Autonome Identifikation relevanter Zusammenhänge in hochdimensionalen Zeitreihendaten zur Generierung von Clustern und Graphen im Kontext der Nutzfahrzeugentwicklung

Marcel Mauel[1,2], Prof. Dr. Timo Schürg[1], Prof. Dr. Florian Heinrichs[1], M.Sc. Carsten Binz[2]

[1]Hochschule Darmstadt, Fachbereich Mathematik und Naturwissenschaften & Informatik [2]Daimler Truck AG

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Motivation

Während der Entwicklung und Erprobung von Prototypen fallen bei der Daimler Truck AG durch die Aufzeichnung tausender Sensorinformationen große Mengen multivariater Zeitreihendaten an. Die Fähigkeit, kausale Zusammenhänge aus komplexen Datensätzen zu extrahieren, ist für die Optimierung und Fehlersuche in der Lkw-Entwicklung von entscheidender Bedeutung. So können Teilschritte der Root Cause Analysis durch die Vorauswahl potenzieller Einflussfaktoren optimiert und das Verständnis der betrachteten Datenstrukturen erweitert werden.

Im Gegensatz zu bisherigen Forschungsarbeiten, in denen zur Validierung kausalanalytischer Ansätze häufig geringe Datenmengen mit wenigen Dimensionen analysiert werden, liegen die Daten realer Anwendungen häufig in hoher Dimensionalität als multivariate Zeitreihendaten vor. Das Ziel der Arbeit ist die Entwicklung und Evaluierung eines Ansatzes, der die Identifikation von linearen und nichtlinearen kausalen Zusammenhängen im Falle tausender Signale ermöglicht und die Bestimmung der zeitlichen Verzögerungen (Lags) zwischen gefundenen Kausalitäten erlaubt.

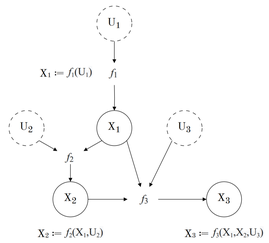

Abbildung 1. Structural Causal Model (vgl. [2])

Causal Discovery

Ziel der Kausalanalyse ist es, aus Beobachtungsdaten Ursache-Wirkungs-Beziehungen zu identifizieren, bei denen die Veränderung eines Signals die Veränderung eines anderen Signals beeinflusst. Die Methoden der Kausalanalyse lassen sich in vier Hauptkategorien unterteilen:

1. **Constraint-basierte Ansätze** verwenden Unabhängigkeitstests, um Kausalbeziehungen zu validieren.
2. **Score-basierte Ansätze** bewerten verschiedene Modelle durch Bewertungsfunktionen und optimieren die Graphenstruktur, um das wahrscheinlichste Modell zu identifizieren.
3. **SCM-basierte Ansätze** treffen mathematische Annahmen über die Beziehungen und die Verteilung der Daten, um diese mathematisch abzubilden.
4. **Granger Causality** [1] basiert auf der Vorhersagekraft eines Signals für ein anderes und der daraus ableitbaren Kausalität - wird häufig mittels Machine Learning umgesetzt.

Aufgrund des exponentiellen Wachstums des Suchraumes möglicher Beziehungen bei steigender Signalanzahl und strenger Annahmen über die zugrunde liegenden wahren Beziehungen sind diese Methoden in der Regel auf niedrigdimensionale Daten beschränkt.

Darüber hinaus stellen die Bestimmung von nichtlinearen Zusammenhängen und die Identifikation von Zusammenhängen mit großer zeitlicher Verzögerung (Lag) große Herausforderungen dar, die die Suche erschweren. Das Temporal Causal Discovery Framework (TCDF) ist ein auf Granger-Kausalität basierender Ansatz, dessen Ziel die Vorhersage einzelner Signale mittels Convolutional- und Attention-Layer ist. Hierbei wird versucht, die Graphenstruktur aus den trainierten Gewichten abzuleiten. Zeitliche Verzögerungen zwischen Kausalitäten werden aus den Attention-Layern extrahiert. Für die Identifikation komplexer Zusammenhänge werden jedoch große Layer benötigt, insbesondere bei steigender Signalanzahl N. Zudem erfordert die Bestimmung des Gesamtgraphen das Training N dieser neuronalen Netze.

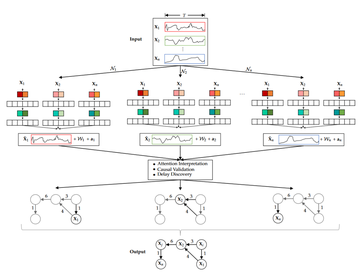

Abbildung 2. Temporal Causal Discovery Framework (TCDF) [5]

Entwickelter Ansatz: tsVOI

Der entwickelte Ansatz *tsVOI - Time-shifted Variation of Information* zielt darauf ab, die Komplexität der explorativen Kausalanalyse in einem dreistufigen Verfahren zu reduzieren. Im ersten Schritt werden lineare und nichtlineare Zusammenhänge mit Hilfe der entropiebasierten Metrik *Variation of Information* identifiziert. Anschließend werden die gefundenen Signalpaare durch die Simulation einer zeitlichen Verschiebung auf die Stärke des Zusammenhangs untersucht, um die Richtung des Informationsflusses sowie die zeitliche Verzögerung zwischen Ursache und Wirkung zu bestimmen. Die Kausalvalidierung verifiziert im dritten Schritt jegliche Beziehung, indem der Einfluss gemeinsamer Ursachensignale unter Nutzung der *Conditional Variation of Information* geprüft und verbleibende ungerichtete Kanten im Kausalgraphen durch die Auswertung beobachteter Unabhängigkeiten mittels des PC-Algorithmus [3] orientiert werden.

Abbildung 3. Ablaufdiagramm der entwickelten Methodik *tsVOI* und eine beispielhafte Anwendung

Die Entropie (H) ist ein Maß zur Berechnung des erwarteten Informationsgehalts von Zufallsvariablen. Sie bildet die Grundlage für die Berechnung der *Variation of Information* (VOI), welche im Gegensatz zur Mutual Information (I) eine Metrik darstellt. Die VOI kann verwendet werden, um den Unterschied zwischen zwei Zufallsvariablen (ZV) X und Y, basierend auf ihrer gemeinsamen Information, zu bewerten, um lineare und nichtlineare Beziehungen aufzudecken.

Für die ZV $X = \{x_1, x_2, \ldots, x_m\}$ mit den Wahrscheinlichkeiten $p_i = P(X = x_i)$ ist die Entropie definiert als:

$$\mathrm{H}(X) = E[I] = -\sum_{i=1}^{m} p_i \log_2 p_i$$

Äquivalent dazu erfolgt die Berechnung der gemeinsamen Entropie der ZV X, Y über die gemeinsame Wahrscheinlichkeitsverteilung. Die Bedingte Entropie und Mutual Information sind definiert als:

$$\mathrm{H}(Y|X) = \mathrm{H}(X,Y) - \mathrm{H}(X)$$
$$\mathrm{I}(X;Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y)$$

Die Variation of Information ergibt sich aus der Differenz zwischen gemeinsamer Entropie und Mutual Information bzw. der Summe der bedingten Entropien:

$$\mathrm{VOI}(X,Y) = \mathrm{H}(X,Y) - \mathrm{I}(X;Y)$$
$$= \mathrm{H}(X \mid Y) + \mathrm{H}(Y \mid X)$$

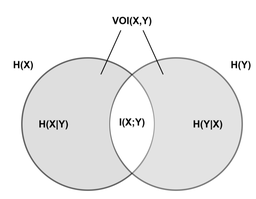

Abbildung 4. Variation of Information in grau schraffiert (vgl. [4])

Ergebnisse

Zur Evaluierung des entwickelten Algorithmus *tsVOI* werden synthetische Daten verwendet, die den realen Sensordaten der Lastkraftwagen nachempfunden sind. Sie ermöglichen die Simulation linearer und nichtlinearer Zusammenhänge, die Nachbildung verschiedener Kausalstrukturen, die Beimischung von Rauschen sowie die Erprobung unterschiedlicher zeitlicher Verzögerungen τ zwischen Ursache und Wirkung. Die Bewertung der identifizierten Kausalstrukturen erfolgte über zwei Varianten des F1-Scores, wobei der Skeleton F1-Score $(\mathrm{F1})$ die gefundenen Strukturen ohne Berücksichtigung der Orientierung der Kanten bewertet und der Directed F1-Score $(\overrightarrow{\mathrm{F1}})$ die Orientierung der Kanten einbezieht. Die Auswertung der identifizierten Lags erfolgte mittels des RMSE.

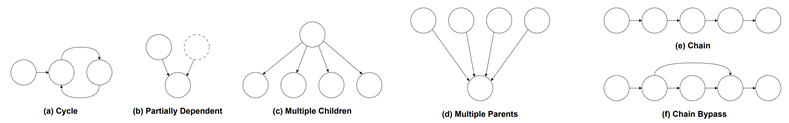

Abbildung 5. Beispiele evaluierter Kausalstrukturen

Die Ergebnisse bestätigen den Einsatz der entropiebasierten Metrik VOI, die in den Untersuchungen die Erkennung nichtlinearer Zusammenhänge ermöglichte. Die simulierten Kausalstrukturen konnten auch nach Variation der Basisstrukturen und des Rauschpegels erfolgreich erkannt werden. Die Anwendung des PC-Algorithmus erweist sich insbesondere bei Strukturen als nützlich, bei denen viele Signale auf das gleiche Signal einwirken. Es ist zu beobachten, dass die Identifikation des Informationsflusses durch Maximierung der geteilten Information bei Verschiebung der Kandidatenpaare und die daraus resultierende Bestimmung der Lags, zuverlässig funktioniert.

		$\tau = 1$		$\tau = 20$		$\tau = 600$	
Structure	Subcategory	F1	$\overrightarrow{\mathrm{F1}}$	F1	$\overrightarrow{\mathrm{F1}}$	F1	$\overrightarrow{\mathrm{F1}}$
Cycle		0.50	0.00	0.80	0.33	0.67	0.50
Partially Dep.	40%	1.00	0.67	1.00	1.00	1.00	1.00
	80%	1.00	0.67	1.00	1.00	1.00	1.00
Mult. Children	Var. Parameters & Time Delay	0.25	0.25	0.25	0.25	0.22	0.22
	Var. Functions	0.62	0.44	0.73	0.62	0.33	0.33
Mult. Parents	3 Parents	1.00	1.00	1.00	1.00	0.50	0.50
	4 Parents	1.00	1.00	1.00	1.00	0.00	0.00
	5 Parents	1.00	1.00	1.00	1.00	0.00	0.00
	4 Parents, Var. Time Delay	1.00	1.00	1.00	0.75	0.00	0.00
	4 Parents, Var. Functions	1.00	1.00	1.00	1.00	0.40	0.40
Chain	6 Links	1.00	1.00	1.00	1.00	0.71	0.71
	6 Links, Var. Time Delay	1.00	1.00	0.53	0.53	0.57	0.57
	6 Links, Var. Functions	0.47	0.44	0.62	0.62	0.60	0.60
Chain BP	6 Links	0.91	0.91	0.91	0.91	0.67	0.67
	6 Links, Var. Time Delay	0.91	0.91	0.60	0.60	0.50	0.50
	6 Links, Var. Functions	0.50	0.50	0.63	0.63	0.55	0.55

Tabelle 1. Erkennung von Kausalstrukturen mit Lag $\tau \in \{1,20,600\}$, Noise a=0.00

Zusammenfassung

Die Untersuchungen auf synthetischen Daten zeigen auf, dass der entwickelte Ansatz in der Lage ist, relevante Zusammenhänge, unter Variation verschiedener Faktoren der Datenbasis erfolgreich zu identifizieren. Im Rahmen der Arbeit erfolgte eine Evaluierung verschiedener Faktoren, darunter die Anpassung des Rauschniveaus, die Integration nichtlinearer Transformationen sowie die Auswertung unterschiedlicher Lags. Es konnte dabei gezeigt werden, dass die Identifikation von Lags eine hohe Robustheit gegenüber verrauschten Daten aufweist. Der Algorithmus ist in der Lage, die Richtung des Informationsflusses korrekt zu bestimmen, sofern Abhängigkeiten erkannt werden. Die erzielten Ergebnisse sind vielversprechend, insbesondere im Hinblick auf die Fähigkeit des Algorithmus, sowohl lineare als auch nichtlineare Abhängigkeiten zu erkennen. Eine kritische Betrachtung des Ansatzes im Hinblick der Übertragung auf reale Anwendungsfälle zeigt, dass Einschränkungen hinsichtlich der Abwesenheit externer Einflussfaktoren bestehen. Die Arbeit demonstriert die Möglichkeit der Kausalanalyse auf hochdimensionalen Zeitreihendaten und zeigt das Potential auf, die erkannten Zusammenhänge zur Vorselektion relevanter Signale in der Root Cause Analysis zu nutzen.

[1] C. W. J. Granger "Investigating Causal Relations by Econometric Models and Cross-spectral Methods" (1969) DOI:10.2307/1912791
[2] Chang Gong u. a. "Causal Discovery from Temporal Data: An Overview and New Perspectives" (2023) DOI: 10.48550/arXiv.2303.10112
[3] Peter Spirtes, Clark Glymour, Richard Scheines "Causation, Prediction, and Search" (1993) ISBN: 978-1-4612-7650-0
[4] Marina Meila "Comparing Clusterings—an Information Based Distance" (2007) DOI: 10.1016/j.jmva.2006.11.013
[5] Meike Nauta, Doina Bucur, Christin Seifert "Causal Discovery with Attention-Based Convolutional Neural Networks" (2007) DOI: 10.3390/make1010019
[6] Charles K. Assaad, Emilie Devijver und Eric Gaussier "Survey and Evaluation of Causal Discovery Methods for Time Series" (2022) DOI: 10.1613/jair.1.13428