Abstract

In commercial vehicle development at Daimler Truck AG, large amounts of multivariate time series data are generated during the development and testing of prototypes by recording thousands of sensor data. The ability to extract relevant dependencies from complex data sets is crucial for optimization and troubleshooting in the development of trucks. The thesis deals with the development and evaluation of a novel approach for the identification of non-linear causal relationships on these data sets. The developed approach *tsVOI - Time-shifted Variation of Information* aims to overcome the challenges of analyzing high-dimensional signals by reducing the complexity in a three-step approach. The first step consists of identifying candidate pairs with a potential causal relationship using the entropy-based metric *Variation of Information*. The subsequent examination of the signal pairs serves to determine the direction of information flow by highlighting the pairs' time delays that maximize their dependence. Finally, causal validation verifies any relationships by checking the influence of common causal signals and orienting any remaining undirected edges in the causal graph by analyzing the observed independencies. The evaluation of the approach on synthetic data shows promising results. The algorithm proved to be robust to different levels of noise when identifying the time delays and was able to effectively capture both linear and non-linear dependencies. Despite the observed strengths, such as the ability to identify relevant lags and causalities, limitations exist, especially with regard to the assumption of the absence of unobserved confounding factors. The assumption can limit the performance of the approach in real scenarios where spurious correlations could occur due to hidden confounding factors. The findings of the research illustrate the potential of the proposed method to optimize sub-steps of root cause analysis and to expand the understanding of analyzed data structures.

Zusammenfassung

Im Rahmen der Entwicklung und Erprobung von Prototypen im Bereich der Nutzfahrzeuge entstehen bei der Daimler Truck AG durch die Aufzeichnung tausender Sensorinformationen große Mengen multivariater Zeitreihendaten. Die Fähigkeit, aus komplexen Datensätzen relevante Zusammenhänge extrahieren zu können, ist für die Optimierung und Fehlersuche in der Entwicklung von Lastkraftwagen von entscheidender Bedeutung. Die vorliegende Arbeit befasst sich mit der Entwicklung und Bewertung eines neuartigen Ansatzes zur Identifikation nichtlinearer Kausalzusammenhänge auf diesen Datenmengen. Der entwickelte Ansatz *tsVOI - Time-shifted Variation of Information* zielt darauf ab, die Herausforderungen der Analyse von hochdimensionalen Signalen zu bewältigen, indem die Komplexität in einem dreistufigen Verfahren reduziert wird. Dieses besteht im ersten Schritt, unter der Nutzung der entropiebasierten Metrik *Variation of Information*, aus der Identifikation von Kandidatenpaaren mit potenziellem Kausalzusammenhang. Anschließend erfolgt eine Untersuchung der Signalpaare hinsichtlich etwaiger zeitlicher Verzögerungen und deren Zusammenhang, um die Richtung des Informationsflusses zu bestimmen. Die Kausalvalidierung verifiziert abschließend jegliche Beziehungen, indem der Einfluss gemeinsamer Ursachensignale überprüft und verbleibende ungerichtete Kanten im Kausalgraphen durch die Auswertung beobachteter Unabhängigkeiten orientiert werden. Die Evaluierung des Ansatzes anhand synthetischer Daten liefert vielversprechende Ergebnisse. Der Algorithmus erwies sich bei der Bestimmung der zeitlichen Verzögerungen als robust gegenüber verschiedenen Rauschstärken und war in der Lage, sowohl lineare als auch nichtlineare Abhängigkeiten effektiv zu erfassen. Trotz der beobachteten Stärken, wie der Fähigkeit zur Identifikation relevanter Lags und Kausalitäten, existieren Einschränkungen, insbesondere im Hinblick auf die Annahme der Abwesenheit unbeobachteter Störfaktoren. Die Annahme kann die Leistungsfähigkeit des Ansatzes in realen Szenarien, bei denen Scheinkorrelationen durch versteckte Störfaktoren auftreten könnten, einschränken. Die Auswertungen der Untersuchungen demonstrieren das Potenzial des vorgestellten Ansatzes, Teilschritte der Ursachenforschung zu optimieren und das Verständnis analysierter Datenstrukturen zu erweitern.