ABSTRACT

Data farming is a methodology that enables artificial data generation where real-world data is difficult or impossible to obtain. Artificial data generation using full factorial simulation means generating predictions for all combinations of possible values for input variables. Therefore, computational intensity is raising exponentially with the number of considered influences on a phenomenon.
In data farming, the computing effort is reduced, among other things, by the fact that only a partial simulation takes place. Variable combinations to be simulated are determined using a strategic selection process. This can enable data generation in situations where it might not be economical otherwise.

In order to still have a broad database available for further analyzes and evaluations (e.g. the observation of main influencing factors and pareto-optimal solutions), this technology is combined with machine learning methods in the concept proposed here. The concept provides a strategic selection of the scenarios and their simulation by software. The data obtained serves as training data for a machine learning model, which can use this as a basis to forecast further data. The main influencing factors and pareto-optimal solutions are to be determined with the database obtained.

A prototype implementation of the model can be evaluated using a complete simulation. The application of the concept to another question was also examined.

Furthermore, the possible implementation of the concept with an active learning approach is discussed.

Keywords— Data Farming, Main Influencing Factors Analysis, Machine Learning, Design of Experiment, Active Learning

ZUSAMMENFASSUNG

Data Farming ist eine Methodik, die künstliche Datenerzeugung in Situation ermöglicht, in denen Real-World-Daten nicht oder nur schwer erhältlich sind. Eine künstliche Datenerzeugung mittels vollfaktorieller Simulation bedeutet, Vorhersagen für *alle* Kombinationen aus Variablenausprägungen zu generieren. Daher ist die Rechenintensität umso höher, je mehr Einflüsse auf ein Phänomen betrachtet werden.
Im Data Farming wird Rechenaufwand unter anderem dadurch gesenkt, dass nur eine teilweise Simulation stattfindet. Mittels eines strategischen Auswahlprozesses werden Variablenkombinationen bestimmt, welche simuliert werden. Dadurch kann die Datenerzeugung in Situationen ermöglicht werden, in denen dies anderweitig womöglich nicht wirtschaftlich wäre.

Um dennoch eine breite Datenbasis für weitergehende Analysen und Auswertungen (etwa der Beobachtung von Haupteinflussfaktoren und pareto-optimalen Lösungen) zur Verfügung zu haben, wird diese Technik im hier vorgeschlagenen Konzept mit Machine Learning-Methoden kombiniert. Das Konzept sieht dabei eine strategische Auswahl der Szenarien und deren Simulation durch eine Software vor. Die erhaltenen Daten dienen einem Machine Learning- Modell als Trainingsdaten, welches auf dieser Basis weitere Daten prognostizieren kann. Mit der erhaltenen Datenbasis sollen Haupteinflussfaktoren sowie pareto-optimale Lösungen ermittelt werden.

Eine prototypische Implementierung des Modells konnte anhand einer vollständigen Simulation evaluiert werden. Die Anwendung des Konzepts auf eine weitere Fragestellung wurde ebenfalls untersucht.

Des Weiteren wird die mögliche Umsetzung des Konzepts mit einem aktiven Lernansatz diskutiert.

Schlagworte— Data Farming, Haupteinflussfaktorenanalyse, Machine Learning, Design of Experiment, Active Learning