Thesis Extract

András Simon

The topic of the thesis is incomplete data handling for machine learning. In the thesis the most popular stat-of-the-art incomplete data handling techniques are introduced and a novel method is presented. The proposed, novel algorithm is called HIDD (Handling Incomplete Data Directly). The HIDD algorithm is based on the idea of Viharos et al., which is a neural network construction for handling incomplete data without imputation. The proposed method is the improvement of the preliminary method by including it into the Levenberg-Marquardt learning algorithm. Beyond this novel learning algorithm some other changes have been made in order to improve the efficiency.

Comprehensive test have been performed in order to prove the convergence of the proposed algorithm and to compare it of other state-of-the art imputation methods. The HIDD method's performance have been tested beside all the three type on incompleteness: MCAR, MAR and MNAR. For first the patterns of missingness are self created on the available, well-known, widely used benchmarking, real-life data set of Iris with deleting certain parts of the data from the original, complete data set. In the second test a supervised learning problem is solved on the Horse Colic data set, which contains originally missing values.

The tests has been shown that, the proposed HIDD algorithm converges well, it can handle the missingness directly in the input and output vectors too. In case of incomplete output the HIDD method converges also well. The proposed algorithm can outperform other, state-of-the-art incomplete data handling techniques in the most of the cases. The algorithm beside the Levenberg-Marquard algorithm contains one additional novelty compared to the preliminary method: impaired data sets are added to the train set with different incompleteness ratio. With this solution the RMSE values can be significantly reduced. Another, side-effect result was that the proposed HIDD algorithm is able to reduce model overfitting, too.

The presented research results proved the applicability of the novel algorithm for training and application of artificial neural networks on incomplete data extending the Levenberg-Marquardt learning algorithm.