Lean Data – Anwendungsspezifische Reduktion großer Datenmengen im Produktionsumfeld

Marina Baucks,*
Marcus Mau,
Peter Ruppelt,
Alexander Puchta und
Jürgen Fleischer

Der Beitrag untersucht das Konzept „Lean Data“, das darauf abzielt, große Datenmengen in Produktionsumgebungen effizient zu reduzieren. Mithilfe einer Toolbox zur datengetriebenen Auswahl spezifischer Reduktionsmethoden wird die Datenkomplexität minimiert, ohne relevante Informationen zu verlieren. Der Ansatz kombiniert Methoden der Datenkompression und -reduktion und berücksichtigt gleichzeitig anwendungsspezifische Anforderungen, wie z. B. den zeitlichen Kontext und die Wiederherstellungsgenauigkeit. Experimente zeigen, dass angepasste Reduktionsmethoden signifikante Speicher- und Analysevorteile bieten.

Einleitung und Hintergrund

Die Digitalisierung hat die Art und Weise, wie wir Daten sammeln und nutzen, grundlegend verändert. Insbesondere in der Industrie 4.0 hat die Vernetzung von Objekten und die Sammlung von Daten auf der Feldebene entscheidend an Bedeutung gewonnen [1]. Dieser Prozess verleiht der Datenerfassung und -verarbeitung eine neue Dimension und eröffnet gleichzeitig eine Vielzahl von Chancen und Herausforderungen im Bereich der Aggregation und der Verwendung von Daten für produzierende Unternehmen [2]. In Produktionssystemen erfolgt die Datensammlung in der Regel durch Sensoren, die in verschiedenen Anlagen und Geräten eingebettet sind. Diese Sensoren sammeln kontinuierlich Informationen, beispielsweise zu Prozessgrößen, Maschinenzuständen oder Werkstückparametern. Die Anbindung dieser Sensoren über ein Netzwerk ermöglicht es, die gesammelten Daten in Echtzeit zu übertragen und zentral zu speichern. Diese Daten bilden die Grundlage für datengesteuerte Entscheidungen und Prozesse. Die Qualität und der Umfang der Datengrundlage sind hierbei entscheidend für die Wertschöpfung aus den Daten [3].

Große Datenmengen (englisch: Big Data) können die bestehende Infrastruktur jedoch vor technische Herausforderungen stellen, da die Daten nicht nur erfasst und gespeichert, sondern auch sicher und effizient verarbeitet werden müssen. Um diesen Herausforderungen zu begegnen, wurde das Konzept von „Lean Data“ ins Leben gerufen. Lean Data bezieht sich auf Ansätze und Methoden zur Reduktion der Datenmenge und -komplexität bei gleichzeitigem Erhalt der relevanten Informationen zum Zeitpunkt der Verarbeitung bzw. Speicherung. Es orientiert sich an den Prinzipien des Lean Managements, das Verschwendung minimiert und Effizienz maximiert [4]. Im Kontext von Lean Data werden verschiedene Techniken zur Datenkompression und -reduktion betrachtet, die darauf abzielen, die Verarbeitungs- und Speicheranforderungen großer Datensätze zu verringern, ohne die Qualität und Nützlichkeit der Daten zu beeinträchtigen. Zu den behandelten Methoden gehören sowohl verlustfreie als auch verlustbehaftete Kompressionstechniken sowie Methoden zur Reduktion der Datenmenge und -dimension.

* **Korrespondenzautorin**
Marina Baucks, M. Sc.; wbk – Institut für Produktionstechnik, Karlsruher Institut für Technologie (KIT); Kaiserstr. 12, 76131 Karlsruhe; Tel.: +49 (0) 1523 9502566, E-Mail: marina.baucks@kit.edu

Weitere Autoren
Marcus Mau, M. Sc.; wbk am Karlsruher Institut für Technologie (KIT)
Peter Ruppelt, M. Sc.; wbk am Karlsruher Institut für Technologie (KIT)
Alexander Puchte, M. Sc.; wbk am Karlsruher Institut für Technologie (KIT)
Prof. Dr.-Ing. Jürgen Fleischer; wbk am Karlsruher Institut für Technologie (KIT)

Hinweis
Bei diesem Beitrag handelt es sich um einen von den Advisory-Board-Mitgliedern des ZWF-Sonderheftes wissenschaftlich begutachteten Fachaufsatz (Peer-Review).

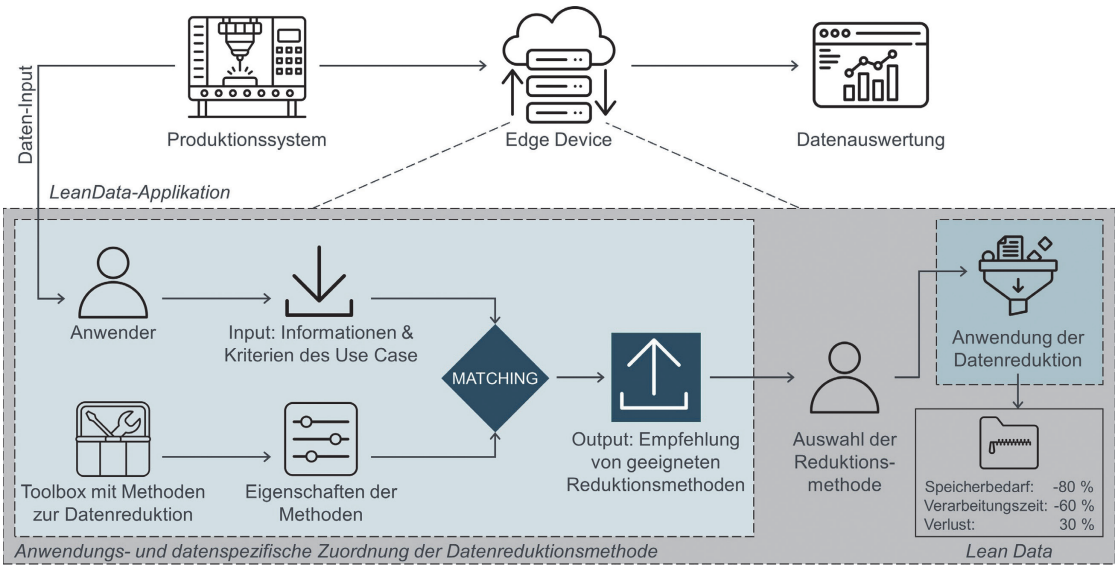

Bild 1. Lean Data Toolbox [I1–I10]

Ziel dieses Beitrags ist es, das Konzept der neu entwickelten Lean Data Toolbox vorzustellen, eine Methodik zur anwendungsspezifischen Auswahl geeigneter Datenreduktionsmethoden für Produktionssysteme. Der Hauptbeitrag besteht in der Entwicklung eines datenbasierten Verfahrens, welches mithilfe charakteristischer Merkmale der Daten und ihres Verwendungszwecks eine optimierte Auswahl einer geeigneten Datenreduktionsmethode ermöglicht. Dieses Konzept wird anhand einer experimentellen Validierung im Kontext der datengetriebenen Zustandsüberwachung demonstriert.

Im weiteren Verlauf des Beitrags wird zunächst das Konzept der Lean Data Toolbox erläutert, das eine datengestützte und anwendungsspezifische Auswahl von Reduktionsmethoden ermöglicht. Dazu wird die Funktionsweise der Toolbox, einschließlich der eingesetzten Matching-Algorithmen und der relevanten Datenmerkmale, vorgestellt. Der darauf folgende Abschnitt beschreibt die zugrunde liegenden Merkmale der Produktionsdaten und deren Ausprägungen, die für die Auswahl geeigneter Reduktionsmethoden entscheidend sind. Darauf aufbauend werden verschiedene Datenreduktionsmethoden detailliert diskutiert, einschließlich ihrer Wirkprinzipien und typischer Einsatzmöglichkeiten. Der darauf folgende Abschnitt enthält eine experimentelle Validierung der Lean Data Toolbox, bei der exemplarisch die datengetriebene Zustandsüberwachung als Anwendungsfall untersucht wird. Hier werden sowohl die Methodik als auch die Ergebnisse der Evaluation dargestellt, um die Praktikabilität des Konzepts zu verdeutlichen. Abschließend sind eine Zusammenfassung der Erkenntnisse sowie ein Ausblick auf mögliche Weiterentwicklungen und zukünftige Anwendungen im Produktionsumfeld zu finden.

Konzept der Lean Data Toolbox

Um Lean Data für Produktionssysteme anwendbar zu machen, müssen die Anlagen als Datenquellen, die Speichersysteme sowie die Anforderungen der Anwender miteinander verknüpft werden. Dazu wird eine Toolbox als Edge-Anwendung entwickelt, welche in Bild 1 dargestellt ist.

Um die Maschinendaten, die auf dem Edge Device eingehen, gemäß dem Lean-Data-Prinzip zu reduzieren, stellt der Nutzer Informationen zu den Daten selbst und zu ihrem zukünftigen Verwendungszweck bereit. Mithilfe dieser Informationen wird durch einen Matching-Algorithmus eine spezifische Datenreduktionsmethode ermittelt, deren Eigenschaften den Anforderungen, die durch den Nutzer und die Beschaffenheit der Daten gestellt werden, entsprechen. Diese Reduktionsmethode wird nach einer Kontrolle und Bestätigung durch den Nutzer angewandt, sodass die Daten nun in reduzierter Form an das Speichermedium übertragen werden können. Sollen die Daten zu einem späteren Zeitpunkt in ihrem im Vorhinein spezifizierten Verwendungszweck eingesetzt werden, werden die reduzierten Daten, sofern benötigt, mit einer zum ursprünglichen Reduktionsverfahren passenden Methode wiederhergestellt. Soll beispielsweise der Verschleiß einer Maschinenachsenkomponente gemessen werden, könnte der Trend des Motorstroms über einen längeren Zeitraum hinweisgebend sein [5]. Die Zeitreihe muss also nicht unbedingt hochauflösend wiederhergestellt werden, allerdings muss der zeitliche Zusammenhang in den Daten erhalten bleiben. Sollen kurzzeitige Anomalien im Maschinenverhalten analysiert werden, sind wiederum besonders die hochfrequenten Anteile der Daten von Interesse [6].

Anwendungsspezifische Auswahl von Datenreduktionsmethoden

Um zu identifizieren, welche Merkmale von Datensätzen aus Produktionsumgebungen sich dazu eignen, eine passende Datenreduktionsmethode auszuwählen, wurden 21 Use Cases aus dem Industrie- sowie dem Forschungsumfeld untersucht. Dabei wurde unter anderem ermittelt, in welchem Kontext und an welcher Art von Anlage die Daten anfallen, welche Art von Daten vorliegt und was das Ziel der Datenaufnahme ist. Anhand dieser Use Cases sowie der Eigenschaften der Datenreduk-

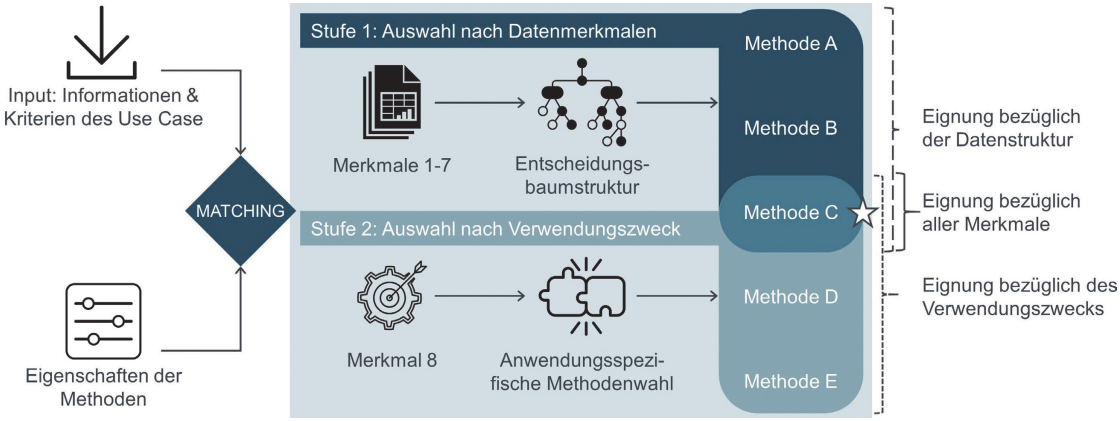

Bild 2. Konzept des Matching-Algorithmus zur Wahl einer geeigneten Reduktionsmethode [I9–I14]

tionsmethoden konnten acht Merkmale festgelegt werden, mittels derer sich zum Use Case passende Reduktionsmethoden ermitteln lassen. Zudem wurde der Betrachtungsraum bezüglich der Datenart auf Zeitreihendaten begrenzt, da diese am häufigsten im Produktionskontext vorliegen. Die ermittelten Merkmale zeichnen sich folgendermaßen aus:

- Priorisierung von Wiederherstellungsgenauigkeit bzw. Kompressionsrate durch den Anwender,
- Relevanz der Beibehaltung des zeitlichen Kontextes in der Datenreihe,
- Redundanz in den Daten,
- Dimensionalität der Zeitreihe (univariat/multivariat),
- kausale Zusammenhänge innerhalb von multivariaten Zeitreihen,
- zeitliche Verschiebung zwischen zyklischen Komponenten innerhalb eines Signals oder zwischen ähnlichen, miteinander verbundenen Signalen,
- wiederholt auftretende Muster in den Daten sowie
- Verwendungszweck für die Daten.

Merkmal 1 lässt sich auf einer Skala von 1 (...) bis 5 (...) angeben. Alle anderen Merkmale werden durch den Nutzer mit „ja“ (1), „keine Angabe“ (0) oder „nein“ (2) quantifiziert.

Über einen Matching-Algorithmus lassen sich dem jeweiligen Anwendungsfall anhand der Ausprägung seiner Merkmale passende Datenreduktionsmethoden zuordnen. Es wurden in der vorliegenden Implementierung neun Reduktionsmethoden (vgl. Bild 3) untersucht. Das Konzept des Matching-Algorithmus ist in Bild 2 dargestellt.

Während sich die Merkmale 1 bis 7 auf die Daten selbst beziehen, bezieht sich Merkmal 8 auf den späteren Verwendungszweck der Daten. Dieses Merkmal hat einen signifikanten Einfluss auf die Methodenwahl. Allerdings kann es jedoch sein, dass der Verwendungszweck zum Zeitpunkt der Datenaufnahme nicht bekannt ist und daher nicht spezifiziert werden kann. Daher wurde der Matching-Algorithmus zweistufig aufgebaut: In der ersten Stufe werden mittels einer Entscheidungsbaumstruktur auf Basis der Merkmale 1 bis 7 die drei geeignetsten Reduktionsmethoden für die vorliegenden Daten identifiziert.

Die Abbildung der Entscheidungslogik in Form eines von Experten erstellten Entscheidungsbaums wurde aufgrund der Vielzahl an Parametern und deren Abhängigkeiten als nicht praktikabel bewertet. Als vielversprechender Ansatz wurde stattdessen der Einsatz eines maschinellen Lernverfahrens, insbesondere eines Random Forest (RF), identifiziert. Auf Basis dieser Erkenntnis wurde ein Algorithmus entwickelt, der mittels des RF-Ansatzes in der Lage ist, für jeden Use Case die am besten geeigneten Verfahren zu identifizieren. Der entwickelte Algorithmus liefert drei Vorschläge, die nach einem Fit-Score sortiert werden. Dabei wird dasjenige Verfahren als primäre Empfehlung angezeigt, welches den höchsten Score aufweist. Die Trainingsdaten für den Algorithmus wurden aus Zuordnungen der Antwortmöglichkeiten zu den Kompressionsmethoden erstellt. Durch eine Kombination aller zutreffenden Möglichkeiten wurde ein Datensatz mit 144 Einträgen erstellt. Mittels eines Grid-Searches wurden optimale Werte für die maximale Tiefe der Entscheidungsbäume, die minimale Anzahl an Datenpunkten für einen Split zwischen Merkmalen, die Anzahl der Entscheidungsbäume und die minimale Anzahl an Datenpunkten in einem Blattknoten eines Entscheidungsbaums (Hyperparameter „Max Depth“, „Min Samples Split“, „N Estimators“ und „Min Samples Leaf“) identifiziert. Die Gewichtung der Features, also der Antworten aus dem Fragenkatalog, wird durch den Lernalgorithmus automatisiert optimal festgelegt. Es zeigt sich, dass die Antwort auf „Kompressionseffizienz vs. Datenrekonstruktion“ gefolgt von „Vorhandensein einer zeitlichen Verschiebung“ und „Erhaltung des zeitlichen Kontexts“ den höchsten Einfluss auf das Ergebnis der Vorhersage des Random Forests hat.

In der zweiten Stufe werden die Methoden, die sich am besten für den angegebenen Verwendungszweck eignen, herangezogen. Wenn es Überschneidungen zwischen den „Top 3“-Methoden der ersten Stufe (basierend auf den Datenmerkmalen) und den Methoden der zweiten Stufe (basierend auf dem Verwendungszweck) gibt, werden die gemeinsamen Methoden als priorisierte Optionen vorgeschlagen. Falls es jedoch keine solchen Überschneidungen gibt, kann der Nutzer entscheiden, ob er eine Methode bevorzugt, die besser zur Datenstruktur passt, oder eine, die stärker auf den geplanten Verwendungszweck abgestimmt ist.

Datenreduktionsmethoden

Ansätze zur Datenreduktion können in die Kategorien der Dimensionsreduktion,

der Instanzenreduktion und der Datenkompression klassifiziert werden. Die Methoden der Dimensionsreduktion generieren eine neue Repräsentation der Daten, bei der weniger relevante Merkmale eliminiert werden. Die Kriterien für die Irrelevanz variieren je nach Kontext und dienen dazu, dem „Curse of Dimensionality“ entgegenzuwirken [7]. Mit zunehmender Anzahl von Attributen steigen die Komplexität der Informationen und die für Data-Mining-Algorithmen erforderliche Rechenleistung exponentiell [8]. Eine Reduktion der Merkmale kann den Analyseaufwand, den Speicherbedarf und gegebenenfalls auch das Ergebnis optimieren, da globale Optima in einem verringerten Suchraum schneller gefunden werden können. Methoden der Dimensionsreduktion implizieren jedoch einen Datenverlust und zielen darauf ab, die Daten für eine effizientere Analyse in einer vereinfachten Form zu repräsentieren [8, 9].

Die Instanzenreduktion fokussiert sich auf die Verringerung der Datentupel eines Datensatzes. Hierbei wird zwischen parametrischen und nichtparametrischen Methoden unterschieden. Parametrische Methoden erforschen mathematische Beziehungen zwischen Datenpunkten und definieren Funktionen anhand von Interpolationsverfahren. Die Parameter dieser Funktionen repräsentieren die Daten temporär und erlauben bei Bedarf eine Rekonstruktion der Informationen. Nicht-parametrische Methoden hingegen generieren neue, aggregierte Merkmale oder eliminieren Einträge, um Strukturen und Muster direkt aus den Daten abzuleiten [7].

Die Datenkompression befasst sich mit der temporären Reduktion von Datenmengen, um Speicherplatz zu sparen und Übertragungsgeschwindigkeiten zu erhöhen [10]. Sie identifiziert Wiederholungen innerhalb der Datensätze und kreiert eine alternative, redundanzfreie Darstellung. Das primäre Ziel ist die vollständige und unveränderte Rekonstruktion der Originaldaten aus der komprimierten Darstellung, wobei zwischen verlustfreier und verlustbehafteter Kompression unterschieden wird und letztere mit dem Ziel höherer Kompressionsraten Verluste zulässt [11].

Zusammenfassend zielen die Methoden der Dimensions- und Instanzenreduktion darauf ab, die Effizienz der Datenanalyse zu steigern, wobei ein Verlust von Daten impliziert wird, während die Datenkompression versucht, Speicherressourcen zu schonen und alle Daten unverändert zu erhalten [8, 9].

Anwendungsfall	Ziel der Anwendung	Anforderungen an die Datenreduktion
Datenkompression	Verringerung des Speicherbedarfs	hohe Kompressionsrate, hohe Wiederherstellungsgenauigkeit
Datengetriebene Zustandsüberwachung	Erkennung von Trends im zeitlichen Verlauf der Daten	Erhalt der zeitlichen Abhängigkeit der Daten, keine hohe Wiederherstellungsgenauigkeit erforderlich
Datengetriebene Qualitätsbewertung	Bewertung der Produktqualität anhand der statistischen Verteilung der Daten	hohe Übereinstimmung der statistischen Verteilung innerhalb der Daten vor und nach der Datenreduktion

Tabelle 1. Exemplarische Lean-Data-Anwendungsfälle

Metriken zur Bewertung der Datenreduktion

Zur Evaluierung von Algorithmen sind quantitative Maße unerlässlich und bilden die Grundlage für die Auswahl der optimalen Methode. Abhängig von den spezifischen Anforderungen und Funktionen können unterschiedliche Metriken geeignet sein und sollten in gewichteter Kombination betrachtet werden [12, 13].

Als Maß für die Speichereinsparung gilt die Kompressionsrate f_K, die als Verhältnis zwischen originalem und reduziertem Datenumfang definiert ist [14] (vgl. auch Formel 1).

Mit Ausnahme der verlustfreien Kompressionsmethoden implizieren alle Methoden eine Verfälschung der Daten. Es ist daher entscheidend zu evaluieren, inwiefern die essenziellen Informationen erhalten bleiben, wofür der spezifische Anwendungsfall für die Definition der relevanten Daten maßgeblich ist. Generelle Qualitätsbewertungen können durch Ähnlichkeitsmaße durchgeführt werden, die die Abweichung zwischen den originalen und den rekonstruierten Daten quantifizieren. Hierbei kann grundlegend zwischen formbasierten und strukturbasierten Ähnlichkeitsmaßen unterschieden werden [15]. Formbasierte Maße konzentrieren sich auf den Vergleich der individuellen Datenpunkte, während strukturbasierte Maße die Bewahrung der übergeordneten Strukturen, Verläufe und Trends evaluieren.

Um die performantesten Datenreduktionsmethoden zu identifizieren, ist eine Verknüpfung der Kompressionsrate (f_K) und Qualitätsmetriken (f_Q) erforderlich. Hierzu wurde die Rate-Distortion-Optimization aus [13] verwendet, welche die Kennzahlen mittels eines Parameters λ gewichtet und zu einem Index F_{RD} verbindet:

$$F_{RD} = f_K + \lambda \cdot (f_Q - f_K) \quad (2)$$

Experimentelle Untersuchung der Datenreduktionsmethoden

Um die Funktionsweise des Lean-Data-Ansatzes zu demonstrieren, wurden verschiedene Reduktionsmethoden auf einen exemplarischen Datensatz angewandt. Abhängig vom jeweiligen Anwendungsfall wurde das Resultat mittels einer individuellen Kombination von Metriken bewertet, um eine Gruppe geeigneter Methoden zu ermitteln.

Im Experiment diente ein Drehmomentsensor als Quelle der Beispieldaten. Für die Analyse wurde eine Methodendatenbank mit insgesamt elf Verfahren erstellt, von denen sieben Methoden jeweils einen oder zwei eigene Parameter zur Anpassung der Verarbeitungseigenschaften aufwiesen.

Um die Ergebnisse auszuwerten, wurden drei exemplarische Anwendungsfälle mit unterschiedlichen Anforderungen an die verbleibende Datenqualität definiert. Diese werden in Tabelle 1 dargelegt. Es wurde angenommen, dass keine Information über einen späteren Anwendungszweck vorliegt und keine Eingrenzung relevanter Daten getroffen werden kann. Die Daten sollen nur komprimiert wer-

Formel 1

$$f_K = \frac{Speicherbedarf(komprimierte\ Daten)}{Speicherbedarf(originale\ Daten)} \quad (1)$$

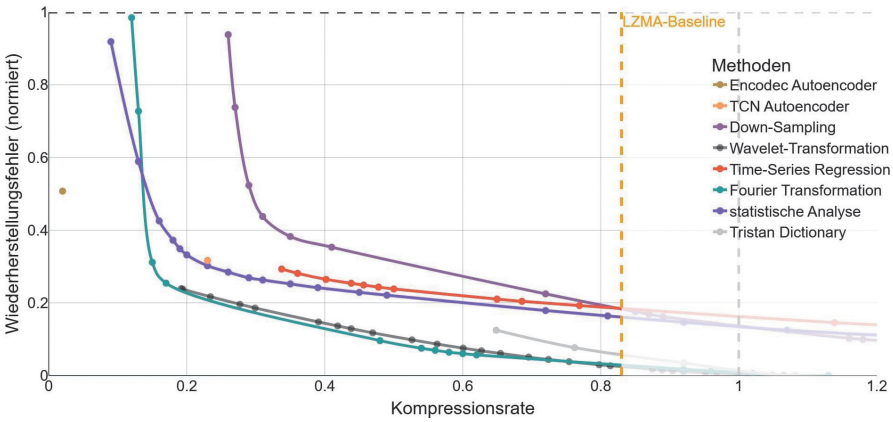

Bild 3. Ergebnisse der Suche nach der geeignetsten Datenreduktionsmethode im gesamten Parameterraum für den Anwendungsfall der datengetriebenen Zustandsüberwachung

den. Bei den hier ausgewählten Anwendungsfällen handelt es sich um häufig auftretende Szenarien aus dem Produktionsumfeld. Die Lean Data Toolbox kann jedoch in jedem beliebigen Anwendungsfall eingesetzt werden, solange die Anforderungen an die Kompressionsrate bzw. die Wiederherstellungsgenauigkeit spezifiziert werden können.

Für alle Anwendungsfälle war zunächst die Reduktionseigenschaft einer Methode von Interesse, welche im Experiment durch die Kompressionsrate f_K bestimmt wurde. Die Qualitätsmetrik f_Q hingegen wurde für jeden Anwendungsfall individuell gewählt.

Für den Anwendungsfall der reinen Datenkompression mit dem Hauptziel der Verringerung des Speicherplatzes und Wiederherstellbarkeit möglichst aller Informationen wurde der Fokus auf eine originalgetreue Wiedergabe der ursprünglichen Daten gerichtet. Dies wurde durch die Bewertung anhand einer formbasierten Ähnlichkeitsmetrik [16] in Form des Root Mean Squared Error (RMSE) umgesetzt.

Für die Analysemethoden, die im Anwendungsfall der datengetriebenen Zustandsüberwachung eingesetzt werden, können weniger aussagekräftige Elemente entfernt werden, wobei ein Datenverlust akzeptiert wird. Ein typisches Beispiel hierfür ist die Überwachung des Verschleißzustands von Maschinenkomponenten, bei der Zeitreihen wie Stromaufnahme oder Drehmoment analysiert werden, um frühzeitig Trends oder Anomalien zu erkennen. Hier zielte die Bewertung des Informationsgehalts darauf ab, wie gut die für die Zustandsüberwachung relevanten Daten erhalten bleiben. Forschungsarbeiten im Bereich der Predictive Maintenance [17–19] geben an, dass Modelle, die in diesem Kontext eingesetzt werden, oftmals den Verlauf des Trends innerhalb der Daten beobachten. Daher wurden die in der Aufzeichnung enthaltenen Trendinformationen priorisiert und eine originalgetreue Wiedergabe einzelner Datenpunkte vernachlässigt. Zur Bewertung wurde eine strukturbasierte Ähnlichkeitsmessung in Form der Korrelationsmetrik herangezogen.

Die datengetriebene Qualitätsbewertung adressiert die Produktqualität [20] und sollte im Experiment einen Anwendungsfall des Machine-as-a-Service abbilden. Hierbei wurden statistische Daten analysiert, um Lastbereiche zu untersuchen und vertrauliche Informationen im Produktionsablauf zu verschleiern [21, 22]. Die Qualität wurde durch den Vergleich der statistischen Verteilung eines Datensegments vor und nach der Verarbeitung definiert.

Unabhängig vom Anwendungsfall wurde die jeweilige Qualitätsmetrik mittels der Rate Distortion Optimization Gleichung 2) mit der Kompressionsrate verknüpft und beide durch den Parameter λ gewichtet. Werte von $\lambda \to 0$ stellen eine reine Bewertung durch die Kompressionsrate dar und vernachlässigen den entstehenden Fehler, während Werte von $\lambda \to 1$ ausschließlich die Wiederherstellungsgenauigkeit bewerten und die Methode mit den geringsten Abweichungen die beste Bewertung erhält. Da mit zunehmender Abweichung zwischen Original- und wiederhergestellten Daten der Nutzen der Methode abnimmt, wurde standardmäßig ein konservativer Wert von $\lambda = 0{,}75$ gewählt.

Ergebnisse des Experiments

Im Rahmen der Durchführung wurde eine Grid Search verwendet, um neun Datenreduktionsmethoden mit ihren jeweiligen Parameterkonfigurationen auf den Datensatz anzuwenden. Die getesteten Methoden gehören zu den Kategorien der Instanzenreduktion (Downsampling, Wavelet-Transformation, Time-Series Regression, Fourier-Transformation, statistische Analyse; alle verlustbehaftet) und der Datenkompression (Encodec Autoencoder [23], TCN Autoencoder [24], Tristan Dictionary [25]; verlustbehaftet bzw. LZMA; verlustfrei). Die Resultate wurden tabellarisch erfasst und jede Konfiguration anhand der Metrikkombinationen der definierten Anwendungsfälle evaluiert. Dabei wurden jeweils die verlustbehaftete Kompressionsmethode mit der verlustfreien LZMA-Methode verglichen, die eine Kompressionsrate von 0,834 erreicht. Beispielhaft sind die Ergebnisse für den Anwendungsfall aus dem Bereich der datengetriebenen Zustandsüberwachung, bei der das Korrelationsmaß zwischen Original- und wiederhergestellten Daten zur Bewertung der angewandten Datenreduktionsmethoden herangezogen wird (Bild 3). Die vertikale gestrichelte Gerade stellt die LZMA-Baseline, also die höchste Kompressionsrate, die verlustfrei erreicht werden kann, dar. Insgesamt ist das Ranking der Methoden von der Gewichtung λ und den verwendeten Metriken abhängig.

Es lässt sich erkennen, dass die statistische Analyse, die Fourier-Transformation und das Downsampling bei einer niedrigen Kompressionsrate eine gute Wiederherstellungsgenauigkeit aufweisen. Bei höheren Kompressionsraten erreicht aber keine der verlustbehafteten Methoden mit der verlustfreien LZMA-Methode vergleichbare Werte. Die Fourier-Transformation oder das Downsampling stellen dennoch geeignete Kompressionsmethoden dar, da der spätere Verwendungszweck der Daten, in diesem Falle die datenge-

triebene Zustandsüberwachung, einen Informationsverlust zulässt.

Es ist zu beachten, dass andere Datensätze zu anderen Ergebnissen führen können, da einige Kompressionsmethoden beispielsweise periodische Daten besonders effizient komprimieren können.

Zusammenfassung und Ausblick

Insgesamt konnte gezeigt werden, dass durch gezielte Reduktionsmethoden wie Dimensions- und Instanzenreduktion sowie Kompression große Datenmengen reduziert werden können, ohne wichtige Informationen zu verlieren. Dabei ist jedoch die Angabe von spezifischen Informationen über die zu reduzierenden Daten erforderlich. Durch die Kenntnis des späteren Verwendungszwecks der Daten lässt sich die Methodenauswahl noch weiter präzisieren. Die vorgestellte Toolbox wählt auf Basis der eingegebenen Informationen methodisch die passende Reduktionsmethode für verschiedene Anwendungsfälle aus.

Zukünftige Entwicklungen sollten sich auf die Weiterentwicklung und Automatisierung der Auswahl von Datenreduktionsmethoden fokussieren. Insbesondere die Integration von Machine Learning zur Optimierung der Datenanalyseprozesse bietet großes Potenzial. Ebenso bleibt es ein wichtiges Forschungsfeld, die Übertragbarkeit der Ergebnisse auf andere Anwendungsfälle und Datentypen, wie z. B. Bild- oder Videodaten, zu evaluieren.

Darüber hinaus könnte die Untersuchung zusätzlicher Anwendungsbereiche, etwa die datengetriebene Optimierung in Echtzeitsteuerungen oder Predictive-Maintenance-Szenarien in unterschiedlichen Branchen, den Lean-Data-Ansatz weiter verbreitern. Schließlich stellt auch die Bewertung der ökologischen Auswirkungen durch reduzierte Speicher- und Rechenressourcen eine interessante Perspektive für zukünftige Arbeiten dar.

Mehrwert aus den Produktionsdaten

Die IIoT-Building-Blocks der *iT Engineering Software Innovations GmbH* sind eine Lösung für Maschinenhersteller, um aus ihren Produktionsdaten einen Mehrwert zu gewinnen und neue Services zu gestalten. Die verschiedenen Apps bieten eine modulare Gesamtlösung, mit der Daten gespeichert, analysiert und aus ihnen gelernt werden kann. Die Lean-Data-App hilft, nicht wertschöpfende Daten zu vermeiden und den Wert der gesammelten Daten zu steigern. Sie bietet Unterstützung bei der Auswahl und Anwendung geeigneter Reduktionsmethoden. Weitere Informationen finden Sie auf unserer Website iiotbuildingblocks.io.

Literatur

1. Wischmann, S.; Wangler, L.; Botthof, A.: Industrie 4.0: Volks- und betriebswirtschaftliche Faktoren für den Standort Deutschland. 2015 (https://vdivde-it.de/system/files/pdfs/industrie-4.0-volks-und-betriebswirtschaftliche-faktoren-fuer-den-standort-deutschland.pdf [Abgerufen am 24.1.2025])
2. Vogel-Heuser, B.; Bauernhansl, T.; Hompel, M. ten (Hrsg.): Handbuch Industrie 4.0. Bd.4: Allgemeine Grundlagen. Springer, Berlin, Heidelberg 2017 DOI:10.1007/978-3-662-53254-6
3. Bauernhansl, T.; Hompel, M. ten; Vogel-Heuser, B. (Hrsg.): Industrie 4.0 in Produktion, Automatisierung und Logistik: Anwendung, Technologien, Migration. Springer Vieweg, Wiesbaden 2014 DOI:10.1007/978-3-658-04682-8
4. Corbett, C.; Chen, J.: Big Data Efficiency, Information Waste and Lean Big Data Management: Lessons from the Smart Grid Implementation. In: CONF-IRM 2015 Proceedings. (https://aisel.aisnet.org/confirm2015/8 [Abgerufen am 24.1.2025])
5. Zhang, R.; Gu, F.; Mansaf, H. et al.: Gear Wear Monitoring by Modulation Signal Bispectrum Based on Motor Current Signal Analysis. Mechanical Systems and Signal Processing 94 (2017) 9, S. 202–213 DOI:10.1016/j.ymssp.2017.02.037
6. Mai, J.; Chuah, C.-N.; Sridharan, A. et al.: Is Sampled Data Sufficient for Anomaly Detection? In: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, ACM 2006, S. 165–176 DOI:10.1145/1177080.1177102
7. Han, J.; Kamber, M.; Pei, J.: Data Mining: Concepts and Techniques. Elsevier/Morgan Kaufmann, Boston 2012 DOI:10.1016/C2009-0-61819-5
8. García, S.; Luengo, J.; Herrera, F.: Data Preprocessing in Data Mining. Intelligent Systems Reference Library 72 (2015), Springer International Publishing, Cham 2015 DOI:10.1007/978-3-319-10247-4
9. Chen, L.; Özsu, M. T.; Oria, V.: Using Multi-Scale Histograms to Answer Pattern Existence and Shape Match Queries. In: International Conference on Statistical and Scientific Database Management, 2005 (https://api.semanticscholar.org/CorpusID:12693997 [Abgerufen am 24.1.2025])
10. Sayood, K.: Introduction to Data Compression. Elsevier, Boston 2006 DOI:10.1016/C2015-0-06248-7
11. Mohan, B. S. S.; Govindan, V. K.: IDBE – An Intelligent Dictionary Based Encoding Algorithm for Text Data Compression for High Speed Data Transmission Over Internet. 2006 DOI:10.48550/ARXIV.CS/0601077
12. Maillo, J.; Triguero, I.; Herrera, F.: Redundancy and Complexity Metrics for Big Data Classification: Towards Smart Data. IEEE Access 8 (2020), S. 87918–87928 DOI:10.1109/ACCESS.2020.2991800
13. Sullivan, G. J.; Wiegand, T.: Rate-distortion Optimization for Video Compression. IEEE Signal Processing Magazine 15 (1998) 6, S. 74–90 DOI:10.1109/79.733497
14. Chiarot, G.; Silvestri, C.: Time Series Compression Surveys. ACM Computing Survey 55 (2023) 10, S. 1–32 DOI:10.1145/3560814.
15. Lin, J.; Li, Y.: Finding Structural Similarity in Time Series Data Using Bag-of-Patterns Representation. In: Winslett, M. (Hrsg.): Scientific and Statistical Database Management. Lecture Notes in Computer Science, Bd. 5566, Springer, Berlin, Heidelberg 2009, S. 461–477 DOI:10.1007/978-3-642-02279-1_33
16. Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (Hrsg.): Multivariate Analysemethoden: eine anwendungsorientierte Einführung; mit 6 Tabellen. Springer, Berlin 2006 DOI:10.1007/978-3-642-14987-0
17. Zonta, T.; Da Costa, C. A.; Da Rosa Righi, R.; De Lima, M. J. et al.: Predictive Maintenance in the Industry 4.0: A Systematic Literature Review. Computers & Industrial Engineering 150 (2020) 17 DOI:10.1016/j.cie.2020.106889
18. Carvalho, T. P.; Soares, F. A. A. M. N.; Vita, R. et al.: A Systematic Literature Review of Machine Learning Methods Applied to Predictive Maintenance. Computers & Industrial Engineering 137 (2019) DOI:10.1016/j.cie.2019.106024
19. Pech, M.; Vrchota, J.; Bednář, J.: Predictive Maintenance and Intelligent Sensors in

Smart Factory: Review. Sensors 21 (2021) 4 DOI:10.3390/s21041470
20. Wang, J.; Xu, C.; Zhang, J.; Zhong, R.: Big Data Analytics for Intelligent Manufacturing Systems: A Review. Journal of Manufacturing Systems 62 (2022) 2, S. 738–752 DOI:10.1016/j.jmsy.2021.03.005
21. Hsu, J.-Y.; Wang, Y.-F. ; Lin, K.-C. et al.: Wind Turbine Fault Diagnosis and Predictive Maintenance Through Statistical Process Control and Machine Learning. IEEE Access 8 (2020), S. 23427–23439 DOI:10.1109/ACCESS.2020.2968615
22. Kamat, P.; Sugandhi, R.: Anomaly Detection for Predictive Maintenance in Industry 4.0-A Survey. E3S Web of Conferences 170, 2020 DOI:10.1051/e3sconf/202017002007
23. Défossez, A.; Copet, J.; Synnaeve, G.; Adi, Y.: High Fidelity Neural Audio Compression. 2022 (https://arxiv.org/pdf/2210.13438 [abgerufen am 24.1.2025])
24. Bai, S.; Kolter, J. Z.; Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. 2018 (https://arxiv.org/pdf/1803.01271 [Abgerufen am 24.1.2025])
25. Marascu, A.; Pompey, P.; Bouillet, E. et al.: TRISTAN: Real-time Analytics on Massive Time Series Using Sparse Dictionary Compression. In: Proceedings of the 2014 IEEE International Conference on Big Data, IEEE 2014, S. 291–300 DOI:10.1109/BigData.2014.7004244

Die Autor:innen dieses Beitrags

Marina Baucks, M. Sc., studierte Maschinenbau am Karlsruher Institut für Technologie (KIT). Seit 2022 ist sie wissenschaftliche Mitarbeiterin am wbk Institut für Produktionstechnik des KIT.

Marcus Mau, M. Sc., studierte Elektro- und Informationstechnik am Karlsruher Institut für Technologie (KIT). Seit 2024 ist er wissenschaftlicher Mitarbeiter am wbk Institut für Produktionstechnik des KIT.

Peter Ruppelt, M. Sc., studierte Maschinenbau am Karlsruher Institut für Technologie (KIT).

Alexander Puchta, M. Sc., studierte Maschinenbau am Karlsruher Institut für Technologie (KIT). Seit 2020 ist er wissenschaftlicher Mitarbeiter am wbk Institut für Produktionstechnik des KIT und leitet seit 2023 als Oberingenieur die Forschungsgruppe Intelligente Maschinen und Komponenten.

Prof. Dr.-Ing. Jürgen Fleischer studierte Maschinenbau an der Universität Karlsruhe (TH) und promovierte 1989 am Institut für Werkzeugmaschinen und Betriebstechnik (wbk). Von 1992 an war er in mehreren leitenden Positionen in der Industrie tätig, ehe er im Jahr 2003 zum Professor und Leiter des wbk Institut für Produktionstechnik am heutigen Karlsruher Institut für Technologie (KIT) berufen wurde. Dort leitet er den Bereich Maschinen, Anlagen und Prozessautomatisierung.

Abstract

Lean Data – Application-specific Reduction of Large Amounts of Data in the Production Environment. The article examines 'lean data', which aims to reduce large amounts of data in production environments efficiently. To minimize data complexity without losing relevant information, a toolbox for the data-driven selection of specific reduction methods is used. The approach combines data compression and reduction methods and considers application-specific requirements, such as temporal context and recovery accuracy. Experiments show that adapted reduction methods offer significant storage and analysis advantages.

Danksagung

Das Projekt „Lean Data – Effiziente Datenaufnahme und -speicherung für eine nachhaltige Produktion“ wurde durch das Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg im Rahmen des Programmes Invest BW vom 15.10.2021 (VwV Invest BW – Innovation II, Förderkennzeichen: BW1_1269/02) gefördert und vom wbk Institut für Produktionstechnik des Karlsruher Institut für Technologie sowie der iT Engineering Software Innovation GmbH in Kooperation bearbeitet

Schlüsselwörter

Lean Data, Big Data, Data Reduction, Maschinendaten, Industrial Internet of Things (IIoT)

Keywords

Lean Data, Big Data, Data Reduction, Machine Data, Industrial Internet of Things (IIoT)

Bibliography

DOI:10.1515/zwf-2024-0148
ZWF 120 (2025) Special Issue; page 152 – 158

ISSN 0947-0085 · e-ISSN 2511-0896