Datenwelten entdecken: Muster erkennen mit Korrelation, PCA, Heatmaps und mehr

Cornelia Thieme, Hexagon, Garching

Wenn Sie einen großen Datensatz aus Messungen, Herstellprozessen oder Simulationen erhalten und schnell einen Überblick benötigen, welche Informationen darin stecken – dann helfen statistische Methoden. Vielleicht handelt es sich um Daten als Input fürs Machine Learning, wo der Datenaufbereitung eine wichtige Rolle zukommt. Oder Sie möchten herausfinden, welche Parameter aus einem Fertigungsprozess oder einer FEM-Simulation das Ergebnis beeinflussen. Der Vortrag zeigt verschiedene Darstellungsmöglichkeiten und ihren Nutzen, um einen Datensatz schnell beurteilen zu können.

Um alle Daten auf einen Blick darzustellen, Ausreißer und fehlerhafte Daten zu finden, eignen sich Color Maps und Parallel Coordinates Plots. So kann man in einem ersten Schritt die Daten bereinigen.

Um Abhängigkeiten dazustellen, kann die Korrelationsanalyse verwendet werden. Sie zeigt nicht nur, wie hoch die Abhängigkeit der Ergebnisse von den Parametern ist, sondern auch die Abhängigkeit der Parameter untereinander, also die Qualität der DOE. Vielleicht ist aufgrund physikalischer Voraussetzungen eine gleichmäßige Parameterverteilung (DOE) über den gesamten Datenraum gar nicht möglich.

Die Daten können in verschiedene Cluster eingeteilt werden. Wenn die Daten z.B. von zwei unterschiedlichen Maschinen stammen, könnte der Clusterplot nahelegen, die Daten als zwei separate Datensätze auszuwerten.

Die Hauptkomponentenanalyse (PCA) ist eine statistische Methode zur Reduzierung der Dimensionalität von Datensätzen. Es wird ein Hauptachsensystem durch die Daten gelegt, dadurch werden die wichtigsten Abhängigkeiten erkennbar.

Heatmaps bieten eine effektive visuelle Darstellung von Datensätzen. Sie zeigen die Intensität von Werten in einer Farbskala mit Höhenlinien an und können zur Identifizierung von Mustern, Ausreißern und Clustern verwendet werden.

Der Vortrag zeigt diese und weitere Methoden anhand von praktischen Beispielen in der Machine Learning Software Odyssee.