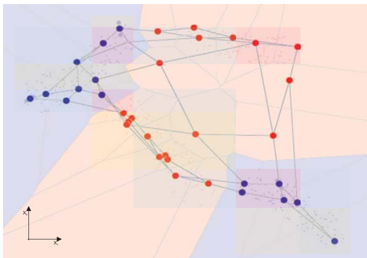


Example 2: The calibrated Self-Organizing Map (4 x 4 neurons) trained on the complete case solves the classification problem insufficiently. In the overlapping region (gray area) the calculation of the winner neuron is fully random.

Complete case analysis



Example 3: a two-dimensional artificial data set with 95% missing values; data vectors containing one missing value are marked by small dots; complete vectors are marked by solid dots. In some regions only attribute x_1 is discriminating (blue areas), while in other regions only x_2 is discriminating (yellow areas) and in other regions both attributes are discriminating (red areas); x_1 or x_2 is discriminating (green areas)

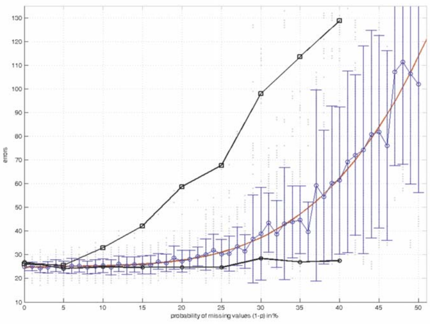


The calibrated Self-Organizing Map (6 x 6 neurons) leads to mean classification rates of 74%.

SOM performs better than fuzzy c-means but cannot achieve the results of available case analysis. The error variance is increasing with increasing missingness.

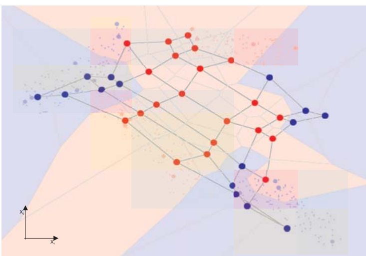
[1] Mangasarian, Olvi L. and Wolberg, William H.: Cancer diagnosis via linear programming, SIAM News, Volume 23,Number 5, pp 1 & 18.; 1990

[2] 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

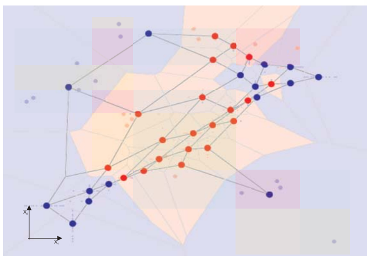


Example 4: Wisconsin breast cancer data set [1]; MCAR missing values number of misclassified samples versus probability of missing values
light gray dots: SOM-algorithm, complete case
squares with error bars: SOM, mean ± standard deviation
thick red line: SOM, trend polynom
thick black line with squares: fuzzy c-means, complete case [2]
thick black line with circles: fuzzy c-means, available case [2]

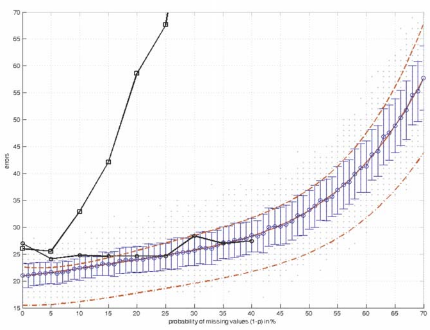
Available case analysis



Example 3: The calibrated Self-Organizing Map (6 x 6 neurons) trained on available cases leads to mean classification rates of 77%. Neuron-neuron distances near zero as in the complete case SOM are avoided.



Example3: Application of our modified SOM algorithm using an imputation rule during training leads to mean classification rates of 81%.



Example 4: Wisconsin breast cancer data set [1]; MCAR missing values number of misclassified samples versus probability of missing values
light gray dots: modified SOM-algorithm, available case
circles with error bars: modified SOM, mean ± standard deviation
thick red line: modified SOM, (dashed : unmodified SOM), trend polynom
dash-dot line: modified SOM with 16x16 neurons, trend polynom;
thick black line with squares: fuzzy c-means, complete case [2]
thick black line with circles: fuzzy c-means, available case [2]

The SOM (4 x 4 neurons) with standard available case method has the same performance like fuzzy c-means [2]. Our modified algorithm is slightly better than the fuzzy c-means. With more neurons (16 x 16 map) the performance is expectingly higher.

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 [2] 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.