Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

IDP Institut für Datenanalyse
und Prozessdesign

Imputation of Missing Values: Deep Learning vs. Traditional Methods

Diplomand/in
Roberto Barbieri

Dozent
Matthias Templ

In vielen Datensätzen, vor allem der Bereiche Statistik, Vermessungs-analyse und Industriemessung, treten fehlende Werte auf. Oft wird das Löschen dieser Werte nach dem «Listwise Deletion»-Ansatz durchgeführt, sodass der Datensatz in der Grösse und damit auch in der Genauigkeit der Information reduziert wird. Die beste Möglichkeit, fehlende Werte in einem Datensatz zu vermeiden, besteht darin, zu verhindern, dass diese entstehen. Dies ist allerdings selten möglich. Um trotzdem gute Ergebnisse bei der Datenanalyse zu erzielen, ist der Prozess der Imputation wichtig, bei welchem fehlende Werte in einem Datensatz durch Substitutionen ersetzt werden. Diese Masterarbeit beschäftigt sich mit verschiedenen Methoden der Imputation, mit dem Ziel, deren Stärken für diverse Datensatztypen aufzuzeigen. Das Hauptziel ist es, die Unter-schiede zwischen traditionellen und «Deep Learning»-Methoden der Imputation aufzuzeigen. Da Datensätze mit unterschiedlichen Eigenschaften verschiedene Methoden der Imputation erfordern, werden in dieser Arbeit mehrere Möglichkeiten gezeigt, wie Substitutionen für feh-lende Werte gefunden und verglichen werden können. Die Simulations-studie in dieser Arbeit umfasst somit Variationen der Anzahl der Beobachtungen, der Anzahl der Variablen, des Prozentsatzes fehlender Werte sowie des Kontaminationsgrades durch Ausreisser. Darüber hin-aus wird die Auswertung der Ergebnisse auf vier verschiedene Arten vorgenommen und folgende Evaluationsmasse verwendet: Der mittlere absolute prozentuale Fehler, der prozentuale Änderungsfehler, die Korrelationsänderung und die Deckungsrate. Zusätzlich zu den modellbasierten Daten werden zwei reale Datensätze verwendet, welche Informationen über die Qualität der Imputationen liefern. Die Ergebnisse zeigen, dass die neuronalen Netzwerk-Methoden, die «Self organizing Map» von Kohonen und die erarbeitete Miss Keras-Methode, welche auf Keras basiert, nicht nur gute Ergebnisse bei Problemen der Imputation erzeugen. Zudem haben die Deep Learning-Methoden aufgrund der hohen Rechenzeit keine Vorteile gezeigt. Traditionelle Methoden sind schneller und können den Vorteil haben, dass keine vorherige Imputation erforderlich ist. Die Schlussfolgerung der in dieser Masterarbeit durchgeführten Analyse ist, dass bei den meisten Datensätzen, wenn die korrekte traditionelle Methode verwendet wird, die berechneten Imputationen dieselbe Qualität wie bei komplexeren Methoden erreichen.

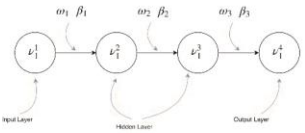

Abbildung 1 zeigt eine einfache Struktur eines neuronalen Netzwerks mit vier Schichten. Eine Eingabe-Schicht, zwei Verborgene Schichten und eine Ausgabe-Schicht. Die allgemeinen Positionen der Neuronen, der Gewichte und der Bias sind dargestellt.

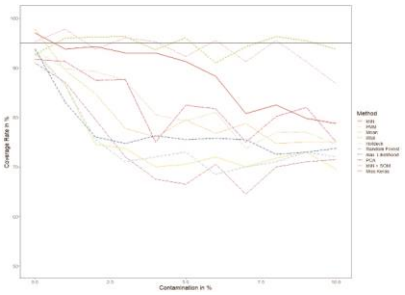

Abbildung 2 zeigt die Deckungsraten von 100 Wiederholungen für die verschiedenen Methoden in kontaminierten Datensätzen. Es ist deutlich, dass nur wenige Methoden wie IRMI-, SOM- und kNN-Imputation annehmbare Deckungsraten von nahezu 95 % erreichen.