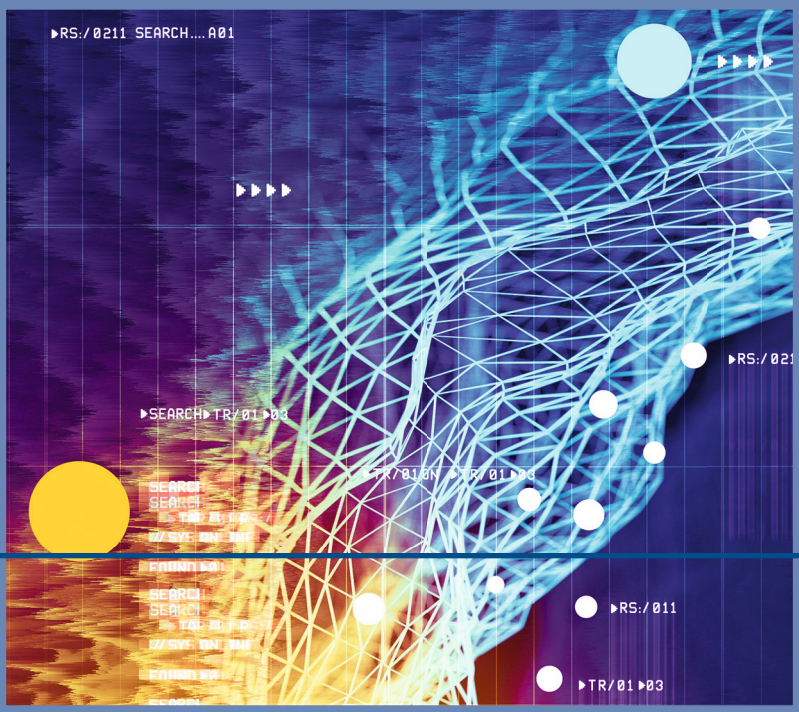
Ralf Otte
Boris Wippermann
Sebastian Schade
Viktor Otte

inkl. Ausblick
auf Small Data
und Mind Data

Von Data Mining bis Big Data

Handbuch für die industrielle Praxis



HANSER

Otte/Wippermann/Schade/Otte

Von Data Mining bis Big Data

Ralf Otte
Boris Wippermann
Sebastian Schade
Viktor Otte

Von Data Mining bis Big Data

Handbuch für die industrielle Praxis
inklusive Small Data und Mind Data

Mit 204 Bildern und 52 Tabellen

HANSER

Die Autoren:
Prof. Dr.-Ing. Ralf Otte ist Hochschullehrer für Prozessautomatisierung und Künstliche Intelligenz an der Technischen Hochschule Ulm.
Boris Wippermann ist Principal bei der h&z Unternehmensberatung AG in München.
Sebastian Schade, ist Lead Consultant bei der INFOMOTION GmbH Frankfurt.
Prof. Dr.-Ing. habil. Viktor Otte ist emeritierter Hochschullehrer der Universität Wuppertal im Fachgebiet Maschinenbau und langjähriger Berater für Data-Mining-Projekte.

Bibliografische Information der Deutschen Nationalbibliothek:

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Lektorat: Dipl.-Ing. Volker Herzberg
Herstellung: Björn Gallinge, Carl Hanser Verlag/le-tex publishing services GmbH, Leipzig
Coverkonzept: Marc Müller-Bremer, www.rebranding.de, München
Coverrealisation: Max Kostopoulos
Satz: Kösel Media GmbH, Krugzell
Druck und Bindung: UAB BALTO print, Vilnius (Litauen)
Printed in Lithuania

Print-ISBN: 978-3-446-45550-4
E-Book-ISBN: 978-3-446-45717-1

Vorwort

Das vorliegende Buch entstand im Ergebnis einer über 25-jährigen Beschäftigung der Autoren mit Data-Mining-Projekten im industriellen Umfeld und einer mehrjährigen Tätigkeit im Big-Data-Business im kommerziellen Bereich.

Als 2004, also vor über 15 Jahren, das Buch „Data Mining für die industrielle Praxis“ im HANSER Verlag von zwei der heutigen Autoren erschien [OTT04], reagierte es auf einen zunehmenden Trend in Wirtschaft, Verwaltung, Forschung und Industrie, die Informationen vorliegender Prozessdaten für die Prozessoptimierung besser zu nutzen. „Graben in Daten“ war damals bereits angesagt. Die Effizienz der vorhandenen Rechnersysteme war gestiegen, neue mathematische Verfahren zur statistischen Datenanalyse waren hinreichend erprobt und der Bedarf der potenziellen Anwender dieser neuen Technologie war groß. Schon bis zum Erscheinen des damaligen Buches, aber insbesondere in der darauffolgenden Zeit, entstanden vielfältige Data-Mining-Projekte, die nahezu alle Facetten dieser relativ neuartigen Technologie tangierten. Es musste auch viel gelernt werden. Bei der Projektrealisierung wurden viele handwerkliche Fehler gemacht, insbesondere die Phase der Datenbereitstellung umfangs- und zeitmäßig unzureichend geplant, die Kosten unterschätzt und die Projekteinführung in den industriellen Prozess zu optimistisch prognostiziert. Jeder, der ein derartiges Projekt real begleitet oder sogar selbst verantwortlich geführt hat, weiß davon ein Lied zu singen. Letztlich hat das dem Siegeszug von Data Mining oder Data Science aber nicht geschadet.

Die Aufgaben zur Optimierung vorhandener Prozesse sind heute noch genauso aktuell wie vor 15 Jahren und deshalb haben sich die Autoren des damaligen Buches in Abstimmung mit dem Verlag entschlossen, trotz neuer Entwicklungen und Trends in der Datenverarbeitung wie Big Data oder Small Data dem klassischen Data Mining in diesem Buch erneut hinreichend Platz einzuräumen. Will man Data Mining anwenden, so muss man die dazugehörigen Verfahren bzw. Methoden beherrschen oder zumindest theoretisch verstehen. Obwohl auch dieses Buch kein Lehrbuch sein will, so erschien es uns zweckmäßig, die im Buch von 2004 beschriebenen mathematischen Grundlagen für das vorliegende Buch komplett zu übernehmen, ergänzt durch einige zweckmäßig ausgewählte Beispiele. Der Leser

unseres Vorgängerbuches wird also in den Kapiteln 2 bis 3 nichts wesentlich Neues finden. Es lohnt sich aber auf jeden Fall, sich mit den verschiedenen, problemangepassten Möglichkeiten der Datenselektion und -zusammenführung, der Datenvorverarbeitung, der Datentransformation und der statistischen Datenanalyse wie z. B. multivariaten Verfahren, Künstlichen Neuronalen Netzen (KNN), Support-Vektor-Maschinen, Selbstorganisierenden Merkmalskarten (SOM) oder Clusterverfahren und den daraus ziehbaren Schlussfolgerungen, wie sie z. B. durch verschiedene Regelgenerierungsverfahren oder Visualisierungen hochdimensionaler Datenräume möglich werden, nochmals zu beschäftigen. Zur Klassifizierung von Daten im Sinne von Lernprozessen hat sich die Informatik in der letzten Zeit sehr stark auf Künstliche Neuronale Netze konzentriert, vor allem angeregt durch die neuen Möglichkeiten des Deep Learning mit tiefen Neuronalen Netzen. Dazu wird jedoch Spezialliteratur empfohlen, die Grundlagen der neuronalen Lernprozesse muss man aber verstanden haben. Auch deshalb ist es zweckmäßig, sich nochmals mit der Theorie zu beschäftigen. In Kapitel 4 werden einige Auswertmöglichkeiten für praktische Anwendungsfälle dargestellt, die heute wie vor 15 Jahren noch hochaktuell sind.

Viele der hier beschriebenen mathematischen Verfahren findet die Leserin bzw. der Leser auch im KI-Buch „Künstliche Intelligenz für Dummies“ aus dem Jahre 2019 von einem unserer Autoren [OTT19]. Da die Künstliche Intelligenz heute zu einem Großteil auf maschinelles Lernen setzt, sind viele Verfahren und auch Anwendungsbeispiele in beiden Büchern nahezu deckungsgleich. Manche Algorithmen, wie die zu Unrecht fast nirgends erwähnten Selbstorganisierenden Merkmalskarten (SOM), sind in dem hier vorliegenden Buch allerdings detaillierter erläutert, andere, eher übliche KI-Verfahren, wie Entscheidungsbäume, Backpropagation-Netze oder Deep-Learning-Faltungsnetze (CNN) sind in dem KI-Buch näher beschrieben. Der Schwerpunkt des hier vorliegenden Buches liegt eindeutig in der Datenanalyse und nicht so sehr in der Beschreibung der Künstlichen Intelligenz selbst.

In den letzten Jahren hat sich ein gravierender Wandel in den Datenverarbeitungstechnologien ergeben. Während klassische Data-Mining-Projekte immer mit einem gewissen Zeitpolster bearbeitet werden konnten und weiterhin auch können, gibt es zunehmend Aufgaben in allen Bereichen moderner Gesellschaften, bei denen riesige, temporär anfallende Daten (Volume) unterschiedlichster Formate (Variety) in extrem kurzer Zeit (Velocity) verarbeitet werden müssen. Die drei „Vs“ sind zum Charakteristikum einer neuen Form der Datenverarbeitung geworden, die als Big Data bezeichnet wird. Alle Datenverarbeitung bleibt aber nur sinnvoll, wenn den ermittelten Informationen ein inhaltlicher Wert (Value), ein Nutzen, zugeordnet werden kann. Betrachtet man die täglich anfallenden riesigen Datenmengen in sogenannten Sozialen Netzen oder, um ein diametrales Beispiel zu nennen, im Gesundheitswesen, so wird schnell klar, dass die abgeleiteten Informationen bzgl.

ihre Korrektheit oder Zuverlässigkeit sehr unterschiedlich bewertet werden müssen. Hinweise zum Kaufverhalten des Nutzers eines Online-Shops haben durchaus einen anderen Qualitätsanspruch als eine medizinische Diagnose, die aus Vergleichsdaten ähnlicher Fälle computergestützt erstellt wird. Ansprüche an Prädiktion und Wertvorstellungen sind also kontextabhängig. Wir wollen sie hier nicht betrachten, sondern uns stattdessen auf die technologischen Herausforderungen der drei Vs konzentrieren.

Kapitel 5 beschreibt zunächst Architekturkonzepte zum Umgang mit vielen Daten, geht dann auf die Verarbeitungsmöglichkeiten in Form eines deskriptiven oder präskriptiven Datenmodells ein, benennt Möglichkeiten der Hardware- und Datenvirtualisierung und den weiteren Trend zur Entkoppelung von Speicherung und Verarbeitung von Daten. Nach diesen grundsätzlichen Bemerkungen werden in Kapitel 6 die Komponenten der Big-Data-Pionierlösung Hadoop und ihre Weiterentwicklungen dargestellt, auf Apache Spark, eine weitere Evolutionsstufe, eingegangen (die sich gegenwärtig als Standard herausbildet) und anschließend die vielfältigen Modifikationen und Neuentwicklungen beschrieben, die verschiedenen Nutzern den Zugang zu Big-Data-Diensten erleichtern sollen. Dieses Kapitel informiert auch über solche Schlüsseltechnologien wie Apache Kafka oder Kappa, die insbesondere für eine Streamingverarbeitung genutzt werden. Von besonderem Interesse könnte für Leser auch der zunehmende Einsatz von NoSQL-Datenbanken sein, die das übliche relationale Speicherkonzept auch zugunsten von massiv verteilten WEB-Anwendungen aufgegeben haben. Es werden verschiedene Einsatzfälle beschrieben. Hier sind hunderte Open-Source-Anbieter auf dem Markt. Abschließend informiert das Kapitel über die Möglichkeit, für wiederkehrende Aufgaben klar definierte Lösungsmuster, sogenannte Stacks, einzusetzen. Damit sind die Voraussetzungen genannt, um die in den folgenden Abschnitten spezialisierten Ausführungen zu Datenarchitekturen, zum Aufbau von Data Lakes und zum Cloud Computing einordnen zu können. Die Frage, wie Big Data und Künstliche Intelligenz (KI) zusammenwirken, wird abschließend diskutiert. Big Data wird zu einer Quelle für KI und für den Data Scientist, der letztlich für den Mehrwert der Datenanalyse zuständig ist.

Kapitel 7 und 8 schildern Anwendungen. Die gesamten Ausführungen dieser Kapitel sind sehr kompakt gehalten und man bekommt viele Detailinformationen, die auch als Kompass zum Navigieren in diesem äußerst vielseitigen und sich gegenwärtig noch ausdehnenden Terrain dienen. Die Autoren wissen natürlich, dass die dargestellten Informationen nur eine Übersicht sein können und kein Lehrbuch zum Studium einzelner Facetten dieses riesigen Gebietes der Datenverarbeitung. Die angegebenen Quellen, im Schwerpunkt Internetadressen, helfen aber weiter, wenn man die Absicht hat, sich tiefer in die Problematik von Big Data einzuarbeiten. In vielen, in den vergangenen Jahren realisierten Beispielen hat sich gezeigt, dass die ausreichende Vorbereitung und Planung von Data Mining oder Big-Data-

Projekten oft sträflich unterschätzt wird, oftmals sogar bereits die Aufgabenstellung nicht hinreichend an den unternehmerischen Belangen, sondern zu sehr aus einer IT-Perspektive heraus ausgerichtet ist. Hierbei sind also nicht nur die Informatiker und Kenner der inhaltlichen Probleme gefragt, sondern vor allem die Geschäfts- und Prozessverantwortlichen sowie Entscheider. Hier wird auch das Urteil darüber gefällt, ob das Data-Projekt erfolgreich durchgeführt werden wird, ob es beginnt, wie es realisiert oder ob es bereits in einer frühen Phase gestoppt wird. Wir haben uns deshalb entschlossen, diese Problematik sehr ausführlich darzustellen. Die dafür zuständigen Autoren kennen die zu behandelnden Aufgaben genau, da sie sich in ihrem beruflichen Alltag mit allen diesen Fragen der Projektplanung, Projektrealisierung und Projekteinführung täglich beschäftigen müssen. Die gegebenen methodischen Hinweise sind deshalb äußerst praxisnah und oft auch wie ein Kochrezept abzuarbeiten. Dem Leser, der ein Data-Projekt von der Planung bis zur Einführung bearbeiten muss, sind die in diesen Kapiteln gegebenen Hinweise deshalb nachdrücklich zu empfehlen, insbesondere auch deshalb, weil hier verschiedene betriebliche Anwendungsfelder behandelt werden. Einige der Beispiele zu Anwendungsfällen von Data Mining sind wieder dem älteren Buch „Data Mining für die industrielle Praxis“ entnommen, da sie ihre didaktische Funktion, die in Kapitel 2 und 3 dargestellte Theorie am praktischen Beispiel zu erläutern, noch immer gut erfüllen können. Weitere, kleine Ausführungsbeispiele findet man in dem bereits genannten Fachbuch „Künstliche Intelligenz für Dummies“ von einem unserer Autoren.

Verfolgt man die Entwicklung zu Datenverarbeitungstechnologien, so taucht seit einiger Zeit ein neuer Begriff auf, das sogenannte Small Data. Hier wird das Paradigma geändert: Nicht noch immer mehr und immer schneller, sondern zurück zu den Anfängen. Small Data erklärt sich aus der Absicht, nur wenige Daten zu nutzen oder auch nur zur Verfügung zu haben, aber auch aus ihnen wesentliche Informationen zur Lösung eines Problems ableiten zu müssen. In der Philosophie des Ansatzes versteckt sich auch die Sorge, dass extrem viele Daten möglicherweise den Problemkern verschleiern könnten, man also besser erst einmal versucht, die Dinge zu vereinfachen und mit dem zu arbeiten, was ohnehin zur Verfügung steht. Kapitel 9 widmet sich diesem Thema, wobei schnell klar wird, dass es sich eigentlich nicht um eine neue Technologie, sondern tatsächlich nur um einen Ansatz handelt, der etwas in Vergessenheit geraten ist. Neue Erkenntnisse über die Welt gewinnt man durch Induktion und Deduktion, also durch Extraktion versteckter Inhalte aus riesigen Datenbergen mittels statistischer Verfahren (induktives Lernen), aber auch durch logische (oder kreative) Schlussfolgerung aus oft nur wenigen (oder gar keinen) Daten (deduktives Schließen). Der Mensch kann beides, die modernen Datenverarbeitungstechnologien haben den Akzent seit Data Mining und Big Data aber mehr auf die Statistikvariante gelegt. Es wird also Zeit, sich wieder der Deduktion zu erinnern und zu hinterfragen, wie eigentlich der Mensch mit

seinem Gehirn Small Data und letztlich Deduktion praktiziert. Wie also wird die Zukunft eines technisch optimierten Small Data aussehen?

Hier nun haben die Verfasser dieses Kapitels den bisher beschrittenen Weg der Analyse und Zusammenfassung vorliegender, bewährter Fachinhalte verlassen und – mit dem Risiko „zweifelnder Akzeptanz" beim Leser – einen vollkommen neuen, hypothetischen Ansatz gewählt. Sie bezeichnen die im menschlichen Gehirn ablaufenden Datenverarbeitungsprozesse als Mind Data und führen aus, dass sich Mind Data nur unter Nutzung von Bewusstsein realisieren lässt. Damit entsteht die grundsätzliche Frage, wie maschinelles, technisches Bewusstsein erzeugt werden kann. Es wird versucht, Antworten darauf aus einem Konzept abzuleiten, welches Bewusstsein als imaginären physikalischen Prozess beschreibt. Dieses Problem tangiert die aktuellen Forschungsarbeiten zur Künstlichen Intelligenz und es wird sofort klar, dass hier auch die neuesten Ansätze der KI hineinspielen, Künstliche Neuronale Netze mit neuromorpher Struktur oder Quantentechnologien aufzubauen.

Damit schließt sich der Kreis dieses Buches, angefangen mit Erläuterungen zu statistischen Verfahren zur Informationsgewinnung im Rahmen von Data Mining, über den Einsatz modernster Datenverarbeitungstechnologien zur Verarbeitung extrem großer Datenmengen in extrem kurzer Zeit durch Big Data bis zur Wiedergeburt der Erkenntnis des Small Data, dass auch in geringen Datenbeständen viel Information enthalten sein kann, die sich mit stark reduzierten Data-Technologien und auch alten, konventionellen Methoden gewinnen lässt. Daraus folgt zwangsläufig der Ansatz eines Mind Data, also einer Datenverarbeitung, die von den zukünftigen Entwicklungen der KI zum Aufbau eines rudimentären, technischen Bewusstseins geprägt sein wird.

Die Autoren hoffen, Ihnen, liebe Leser, damit einen verständlichen und interessanten Überblick über Technologien gegeben zu haben, die im „Zeitalter der Digitalisierung" schon heute und künftig noch nachhaltiger unser Leben bestimmen werden.

Das vorliegende Buch ist das gemeinschaftliche Werk von vier Autoren. Selbstverständlich gibt es bei den Fachkapiteln dedizierte Verantwortlichkeiten. Insbesondere verantwortlich für die Fachkapitel 2, 3, 4 und 9 und für zahlreiche Industriebeispiele in diesem Buch, unter anderem in den Kapiteln 7 und 8, sind Ralf Otte und Viktor Otte. Das Kapitel 4.1 basiert auf Ausarbeitungen von Christian Rohrdantz aus Konstanz, wofür wir uns an dieser Stelle nochmals ganz herzlich bedanken wollen. Unser Autor Sebastian Schade ist als Big-Data-Experte bei den Kapiteln 5 und 6 und bei zahlreichen anderen Praxisbeispielen in verschiedenen Kapiteln federführend gewesen, und Boris Wippermann, als Experte in der Verknüpfung von Digitalisierung und Unternehmensoptimierung, zeichnet hauptverantwortlich für die Anwendungskapitel 7 und 8.

In der Drucklegung des Buches werden leider keine Farben wiedergegeben. Die Autoren weisen extra darauf hin, da bei verschiedenen Bildern die dargestellten Grauwerte nicht immer als unterschiedlich identifiziert werden können.

Wenn Sie Fragen, Anregungen oder auch Verbesserungsvorschläge an einen unserer Autoren haben, freuen wir uns über eine Mail an den Hauptverantwortlichen des Buches unter *ralf.otte@email.de*.

Am Ende des Vorwortes möchten wir allen danken, die zur Erstellung des Buches beigetragen haben, aber ganz besonders auch unseren Familienangehörigen, die durch Verzicht, aber auch Aufmunterung am Gelingen des Projektes beteiligt waren.

Die Autoren

Weinheim, München, Frankfurt/Main und Magdeburg im Frühjahr 2020

Inhalt

1 Einführung

Seit vielen Jahren lassen sich im Investitionsverhalten von Industrienationen veränderte Trends erkennen. Während zahlreiche Neuinvestitionen in aufwärtsstrebenden asiatischen Ländern durchgeführt werden, sind Neuinvestitionen vor allem von Großprojekten in den USA, Japan, Europa und insbesondere auch in Deutschland erheblich zurückgegangen. Dieser Trend erscheint zumindest für die nächsten Jahre unumkehrbar zu sein, da aus wirtschaftlichen, aber auch aus sozialpolitischen Überlegungen die Produktionsanlagen möglichst nahe an den Verbrauchermärkten installiert werden und somit gerade in etablierten Industrienationen Neuinvestitionen nur noch in geringem Ausmaß notwendig sind. Deshalb werden hier in den nächsten Jahren der Schwerpunkt und die Herausforderungen darin bestehen, die vorhandenen industriellen Anlagen noch effizienter zu nutzen.

Derart notwendige Effizienzsteigerungen lassen sich durch verfahrenstechnische oder konstruktive Verbesserungen, durch bessere Materialien, bessere Rohstoffe, aber natürlich auch durch eine verbesserte „Fahrweise“ der Anlagen erreichen. Und gerade diese verbesserte Fahrweise ist unter Kosten-Nutzen-Aspekten ein vielversprechender Ansatz.

Auf der anderen Seite sind wir im Zeitalter der Daten angekommen. Daten sind zum Rohstoff geworden, der „gefördert“ und verarbeitet werden muss. Big Data ist in aller Munde, jeden Tag werden tausende Terabyte von Daten auf der Welt erfasst und gespeichert. Im kaufmännischen und technischen Umfeld, aber auch in sozialen Bereichen wird die Nutzung dieser erfassten Daten mit Data Mining und Big-Data-Technologien zum entscheidenden Wettbewerbsvorteil.

Ein Blick in heutige Produktionsbereiche zeigt die Lage: Pro Sekunde entstehen hier Tausende von Informationen, Sensoren messen Geschwindigkeit und Qualität der Fertigung, Steuerungen horten wichtiges Wissen über den Zustand von Maschinen und Anlagen. Ursprünglich haben Unternehmen diese Informationsquellen gezielt für die Automatisierung ihrer Fertigung eingerichtet. Durch die Vernetzung von Maschinen und das Internet der Dinge werden hier inzwischen Datenmengen generiert, die das Volumen jeder anderen Branche übersteigen. Durch intelligente Datenauswertung lassen sich Daten in Erkenntnisse überfüh-

ren, die Entscheidungen unterstützen oder sogar Grundlage für automatisierte Handlungen und Optimierungen sein können. Welches sind die größten Einflussfaktoren auf die Produktivität meiner Fabrik? Wie kann ich Maschinenausfälle vermeiden? Wie kann ich die Qualität meiner Produkte erhöhen? Wie kann ich Energie einsparen?

Die Beratungsfirma Frost & Sullivan leitet von einem Big-Data-Einsatz enorme Optimierungspotenziale ab: eine Steigerung der Produktionseffizienz von zehn Prozent, eine Reduktion der Betriebskosten um fast 20 Prozent und eine Reduktion der Instandhaltungskosten um 50 Prozent. Roland Berger nennt eine Gesamtsumme von 1,25 Billionen Euro bis 2025, die in Europa durch Digitalisierung in der Industrie zusätzlich „gehoben“ werden könnte.

Das Bewusstsein des Wertes der Daten wächst also zunehmend. In der IT-Branche ist dies eine Selbstverständlichkeit. Technologieunternehmen wie Google oder Facebook begründen ihr gesamtes Geschäftsmodell auf der intelligenten Datenauswertung und erzeugen jährliche Gewinne von mehreren Milliarden US-Dollar. Big-Data-Methoden sind hier schon lange etabliert, Rechenleistung und Speicher sind erschwinglich, die Fortschritte im maschinellen Lernen sind enorm und Frameworks zur Nutzung dieser Verfahren frei verfügbar. Doch der Einsatz von Data Mining und Big-Data-Methoden ist im Maschinen- und Anlagenbau, insbesondere bei kleinen und mittleren Unternehmen, nach wie vor verhalten. Laut einer McKinsey-Studie von 2016 verstehen nur 15 Prozent der Betriebe in der industriellen Fertigung, dass Daten als Teil der Wertschöpfung anzusehen sind. In 50 Prozent der Unternehmen bleiben Daten für die Entscheidungsfindung gänzlich ungenutzt.

Dieses Buch beschreibt Möglichkeiten, aufbauend auf vorhandenen Unternehmens- und Prozessdaten, die Ressourcen noch effizienter einzusetzen, die Störungen drastisch zu senken und Ursachen für Unregelmäßigkeiten im Prozess auf eine intelligente und ökonomische Art aufzufinden. Ähnliches gilt für Forschungs- und Entwicklungsprozesse, bis hin zur Konstruktion und experimentellen Erprobung von Objekten oder Verfahren. Nicht alle Unternehmen sind für diese Art der Datennutzung heute schon hinreichend vorbereitet, aber der Trend zur Datenarchivierung ist in keinem Industriezweig mehr aufzuhalten. Damit gewinnt die Auswertung von Daten immer mehr an Bedeutung. Daten stellen Wissen dar und das erzeugt, richtig genutzt, einen entscheidenden Wettbewerbsvorteil.

2 Warum Data Mining? Wozu Big Data?

Die Welt ist in ein neues, ein „digitales Zeitalter“ eingetreten. Jeden Tag werden unzählige neue Daten erzeugt, gespeichert, archiviert, zunehmend auch genutzt. Allerdings hinkt die Nutzung den jeweiligen technischen Möglichkeiten hinterher, da die richtige Auswertung von Daten zur Generierung von Wissen hochkomplex ist. Die Antwort auf dieses Problem ist seit über 20 Jahren das Data Mining, seit fast 10 Jahren sprechen wir sogar von Big Data. Große Verbreitung hat der Big-Data-Ansatz bereits im Marketing & Sales und im Bereich Social Media gefunden. Dafür gibt es im Wesentlichen zwei Gründe: Genau in diesen Bereichen fallen seit Jahren riesige Datenmengen in digitaler Form an, die computergestützt ausgewertet werden können. Aber auch der zweite Grund ist nicht uninteressant. Auswertefehler sind in diesen Anwendungsfeldern nicht so relevant wie in anderen, kritischeren Lebensbereichen. In diesem Buch wollen wir uns verstärkt die Industriebereiche anschauen, da diese aktuell noch sehr viel Nachholbedarf bei Data Mining und Big-Data-Anwendungen besitzen.

In jedem Industriezweig hat sich eine Vielzahl von Spezialisten herausgebildet, die verfahrenstechnisch, konstruktiv, heuristisch oder analytisch den technischen Prozess zu verbessern helfen. Was in der Industrie allerdings immer noch nicht Stand der Technik ist, ist die Anwendung der Data-Mining-Verfahren zur Prozessanalyse und -optimierung. Nun ist die Mathematik bzw. theoretische Informatik den Anwendungen in der Praxis schon seit jeher um viele Jahre voraus, dennoch ist der Zustand unbefriedigend, da durch einen konsequenten Einsatz von modernen Auswertetechniken Einsparungen im Millionenbereich, wenn nicht Milliardenbereich, allein in Deutschlands Industrie erreicht werden könnten.

Data Mining und Big Data treten dabei nicht in Konkurrenz zu klassischen Verfahren der Prozessoptimierung wie Benchmarking, Schwachstellenanalysen, neuen Logistikkonzepten usw., sondern stellen vielmehr eine interessante Ergänzung zur schnellen Effizienzsteigerung und Erfolgsoptimierung dar. Data Mining tritt aber in Konkurrenz zu Prozesssimulationen, dem Bau von analytischen Modellen, dem Erstellen von Reports und dem „Bauchgefühl“ der Anlagenfahrer. Mit Data Mining werden klassische Fragen eines jeden Prozessverantwortlichen aufgegriffen, die

beispielsweise wie folgt lauten könnte: „Wie müssen in diesem oder jenem Augenblick die 100 Einstellgrößen und Prozessparameter festgelegt werden, um mit den geringstmöglichen Kosten und in der geringstmöglichen Zeit die beste Qualität robust zu produzieren, d. h., wie produziert man ein Produkt polyoptimal?“

Diese leicht nachvollziehbare Frage lässt sich in der Praxis aber gar nicht so leicht beantworten, da oftmals nicht bekannt ist, was bei so vielen Parametern als das Optimum zu betrachten ist. Ist es zum Beispiel der geringste Ausschuss bei gleichzeitig maximalem Durchsatz der Anlage? Und wenn ja, dann bleibt als weitere Frage, wie die dazugehörigen Stellgrößen eingestellt werden müssen, um dieses vektorielle Optimum zu erreichen. Die Anforderungen an die moderne Prozessführung sind mittlerweile so komplex, dass der Mensch bei der Durchführung seiner Tätigkeit immer mehr an seine Grenzen stößt.

Mathematisch gesehen kommt weiterhin hinzu, dass es offensichtlich viele parallele und gleichberechtigte Optima eines Prozesses gibt. Die Arbeit in der Praxis zeigt beispielsweise, dass Schichtführer ihre Produktionslinien nach ihren persönlichen Erfahrungen einstellen, die nachfolgenden Schichtführer die gleichen Linien jedoch nach einer anderen Art fahren. In beiden Fällen wird aber oftmals die gleiche Qualität produziert. Das kann übergeordnete Stellen ratlos machen. Doch aus rein mathematischer Sicht kann ein solches Prozessverhalten durchaus normal sein, da ein Produktionsprozess in der Regel gar nicht eineindeutig, sondern mehrdeutig ist. Mehrdeutigkeit in dem hier verstandenen Sinne bedeutet, dass es mehrere Kombinationen von Einstellgrößen gibt, die das gleiche Prozessergebnis erzeugen und dass umgekehrt die gleichen Einstellparameter an einem anderen Tage zu völlig verschiedenem Prozessverhalten führen können. Beide Fälle treten auf und man erkennt, dass sie eher die Regel als die Ausnahme darstellen.

Hat man sich aber an den Gedanken gewöhnt, dass jeder Produktionsprozess mehrdeutig ist, erkennt man schnell, dass viele klassische Auswertverfahren, wie z. B. analytische Modellbildungen, scheitern müssen. Nicht, weil sie nicht korrekt wären, im Gegenteil. Sie scheitern an ihrer Korrektheit, weil sie mit der Mehrdeutigkeit der Praxis, d. h. mit der schlechten Datengüte, den ständigen Änderungen in den Anlagen, den Umwelteinflüssen und den subjektiven Einstellungen der Anlagenfahrer nicht umgehen können. Sie sind oftmals die falschen Antworten auf die richtigen Fragen.

Letztendlich kommt es nicht auf die genaue Modellbildung an. Am Ende sollte neben jeder Modellbildung eine konkrete Optimierung und damit eine konkrete Parametervorgabe für die „Anlagenfahrer“ entstehen, damit sie den Prozess nach Maßgabe ihrer Kunden kostengünstiger und besser als vorher fahren können. „Anlagenfahrer“ im erweiterten Sinne sind auch alle diejenigen Bearbeiter, die Prozesse steuern müssen, ganz gleich, ob sie im Einkauf, in der Entwicklung, in der Produktion oder im Vertrieb tätig sind.

Data-Mining-Verfahren versuchen, diese Lücke der komplexen und mehrdeutigen Prozessoptimierung zu schließen, denn mit ihnen werden gerade Problemstellungen der Mehrdeutigkeit, der unvollständigen Daten, der täglichen Änderungen und der enormen Prozesskomplexität gelöst. Data Mining ist damit mehr eine neuartige Herangehensweise an eine komplexe Aufgabenstellung als ein bestimmtes mathematisches Verfahren. Es ist letztlich das Synonym für eine Vielzahl von selbstlernenden, unscharfen Verfahren aus dem maschinellen Lernen und dem Soft Computing.

Was aber ist nun Big Data? Bei **[GAB19]** finden wir folgende Erklärung: *Mit „Big Data" werden große Mengen an Daten bezeichnet, die u. a. aus Bereichen wie Internet und Mobilfunk, Finanzindustrie, Energiewirtschaft, Gesundheitswesen und Verkehr und aus Quellen wie intelligenten Agenten, sozialen Medien, Kredit- und Kundenkarten, Smart-Metering-Systemen, Assistenzgeräten, Überwachungskameras sowie Flug- und Fahrzeugen stammen und die mit speziellen Lösungen gespeichert, verarbeitet und ausgewertet werden. Es geht u. a. um Rasterfahndung, (Inter-)Dependenzanalyse, Umfeld- und Trendforschung sowie System- und Produktionssteuerung. Wie im Data Mining ist Wissensentdeckung ein Anliegen. Das weltweite Datenvolumen ist derart angeschwollen, dass bis dato nicht gekannte Möglichkeiten eröffnet werden. Auch die Vernetzung von Datenquellen führt zu neuartigen Nutzungen, zudem zu Risiken für Benutzer und Organisationen. Wichtige Begriffe in diesem Kontext sind „cyber-physische Systeme" und „Internet der Dinge", relevante Ansätze angepasste Datenbankkonzepte, Cloud Computing und Smart Grid.*

Was ist all den im Zitat genannten Anwendungen gemeinsam? Denken wir an die drei Vs aus dem Vorwort. Big-Data-Analyse bezeichnet die Auswertung riesiger Datenmengen, die eine sehr schwache oder ständig wechselnde Struktur oder Größe besitzen, so dass eine klassische Datenverarbeitung schwierig, wenn nicht gar – wie bei Streamingprozessen – unmöglich wird.

Data Mining und insbesondere Big-Data-Analysen lassen sich erst durch den massenhaften Zugang zu Rechnern mit enormer Rechenleistung und Speicherkapazität wirtschaftlich umsetzen, da erst nach Schaffung dieser technischen Voraussetzungen viele Milliarden von Datensätzen vergleichbar und in Relationen zueinander gebracht werden können. Es ist daher nicht verwunderlich, dass Big Data eine sehr junge Disziplin ist, die aber durch die teilweise überaktive mediale Berichterstattung einen hohen Bekanntheitsgrad besitzt.

In bestimmten Anwendungen, wie z. B. der Auswertung von Daten aus sozialen Medien, haben Data Mining und Big Data bereits einen negativen Beigeschmack bekommen. Hier müssen ethische Richtlinien geschaffen werden, um eben nicht alles auszuwerten, was technisch auswertbar ist. In den technischen Prozessen ist das Potenzial jedoch noch lange nicht ausgeschöpft.

2.1 Definition und Einordnung der Begriffe

2.1.1 Was ist Data Mining?

Schaut man in die Literatur, so wird man sehr schnell mit einer Vielzahl von Termini konfrontiert. Konnektionistische Systeme, multivariate Analysen, Soft Computing, Knowledge Discovery in Data Bases (KDD), Wissen, Information, explizite und implizite Regeln, Maschinelles Lernen und Datenbanktechniken. All diese Begriffe haben etwas mit Data Mining zu tun, doch all diese Begriffe beschreiben nur einen Teilaspekt von Data Mining. Aber die Thematik ist nicht neu. Es gibt nahezu unendlich verschiedene Sichten auf ein und dieselbe Faktenlage. Jede dieser Sichten hebt gewisse Eigenschaften hervor und unterdrückt andere.

Bevor man sich mit Möglichkeiten der Datenauswertungen befasst, muss geklärt werden, was Daten überhaupt sind. Jeder hat natürlich eine intuitive Vorstellung davon, aber wenn man den Begriff „Daten“ definieren soll, fällt es einem schwer.

In der alten Norm DIN 44300 waren Daten *„Gebilde aus Zeichen oder kontinuierliche Funktionen, die aufgrund bekannter oder unterstellter Abmachungen Informationen darstellen, vorrangig zum Zweck der Verarbeitung und als deren Ergebnis.“*

Diese Definition ist jedoch etwas unglücklich, da man Daten durch Informationen zu erklären versucht, später Informationen jedoch als ganz bestimmte Daten einführt. Zirkelschlüsse sind nie gut, jedenfalls nicht, wenn man sich in ein neues Fachgebiet einzuarbeiten versucht.

Schauen wir zu Wikipedia: *„In der Informatik und Datenverarbeitung versteht man Daten gemeinhin als (maschinen-) lesbare und -bearbeitbare, in der Regel digitale Repräsentation von Information. Ihr Inhalt wird dazu meist zunächst in Zeichen bzw. Zeichenketten kodiert, deren Aufbau strengen Regeln folgt, der sogenannten Syntax. Um aus Daten wieder die Informationen zu abstrahieren, müssen sie in einem Bedeutungskontext interpretiert werden. So kann eine Ziffernfolge wie „123456“ zum Beispiel in Abhängigkeit vom Kontext für eine Telefonnummer, eine Kontonummer oder die Anzahl von Kfz-Neuzulassungen in einem bestimmten Zeitraum stehen. Die betrachtete Zeichenfolge „123456“ oder auch „11110001001000000“ als solche kann nur als Aneinanderreihung von Ziffern erkannt werden; ihre konkrete Bedeutung wird erst im jeweils passenden Kontext (siehe Semantik) klar.“* [Wikipedia/Daten]

Diese Erklärung von Wikipedia erzeugt auch kein besseres Verständnis, daher soll der Begriff einfacher definiert werden. Versuchen wir einen intuitiven Zugang. Was sind also Daten? Stellen wir uns eine grüne Wiese vor, mit endlos vielen Grashalmen. Sind diese Grashalme Daten? Intuitiv gesehen, ja. Und würde man diese Grashalme mit einem Lichtmikroskop betrachten, würden sich verschiedene Oberflächenstrukturen zeigen. Sind diese Oberflächenbilder in dem Mikroskop nun

auch Daten? Wieder ja. Aber was ist, wenn man mit einem Elektronenmikroskop noch tiefer in die Details der Grashalmoberfläche eindringen könnte, würden jetzt neue Daten entstehen? Natürlich. Und wenn man auf atomare Ebene hinabsteigen würde? Wie viele Daten entstünden dann?

Man muss diese Fragen dringend beantworten, wenn man über Big Data nachdenkt, sonst verliert man den Blick für die Realität. Eine wichtige technische Frage wäre nämlich, ob man ein Cloud-System entwerfen kann, das alle Daten, die sich aus einer grünen Wiese generieren lassen, speichern könnte? Die Antwort wird ein klares Nein sein. Niemals lässt sich ein System entwerfen, das alle Daten unserer Umwelt erfassen kann, auch Big Data kann das nicht.

Denn was sind Daten nach unserer Definition? Ein Datum ist jede Abhebung vor einem Hintergrund, oder im Plural:

Daten sind Unterscheidungen eines beliebigen Objektes von seiner Umgebung.

Beispiele: Ein schwarzer Punkt auf einer weißen Wand stellt ein Datum dar. Eine Eins, eingebettet in eine Vielzahl von Nullen, stellt auch ein Datum dar. Eine schwarze Struktur auf einer gelben Wand stellt bereits sehr viele Daten dar. Eine 7-stellige Nummer in einem Telefonbuch stellt Daten dar. Aber auch eine Schwarzfärbung auf einer Telefonbuchseite stellt Daten dar. Auch ein gefundenes Haar auf einer solchen Seite ist ein Datum, zum Beispiel, wenn man Kriminologe ist und einen Fall aufklären muss. Wohin wir auch blicken, überall finden wir Daten.

Damit ist bereits klar, dass es in unserer Umwelt unendlich viele Daten gibt, denn nicht nur die Abweichungen eines Vordergrundes zu einem Hintergrund wären Daten, sondern auch die Ableitungen der Abweichungen, und sogar die Ableitungen der Ableitungen usw. Big Data hat daher, an den Möglichkeiten gemessen, noch nicht sehr viele Daten erfasst. Es ist gerade erst der Anfang.

Und jetzt kommt die entscheidende Einschränkung, um aus der riesigen Vielfalt aller Daten die richtigen herauszufiltern, denn wir wollen gar keine Daten, wir wollen Informationen.

Informationen sind Daten mit potenzieller Bedeutung für einen Nutzer. Wie groß die Bedeutung ist, ist dabei noch unbekannt.

Die Daten auf der grünen Wiese stellen für die meisten von uns zwar Daten, aber keine Informationen dar. Aus dem unendlich großen Vorrat der Natur werden diejenigen Daten herausgefiltert, erfasst und ausgewertet, die für irgendjemanden eine Bedeutung haben oder haben könnten; man denke an die heutige Datenerfassung im Telefonverkehr, die alles sammelt, was möglich ist. Man muss das verstehen, wenn man von den großen Datenmengen hört, die mittlerweile auf unseren Cloud-Systemen abgelegt sind. All diese Daten sind eigentlich Informationen, haben also für irgendjemanden eine Bedeutung, zumindest eine potenzielle, sonst wären sie nicht erfasst worden.

Im Alltag werden diese gespeicherten Informationen (leider) wieder Daten genannt, aber eigentlich ist das falsch. Und ganz verwirrend wird es, wenn man sich fragt, wie groß denn die Bedeutung einer Information sei. Denn seit Shannon **[SHA48]** wissen wir zwar, wie wir die Informationsmenge, die sogenannte Entropie E, ausrechnen können, die in einem System enthalten ist oder von diesem übertragen werden soll. Wir wissen aber nicht, wie groß die Bedeutung, die semantische Informationsmenge, nun ganz konkret ist. In **[OTT19]** wurde ein Verfahren vorgeschlagen, um auch die kontextabhängige Bedeutung einer Information auszurechnen. Denn es ist klar, dass sich diese Art der Bedeutung nicht allein aus der Information selbst ergibt. An dieser Stelle soll es uns genügen zu verstehen, dass man die Bedeutung einer Information für einen Empfänger dadurch bestimmen kann, dass man ermittelt, wie groß die Zustandsänderung eines Systems ist, wenn es eine Information (mit der Entropie nach Shannon) empfängt. Dieser Ansatz ist rein intuitiv verständlich, denn er bestimmt eine Bedeutung für die Tragweite, die eine Information für den Empfänger besitzt, und er genügt uns für spätere Diskussionen.

Kommen wir zu dem Begriff der Daten zurück. Im Zeitalter der Digitalisierung versteht man unter Daten diejenigen Informationen, die digital erfasst (da sie eine potenzielle Bedeutung besitzen) und auf den Rechnersystemen der Welt abgelegt wurden. Mit den Daten der oben erwähnten grünen Wiese hat das nicht viel zu tun, die Datenmengen der Umwelt bleiben – wie bereits betont – unendlich groß. Im Folgenden werden wir nur noch von denjenigen Daten als „digitale Daten“ sprechen, die mit digitalen Daten-Technologien ausgewertet werden sollen. Die unmittelbare Auswertung analoger Daten spielt heute kaum noch eine Rolle, wobei das aus Sicht der Autoren falsch ist, denn der Mensch verarbeitet nur analoge Daten und insbesondere im Bereich der Künstlichen Intelligenz wird es eine Rückkehr zu analogen Systemen, zum Beispiel neuromorphen Systemen, geben müssen, siehe Ausblick und **[OTT19]**.

Warum speichert man all diese Daten? Was will man aus den Daten (Informationen) herausfiltern? Die Antwort: Man will Wissen extrahieren und genau dies macht man mit Verfahren des Data Mining.

Data Mining ist ein Prozess der (halb-)automatischen Wissensextraktion aus bereits abgespeicherten (strukturierten) Daten (Informationen).

Und bei Big Data ist dieser Prozess ins 1000-fache überhöht und echtzeitfähig. Um Big Data zu verstehen, muss man sich daher zuerst mit Data Mining befassen, weshalb das Buch gängige Data-Mining-Verfahren ausführlich behandelt. Der Weg zu Big Data ist dann ein leichterer, zumindest auf abstraktem Niveau.

Nun wollen wir den Begriff des Data Mining vertiefen. Bei diesem Begriff gibt es verschiedene Betrachtungsstandpunkte. Der mathematisch interessierte Data-Mining-Ingenieur hebt verschiedene Verfahren der Datenanalyse in den Vorder-

grund, der Informatiker stellt interessante Aspekte der effizienten Implementierung von Data-Mining-Algorithmen dar und der betriebswirtschaftliche Anwender erklärt Data Mining mit einer Reihe von sehr interessanten Projekten im Banken-, Versicherungs- oder Handelsbereich. Es verwundert deshalb nicht, dass unterschiedliche Interessenlagen zu unterschiedlichen Sichtweisen auf dieses Thema führen. Und doch gibt es eine Definition mit dem kleinsten gemeinsamen Nenner:

Data Mining ist ein Prozess zum Auffinden von unbekannten und nicht trivialen Strukturen, Zusammenhängen und Trends in Datenbeständen.

Wir wollen uns dieser allgemeineren Auffassung anschließen, aber noch auf einen Unterschied aufmerksam machen. In der technischen Praxis gibt es sehr häufig Aufgaben, die durch nur wenige Daten charakterisiert sind, im kommerziellen Bereich und bei Social Media liegen dagegen oftmals Hunderttausende oder Millionen von Daten über Kunden und deren Verbrauchsverhalten vor. Daher werden verschiedene Auswertemethoden notwendig sein.

Kommen wir auf die obige Definition zurück: Mit dem Begriff *Prozess* wird verdeutlicht, dass es nicht um Algorithmen und mathematische Verfahren geht, sondern um eine komplexe Betrachtungsweise, die Datenerhebung, Datenselektion, Datenvorverarbeitung, Datenanalyse, Interpretation und Anwendung einschließt. Die Wörter *unbekannt* und *nicht trivial* bedeuten, dass nur dann ein (erfolgreicher) Data-Mining-Prozess vorliegt, wenn neue und hinreichend komplexe Ergebnisse erzielt werden. Aus mathematischer Sicht wäre diese Eingrenzung auf neu und nicht trivial nicht notwendig, aus praktischer Sicht ist sie allerdings sehr wohl entscheidend. Ein triviales Ergebnis wäre beispielsweise die Erkenntnis, dass am Sonntag kein Ausschuss produziert wird, weil eben gar nicht produziert wird. Diese Art von Ergebnissen wird natürlich von den verwendeten Data-Mining-Algorithmen auch geliefert. Es wird damit klar, dass an jede Analyse eine Interpretationsphase angeschlossen sein muss.

Um die Begriffe *Strukturen, Zusammenhänge* und *Trends* zu erklären, betrachte man nachfolgende Datentabelle aus einem Produktionsprozess, Bild 2.1:

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Wochentag	x1	x2	x3	x4	x5	x6	x7	x8	...	xi	xj	...	...	y1	y2	y3	y4
5	1920	0,1	8	8	100	290,00	7,7	192	192	192	1400	100	1400	20,8	0,1	5	14
6	8630	0,2	24	24	350	290,00	7,7	3020,5	2828,5	3212,5	400	-80	400	22	0,1	14	9
0	9960	0,3	12	12	580	290,00	6,7	5776,8	2756,3	8989,3	1200	0	1200	23,2	0,5	9	14
1	9770	3,2	120	120	790	290,00	8,3	7718,3	1941,5	16707,6	1200	20	1200	24,4	0,7	14	18
2	9730	0,2	140	140	750	290,00	7,9	7297,5	-420,8	24005,1	1000	-140	1000	25,5	2,5	18	15
3	9730	30,4	52	52	770	290,00	7,4	7492,1	194,6	31497,2	2400	111,429	2400	27,1	3,5	15	6
4	9620	29	20	53,71	600	290,00	11,9	5772	-1720,1	37269,2	1400	1,42857	1286	27,3	1,3	6	9
5	10610	7	380	106,9	600	300,00	9,6	6366	594	43635,2	1300	-22,857	1271	27,5	1,3	9	6
6	10130	4,8	288	144,6	470	303,33	9,1	4761,1	-1604,9	48396,3	2000	-11,429	1500	27,8	0,1	6	4
0	9460	1	1170	310	620	306,67	11,8	5865,2	1104,1	54261,5	2000	-11,429	1614	28	0,1	4	2
1	9150	0,1	1336	483,7	730	310,00	11,9	6679,5	814,3	60941	2000	-25,714	1729	28,4	0,1	2	8
2	8620	23	588	547,7	680	313,33	11,5	5861,6	-817,9	66802,6	2800	14,2857	1986	28,1	0,4	8	13
3	8220	36	484	609,4	710	316,67	12,2	5836,2	-25,4	72638,8	1400	-20	1843	29	0,8	13	7
4	8930	19	540	683,7	710	320,00	11,2	6382,9	546,7	79021,7	2800	-21,429	2043	29,4	0,9	7	10
5	9000	35	704	730	630	300,00	12,1	5670	-712,9	84691,7	2800	8,57143	2257	29,7	0,4	10	9
6	9130	38	648	781,4	490	318,33	11,9	4473,7	-1196,3	89165,4	1400	0	2171	28,6	0,3	9	12
0	10220	35	600	700	650	336,67	11,7	6643	2169,3	95808,4	2000	0	2171	28,5	0,6	12	10
1	8860	28	1240	686,3	650	355,00	10,5	5759	-884	101567	2000	11,4286	2171	27,8	0,1	10	7
2	9000	22	636	693,1	840	373,33	11,9	7560	1801	109127	2000	-5,7143	2057	27,3	0,1	7	8
3	9460	33	616	712	810	391,67	11,9	7662,6	102,6	116790	1800	11,4286	2114	27,5	0,1	8	13
4	8860	5,5	240	669,1	830	410,00	9	7353,8	-308,8	124144	2000	5,71429	2000	28,7	0,3	13	12
5	9430	20	496	639,4	710	310,00	12,1	6695,3	-658,5	130839	2400	-22,857	1943	28,5	0,5	12	29
6	9250	29	548	625,1	670	293,33	11,9	6197,5	-497,8	137037	3000	-14,286	2171	28,2	0,5	29	26
0	11260	6,5	276	578,9	1020	276,67	8,8	11485,2	5287,7	148522	3000	-5,7143	2314	28,4	0,1	26	21
1	11390	4,5	416	461,1	390	260,00	10,6	4442,1	-7043,1	152964	2400	-5,7143	2371	27	0,1	21	16
2	10180	14	404	428	480	243,33	11,4	4886,4	444,3	157850	2400	-8,5714	2429	27	1	16	11
3	8920	16	504	412	410	226,67	11,2	3657,2	-1229,2	161508	2400	0	2514	26,7	0,3	11	8
4	10630	25	540	454,9	460	210,00	11,6	4889,8	1232,6	166397	2000	5,71429	2514	26,7	0,2	8	6
5	10240	34	712	485,7	480	230,00	11,9	4915,2	25,4	171313	2000	14,2857	2457	27,5	0,1	6	5
6	11410	33	624	496,6	400	235,00	11,6	4564	-351,2	175877	2000	17,1429	2314	27,9	0,2	5	4
0	11690	13	520	531,4	760	240,00	11,3	8884,4	4320,4	184761	1800	5,71429	2143	28	0,1	4	4
1	12410	23	520	546,3	600	245,00	10,3	7446	-1438,4	192207	2000	5,71429	2086	26,3	0,1	4	5
2	10860	25	1176	656,6	560	250,00	9,6	6081,6	-1364,4	198289	2000	14,2857	2029	26,1	0,1	5	11
3	8870	28	468	651,4	550	255,00	11,5	4878,5	-1203,1	203167	1400	2,85714	1886	27,2	0,2	11	8
4	11360	2	472	641,7	530	260,00	11,3	6020,8	1142,3	209188	1800	8,57143	1857	27,7	0,1	8	10
5	10120	10	400	597,1	590	270,00	9,6	5970,8	-50	215159	1400	0	1771	28,2	0,4	10	11
6	10480	14	420	568	380	268,33	10,3	3982,4	-1988,4	219141	2000	-2,8571	1771	28,2	0,1	11	6

Bild 2.1 Datentabelle aus einem Produktionsprozess

In den Spalten stehen die einzelnen Prozessvariablen x_1, x_2, x_3, … und die Zielgrößen y_1, y_2, … In den Zeilen stehen die Ausprägungen dieser Variablen für ihre Abtastzeiten, im Beispiel des Bildes 2.1 ist die Abtastzeit einmal pro Tag, siehe Spalte Wochentag (Sonntag bis Samstag sind mit 0 bis 6 gekennzeichnet).

Man kann sich vorstellen, dass es nun mehrere Möglichkeiten gibt, eine solche Tabelle auszuwerten, nämlich spalten- oder zeilenweise. Die Data-Mining-Anwendungen und Algorithmen unterscheiden deshalb folgende Analysemöglichkeiten:

- Zusammenhangsanalyse: Man sucht nach Zusammenhängen zwischen den Spalten der Tabelle, also nach Zusammenhängen der Form $y = f(x_1,x_2,x_3, \ldots)$,
- Strukturanalyse: Man sucht nach Zusammenhängen zwischen den Zeilen der Tabelle, also den Ausprägungen untereinander,
- Trendanalyse: Man sucht Vorhersagemöglichkeiten einer Zielgröße, entweder nur aus den Eingangsgrößen oder aus den Eingangsgrößen und der Zielgröße selbst.

Selbstverständlich gibt es dabei Überschneidungen. Dennoch ist es sehr hilfreich, wenn man am Beginn der Analyse weiß, ob man eher nach Strukturen oder eher

nach Zusammenhängen in den Daten sucht. Um diesen wichtigen Punkt noch besser zu verdeutlichen, sei nachfolgendes einfaches Datenbeispiel aus Bild 2.2 mit zwei Variablen x_1 und y_1 gegeben (Definitionsbereich von x_1 sei 36 bis 45,5; dargestellt sind aus Platzgründen nur Werte von 36 bis 42).

	A	B
1	x1	y1
2	36	3700
3	36	5200
4	36,5	2700
5	36,5	4100
6	36,5	1700
7	37	400
8	37	3900
9	37	4600
10	37,5	2200
11	37,5	2400
12	37,5	3500
13	38	4800
14	38	5200
15	38	800
16	38,5	1700
17	38,5	3400
18	38,5	3400
19	39	2600
20	39	4700
21	39	400
22	39,5	3100
23	39,5	2600
24	39,5	4000
25	40	800
26	40	6100
27	40	2500
28	41,5	5700
29	41,5	7200
30	41,5	4700
31	42	6100
32	42	3700

Bild 2.2
Zwei Variablen x_1 und y_1

Durch einen Scatter-Plot (x-y-Plot oder Streudiagramm) nach Bild 2.3 erkennt man bereits rein visuell Zusammenhänge und Strukturen:

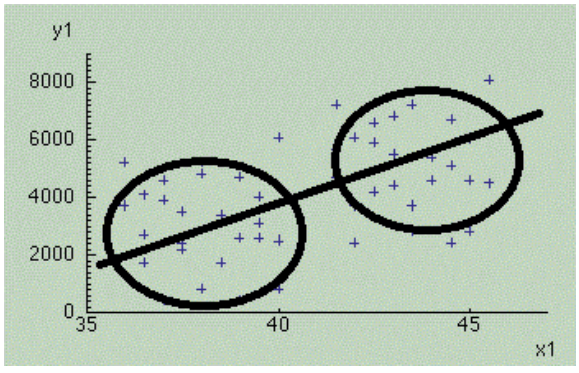


Bild 2.3
Scatter-Plot der Variablen x_1 und y_1 mit eingezeichneten Ellipsen für Strukturen und einer Gerade für Zusammenhänge

Die Gerade in Bild 2.3 zeigt, dass es einen schwachen Zusammenhang zwischen y_1 und x_1 gibt. Die beiden Ellipsen verdeutlichen die beiden Untergruppen des Datensatzes, also eine Struktur.

Ausgehend von diesen rein visuellen Überlegungen lassen sich folgende Datenanalysen durchführen:

- Korrelationsanalyse (ungerichtete Zusammenhangsanalyse)
- Regressionsanalyse (gerichtete Zusammenhangsanalyse)
- Clusteranalyse (Strukturanalyse)

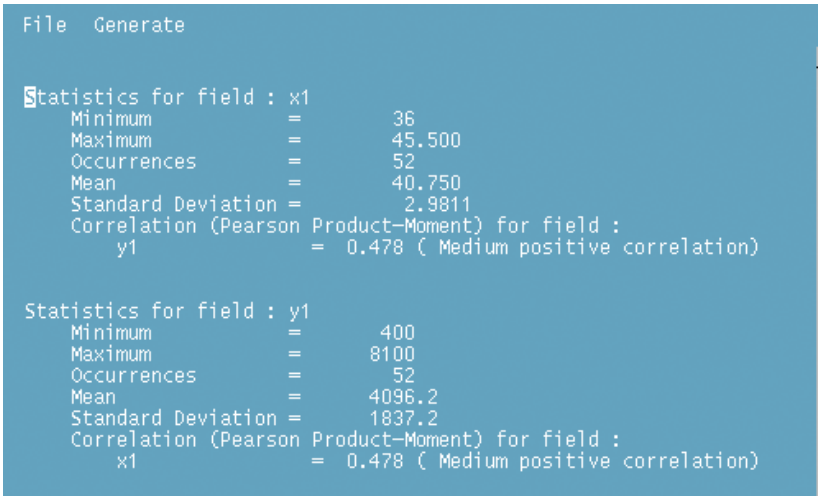


Bild 2.4 Univariate Statistik und Korrelation der Variablen x_1 und y_1

Die Korrelationsanalyse nach Bild 2.4 zeigt eine schwache positive Korrelation zwischen den Variablen x_1 und y_1 von ca. 0,47 (Pearson-Product), d. h., man erkennt folgende, allerdings nur schwach ausgeprägte Regel: Je größer (kleiner) x_1, desto größer (kleiner) ist y_1 bzw. je größer (kleiner) y_1, desto größer (kleiner) ist x_1.[1]

Des Weiteren sieht man in Bild 2.5, dass man mit einer Genauigkeit von 85 % (bei dieser Angabe bezogen auf den Messbereich $y_{1max} - y_{1min}$) den Wert y_1 aus x_1 mit der Technik der neuronalen Netze schätzen kann. Natürlich könnte man den Zusammenhang auch durch eine einfache Regressionsgerade herausfinden.

[1] Korrelationswerte zwischen zwei Variablen, deren absoluter Betrag kleiner als 0,75 ist, werden hier als schwach bezeichnet.

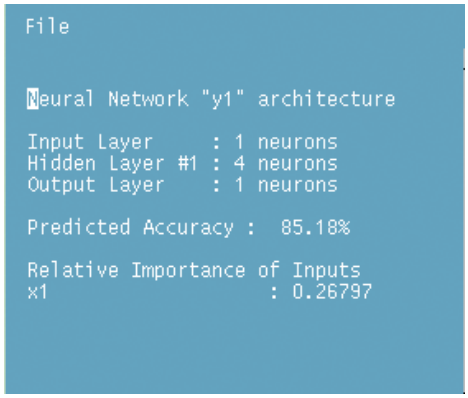


Bild 2.5
Neuronales Vorhersagemodell $y_1 = f(x_1)$

Weiterhin sind statistisch auch zwei Cluster voneinander separierbar, Bild 2.6.

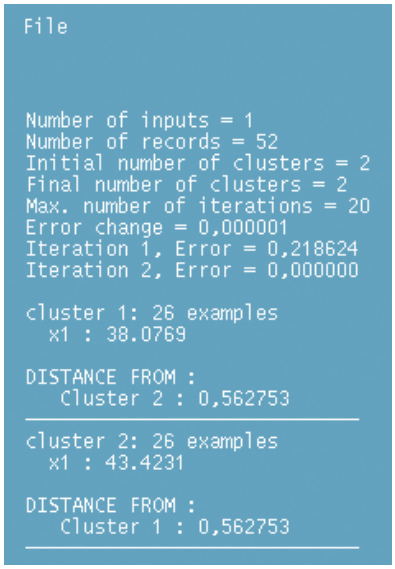


Bild 2.6
Clusteranalyse der Daten x_1 und y_1

Fragt man also nach Zusammenhängen in diesem Datensatz, kann man mittels einer Regression (im Beispiel mit einem neuronalen Modell) erkennen, dass sich y_1 durch x_1 erklären lässt (aber auch umgekehrt). Fragt man nach Strukturen, so sieht man, dass es zwei Untergruppen (Cluster) gibt. Je nach Fragestellung erhält man unterschiedliche Informationen über die in den Daten enthaltenen Zusammenhänge und Strukturen.

Obwohl das dargestellte Beispiel sehr einfach ist, zeigt es prinzipielle Möglichkeiten bei der Auswertung von Datensätzen. Es ist notwendig, vor dem Beginn von Data Mining festzulegen, was man in den Daten suchen möchte. Will man wissen,

wie viele Untergruppen der Datensatz besitzt (wichtig zum Beispiel, um stabile und instabile Produktionszustände voneinander zu unterscheiden), dann benutzt man Methoden der Clusteranalyse. Fragt man nach Abhängigkeiten der Variablen untereinander, sind Methoden der Zusammenhangsanalyse notwendig. Welche Algorithmen letztendlich angewendet werden sollten, wird in den weiteren Kapiteln dargestellt. Data Mining beinhaltet eine solche Vielzahl von möglichen Algorithmen, dass dabei eine Systematik unerlässlich wird (siehe Kapitel 3).

2.1.2 Was ist Big Data?

Wie bereits erläutert, zielt der Begriff Big Data nicht so sehr auf die Auswerteverfahren ab, sondern auf drei andere Dimensionen der Daten: ihre großen Volumina, ihre große Geschwindigkeit bei der Entstehung und Verarbeitung und die große Vielfalt ihrer (unstrukturierten) Speicherformate. Big Data ist daher ein Sammelbegriff zur Erhebung, Speicherung und Auswertung von Massendaten mittels digitaler Technologien.

Schauen wir zuerst auf die Datenmengen. Man sagt, das weltweite Datenvolumen verdopple sich alle zwei Jahre. Andere Schätzungen geben an, dass es 2016 ca. 16 Zettabyte an digitalen Daten gab, im Jahre 2025 soll die Menge auf 163 Zettabyte angewachsen sein.

Der Begriff Zettabyte ist noch nicht ganz so geläufig, er entspricht 10^{21} Bytes oder besser einer Milliarde (10^9) Terabyte (10^{12}). Die Größe Terabyte ist bekannt, da man mittlerweile große Terabyte-Festplatten im Elektrohandel kaufen kann, allerdings sind eine Milliarde Terabyte-Festplatten dann doch schon wieder ziemlich viel.

Aber man kann sich diese Zahlen auch anders verdeutlichen: Im sichtbaren Universum soll es 10^{22} Sterne geben. Und wäre die Erde mit einer 10 Meter hohen Sandschicht bedeckt, würde es ca. 10^{28} Sandkörner geben. Daten im Zettabyte-Bereich sind also schon sehr viele, aber die Natur hält noch größere Zahlen bereit. Dass es in der Natur unendlich viele Daten gibt, hatten wir bereits zu Beginn erkannt.

Die jüngeren Leser werden sich im Laufe ihres Lebens noch an Yottabytes (10^{24} Byte) gewöhnen müssen, nochmals 1000-mal mehr an digitalen Daten als Zettabyte. Es ist klar, dass das Speichern und Auswerten dieser großen Mengen sehr anspruchsvoller Technologien bedarf. Und es sind natürlich Spezialisten nötig, die mit diesen Techniken zurechtkommen. Data-Mining- und Big-Data-Spezialisten werden zu den gefragtesten Fachleuten der Zukunft gehören.

Ein weiterer Faktor für Big Data ist die Geschwindigkeit der Erfassung und Verarbeitung. Es ist einleuchtend, dass die Entwicklung erst dann beendet sein wird,

wenn alle Anwendungen in Echtzeit stattfinden. Das heißt natürlich, dass jede Anforderung ihre eigene Geschwindigkeit benötigt. Will man beispielsweise Steuersünder aus einer riesigen Menge von Daten herausfiltern, hat man sicher eine ganze Nacht Zeit. Beim autonomen Fahren ist jedoch bereits eine Sekunde merkliche Rechenzeit zu viel. Hier fallen riesige Datenmengen in Form von Bilddaten an, darauf muss das Auto ohne merkliche Reaktionszeit reagieren, etwas, was für vollautonome Level-5-Fahrzeuge sicher erst in 30 bis 50 Jahren erreicht werden wird, wenn überhaupt.

Aber gerade auch die besonders große Vielfalt der weltweit möglichen Datenformate charakterisiert Big Data. Nur ein kleiner Teil aller digitalen Daten liegt geordnet in Datenbanken. Der größte Teil ist unstrukturierter Text, Bild oder Ton. Damit besteht eine der größten Herausforderungen darin, aus diesen unstrukturierten und sich ständig ändernden Massendaten so schnell wie nötig Wissen zu generieren, um den angeschlossenen Systemen noch in Echtzeit eine Entscheidungsgrundlage zu bieten.

Warum ist aber um Big Data ein regelrechter Hype ausgebrochen? An den zugrundeliegenden Analyseverfahren des Data Mining kann es nicht liegen, denn die meisten Verfahren zur Auswertung sind in den 1980er Jahren entstanden. All diese Verfahren gehören, wenn sie nicht den statistischen Verfahren zugeordnet werden, in den Bereich des maschinellen Lernens. Aber das maschinelle Lernen ist mindestens eine Generation alt. Der Hype entstand mit den neuen und/oder vermeintlich neuen Möglichkeiten, die Verfahren des maschinellen Lernens auf riesige Datenbestände anzuwenden, um damit neue Muster und Strukturen zu entdecken. Dabei ist natürlich auch viel Überschätzung der Verfahren enthalten, denn Big-Data-Analysen können zwar in Echtzeit eine Vielzahl von Aussagen extrahieren, aber sie basieren auf statistischen Methoden (Kapitel 3 und Kapitel 4), die stets mit einer Unsicherheit (Irrtumswahrscheinlichkeit) verbunden sind.

2.1.3 Data Mining im Kontext anderer Datenanalyseverfahren

Die nachfolgende Übersicht in Bild 2.7 zeigt eine Einbettung von Data Mining in die benachbarten Disziplinen. Data Mining ist dabei das verbindende Element, das bei richtiger Anwendung gestattet, die Optimierungspotenziale in den praktischen Applikationen auszuschöpfen. Betrachten wir die Übersicht genauer:

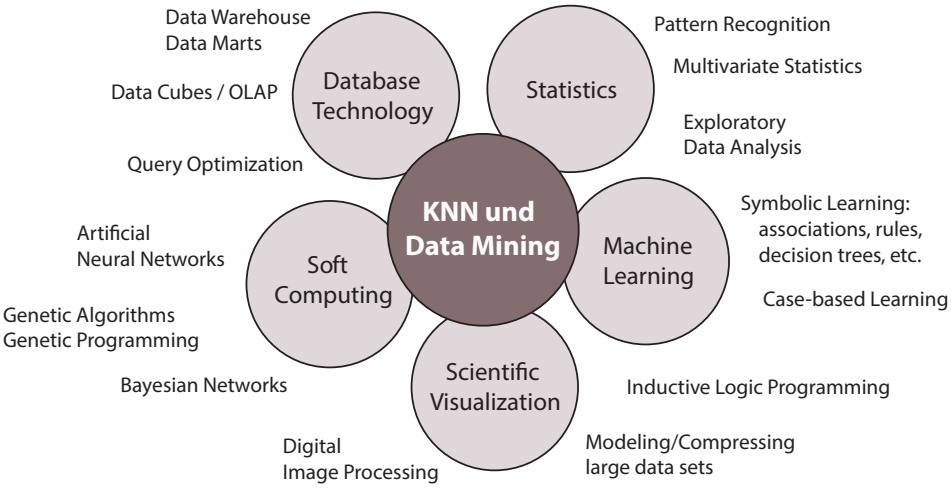


Bild 2.7 Einordnung von Data Mining

1. Data Mining und Data-Base-Technologien

Im Allgemeinen wird Data Mining als die Krone der Data-Base-Technologien bezeichnet, die aufbauend auf einem realisierten Data Warehouse (DW) oder im Big-Data-Umfeld auf einem Data Lake komplexe Zusammenhänge in den Daten finden kann.

Mittlerweile sind die Abfragetechniken (Queries) in einem Data Warehouse oder Data Lake so komplex, dass ein Laie die daraus erhaltenen Informationen mit denen des Data Mining gleichsetzen könnte. Dennoch gibt es einen wesentlichen Unterschied, denn mit den Queries muss man einzeln Frage für Frage stellen und bekommt diese dann im Ergebnis der Query-Abfragen beantwortet. Durch geschickte Abfragen kann man damit interessante Zusammenhänge in den Daten erkennen. Data Mining erkennt interessante Zusammenhänge aber automatisch.

Zum Beispiel könnte man folgende Frage durch Query-Techniken stellen: „Wie ist die Ausschussrate eines Produktes A jeden Montag bei einer Raumfeuchte von 50 % und einer Außentemperatur von 25 Grad Celsius bei der Schicht vom Schichtleiter Herrn Meyer im letzten Jahr gewesen?“ Die Frage lässt sich leicht in ein entsprechendes Datenbank-Statement übersetzen, das Data Warehouse kann daraufhin abgefragt werden und im Ergebnis dieser Anfrage wird eine Liste von Ausschussraten bei den vorgegebenen Bedingungen geliefert. Man kann auch einen Ausschussratenwert erhalten, der dann sofort mit der mittleren Ausschussrate eines Produktes A vergleichbar ist und Aufschluss darüber gibt, ob bestimmte Bedingungen die Ausschussrate signifikant verändert haben. Und dennoch ist das kein Data Mining. Denn es ist ein wesentlicher Unterschied, ob man diese Frage (Query) stellen muss, um zu dem Ergebnis zu kommen, oder ob ein Algorithmus

selbständig ermittelt, dass immer dann, wenn Produkt A am Montag bei Herrn Meyer produziert wird und die Außentemperatur 25 Grad Celsius ist, die Ausschussrate um z. B. 10 % höher liegt als im Durchschnitt.

Mit DW-Techniken kann man im Allgemeinen Hypothesen verifizieren, mit Data Mining lassen sich Hypothesen automatisch erzeugen.

Die seit 20 Jahren eingeführten Online-Analytical-Processing-(OLAP)-Techniken waren eine Weiterentwicklung der Datenbankabfragen. Hat man ein mehrdimensionales Data Warehouse aufgebaut, sind mehrdimensionale Zusammenhänge leicht darstellbar, Bild 2.8.

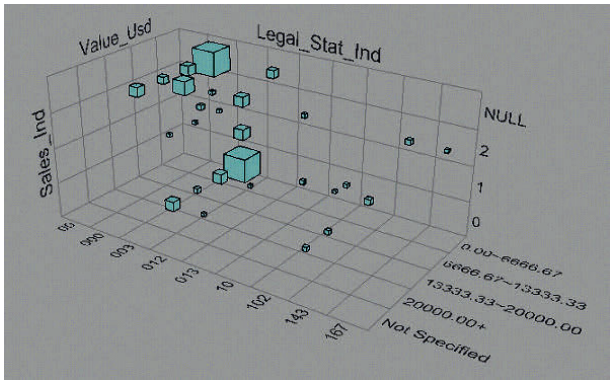


Bild 2.8
OLAP-Darstellung für Verkaufsvolumen in Abhängigkeit dreier Variablen

Im Beispiel aus Bild 2.8 ist die Größe des Verkaufsvolumens in der Größe des Würfels codiert und über die abhängigen Variablen Legal_Stat_Ind (Regionalindex) in der x-Achse, Value_Usd (Größe der Einzelverkäufe in US-Dollar) in der y-Achse und Sales_Ind (Verkaufsindex pro Industriebranche) in der z-Achse dargestellt. OLAP bedeutet für den Anwender nun, dass er interessante Würfel (z. B. sehr große oder sehr kleine) anwählen und sich in diese „hineinzoomen“ kann und dadurch detailliertere Informationen darüber erhält, unter welchen Bedingungen das Verkaufsvolumen besonders groß oder besonders klein war. Diese Form der Datenvisualisierung und der explorativen Analysen ist sehr mächtig und kann als visuelles Data Mining bezeichnet werden, da man die interessanten Gruppen mit dem Auge erkennt und die dazugehörigen Bedingungen durch das System berechnen lässt. Das eigentliche Data Mining wird also visuell durchgeführt, ein durchaus praxisnahes Verfahren, wie in Kapitel 3 bei den sogenannten SOM-Karten gezeigt werden wird.

DW-Methoden werden immer universeller und Data-Mining-Techniken werden zunehmend als Bestandteil in die Warehouse- und Big-Data-Plattformen integriert. Es findet eine gegenseitige Integration und Befruchtung statt.

2. Data Mining, Big Data und multivariate Statistik

Die Abhängigkeiten zwischen Data Mining und Statistiken lassen sich am besten wie folgt beschreiben: Es gibt keinen Data-Mining-Prozess ohne detaillierte Statistik, aber nicht jede detaillierte Statistik ist ein Data-Mining-Prozess. Statistik ist und bleibt die Basis jeder Datenanalyse und damit auch jedes Data-Mining-Prozesses.

Während univariate Analysen noch ganz klar der reinen Statistik zugeordnet werden können, so sehen manche Autoren keine scharfe Grenze mehr zwischen multivariater Statistik und Data Mining. Letztendlich baut Data Mining auf multivariater Statistik auf und kann als fortgeschrittene, nichtlineare, multivariate Datenanalyse betrachtet werden.

Als multivariate oder mehrdimensionale Datenanalysen werden alle Verfahren bezeichnet, die die Abhängigkeiten zwischen mehreren Merkmalen untersuchen. Die multivariate Analyse beschäftigt sich mit der Entwicklung von Modellen „… zur Analyse einer nicht näher spezifizierten Anzahl abhängiger Variablen“ **[SAC97, S. 576]**. Aus der Literatur sind zahlreiche multivariate Datenanalyseverfahren bekannt, von denen die wichtigsten im Kapitel 3 erläutert werden. Insgesamt kann zwischen prüfenden und entdeckenden Analyseverfahren unterschieden werden.

Prüfende Analyseverfahren: Die prüfenden Verfahren der multivariaten Statistik gehen von einem (gerichteten) Zusammenhang zwischen Variablen der Form $y = f(x_1, x_2, x_3 \ldots)$ aus und ihr erklärtes Ziel ist es, diesen Zusammenhang zu überprüfen. Folgende multivariate Verfahren lassen sich den prüfenden Verfahren zuordnen:

- Regressionsanalyse
- Diskriminanzanalyse
- Varianzanalyse
- Kontingenzanalyse

Sie unterscheiden sich im Grunde nur durch das entsprechende Skalenniveau der unabhängigen Variablen x_i und der abhängigen Variablen y, siehe Tabelle 2.1.

Tabelle 2.1 Einteilung der Datenanalyseverfahren

Abhängige Variablen *y*	Unabhängige Variablen $x_1, x_2, x_3, \ldots$	
	metrisch skalierbare Daten	nominal skalierbare Daten
metrisch skalierbare Daten	Regressionsanalyse	Varianzanalyse
nominal skalierbare Daten	Diskriminanzanalyse	Kontingenzanalyse

Beispielsweise wird die Regressionsanalyse dann eingesetzt, wenn sowohl unabhängige als auch abhängige Variablen metrisch skalierbare Daten (numerisch) sind. Eine detaillierte Einführung erfolgt in Kapitel 3.

Entdeckende Analyseverfahren: Bei diesen Verfahren geht es grundsätzlich um die Entdeckung von neuen Zusammenhängen oder Strukturen in vorliegenden Daten. Der Anwender muss keine Trennung in unabhängige und abhängige Variablen durchführen. Oftmals ist eine solche Trennung auch nicht möglich, da das Ziel der Untersuchung gerade darin besteht, neue Zusammenhänge zu ermitteln. Folgende entdeckende Verfahren, die alle auf metrisch skalierbaren Daten aufbauen, können unterschieden werden:

- Korrelationsanalyse
- Faktoranalyse
- Clusteranalyse

Die Korrelationsanalyse überprüft, ob lineare Zusammenhänge zwischen zwei oder mehreren Variablen vorliegen. Die Faktoranalyse untersucht die Frage, ob sich die zugrundeliegende Anzahl von Variablen auf weniger Faktoren zurückführen lässt, d. h., es geht um die Entdeckung neuer zugrundeliegender Variablen. Dadurch soll eine Reduktion und Verdichtung der existierenden Variablen durchgeführt werden.

Die Clusteranalyse nimmt eine Verdichtung von Objekten vor, d. h., sie untersucht die Frage, ob sich die Objekte (Datenbeispiele) so in Gruppen zusammenfassen lassen, dass sich die Mitglieder einer Gruppe möglichst ähnlich, die Gruppen untereinander aber möglichst unähnlich sind. Damit geht es bei der Clusteranalyse um die Entdeckung neuer Strukturen, während die Korrelationsanalyse neue Zusammenhänge entdecken soll.

Allerdings lassen sich Überschneidungen zwischen prüfenden und entdeckenden Verfahren nicht vermeiden, so wird die Regressionsanalyse auch oft zum Entdecken und nicht nur zum Verifizieren neuer Zusammenhänge verwendet. Insbesondere die Sensitivitätsanalyse auf Basis von Regressionsmodellen kann zum Erkennen neuer qualitativer und quantitativer Zusammenhänge führen und gerade Regressionsverfahren auf Basis moderner Techniken – wie neuronale Netze – können für diese Analysen eingesetzt werden.

Letztendlich können damit auch alle Data-Mining-Verfahren (wie neuronale Netze, deep learning, Entscheidungsbäume, Regressionsbäume oder das MIDOS-Verfahren) in die multivariate Systematik und Statistik eingebracht werden.

Data Mining steht dabei aber immer für eine (nichtlineare) automatische Wissensgenerierung, während klassische multivariate Verfahren oftmals nur lineare Methoden beinhalten. Der Begriff „Statistik“ steht im Umfeld des Data Mining für die quantitative Signifikanzuntersuchung des mit Data Mining gefundenen Wissens.

3. Data Mining und maschinelles Lernen

Maschinelles Lernen beinhaltet Techniken, mit denen man strukturierte Muster in Daten erkennen und beschreiben kann, mit dem Ziel, die Daten zu erklären und Voraussagen zu treffen. Maschinelles Lernen beschreibt damit genau die neueren Data-Mining-Algorithmen (ohne auf den gesamten Prozess des Data Mining, wie Datenerzeugung, -auswahl, -vorverarbeitung, -analyse und Interpretation einzugehen).

Data Mining basiert also im erheblichen Maße auf maschinellem Lernen.

Gängige Verfahren sind:

- Entscheidungsbäume
- Regelbäume
- Assoziationsregeln
- Clusteranalysen
- Support-Vector-Machines
- Neuronale Netze

Auf diese Verfahren wird in Kapitel 3 eingegangen. Unter anderem **[GER19], [GUI16]** und **[ALP14]** geben eine gute und detaillierte Darstellung von maschinellen Lernverfahren für das Data Mining. **[OTT19]** gibt eine Einordnung des maschinellen Lernens im Kontext der Künstlichen Intelligenz.

4. Data Mining und (wissenschaftliche) Visualisierungstechniken

Jeder, der in der Praxis Data-Mining-Projekte durchführt, bemerkt sehr schnell, dass der Erfolg eines Projektes auch von den verwendbaren Visualisierungstechniken abhängt. Selbstverständlich gibt es eine Vielzahl von statistischen Kennzahlen; unabhängig davon wurden visuelle Darstellungsformen von Datenstrukturen entwickelt:

- Plots (für numerische Größen)
- Scatter-Plots (für numerische Größen)
- Histogramme (für numerische Größen)
- Verteilungen (für nominale Größen)
- Link Nodes (für nominale Größen)
- Self Organizing Maps (SOM) (für numerische und nominale Größen)

Die nachfolgenden Kapitel geben dazu viele Beispiele.

5. Data Mining und Soft Computing

Der Begriff Soft Computing verdeutlicht einen Trend hin zu unscharfen und robusten Analysemethoden. Seit den 1980er Jahren war dieser Trend in vielen industriellen Einsatzbereichen feststellbar. Der Grund liegt darin, dass es im industriellen Umfeld eine Vielzahl von komplexen Problemstellungen gibt, die nicht mit exakten und scharfen mathematischen Algorithmen behandelt werden können.

Während die Menschen in der Regel keine Schwierigkeiten damit haben, beispielsweise die Begriffe *kühl, weniger kühl, warm* und *heiß* in ihrem jeweiligen Kontext richtig einzuordnen, ist das für Softwareprogramme eine große Herausforderung. Erst durch die Einführung des Begriffes von unscharfen Mengen und der darauf aufgebauten Fuzzy-Theorie konnte man derartige Aufgaben lösen, da unscharfe Begriffe nun algorithmisch behandelbar wurden.

Eine weitere industrielle Aufgabe seit den 1980er Jahren besteht in der robusten Modellierung von nichtlinearen und analytisch nicht beschreibbaren Prozessstrecken. Gerade bei der Analyse komplexer, mehrmals umgebauter oder gealterter Anlagen versagen (scharfe) Modellierungsansätze mittels Differentialgleichungen. Eine algorithmische Antwort auf diese Herausforderungen brachten die neuronalen Netze, deren Informationsverarbeitungsmechanismen denen der biologischen Gehirne nachempfunden wurden.

Mittlerweile haben sich folgende wichtige Soft-Computing-Disziplinen in der Praxis etabliert:

- Fuzzy-Techniken bzw. Fuzzy-Control
- Neuronale Netze und Deep Learning
- Genetische Algorithmen und Evolutionsstrategien

In diesem Buch werden im Bereich der Fuzzy-Datenauswertung nur Fuzzy-Clusterverfahren vorgestellt (Kapitel 3). Für eine Einführung in die Fuzzy-Techniken bzw. Fuzzy-Control wird auf die Literatur verwiesen. Hier sollen beispielsweise die für Ingenieure gut geeigneten Bücher **[KAH95]** und **[BOK96]** genannt werden. **[TSO08]** stellt eine gute Einführung von Fuzzy und Neuro-Methoden dar. **[CRU16]** gibt einen Einblick in Fuzzy-Techniken in LabVIEW.

Neuronale Netze werden im Kapitel 3 vorgestellt. Einen detaillierten und immer noch hochaktuellen Einstieg in die Theorie neuronaler Netze bietet der Klassiker **[ZEL94]**. Auch die Bücher **[BRAU95]** und **[BRA97]** können weiterhin empfohlen werden. Auf den aktuellen Bestsellerlisten rangiert **[RAS17]**, der sehr gut neuronale Netze in Python erklärt. **[OTT19]** beschreibt detailliert, wie ein Backpropagation-Netzwerk arbeitet und erklärt das Prinzip von Deep-Learning-Netzen am Beispiel eines Faltungsnetzes CNN.

Einen guten Überblick über Methoden der genetischen Algorithmen und Evolutionsstrategien geben **[SCHÖ94]** und **[SCHWF02]**, aus Platzgründen werden diese Techniken im vorliegenden Buch nicht erläutert.

6. Data Mining und Big Data

Zur besseren Unterscheidung der in diesem Buch behandelten Hauptbegriffe „Data Mining“ und „Big Data“ soll das bisher Gesagte nochmals wiederholt werden. Big Data stellt Daten sofort und in ihrer vollständigen Breite fachlich unstrukturiert und unverändert zur Verfügung. Strukturiert werden die Daten dann, wie beim Konzept des Data Warehouses (DW) oder Data Marts, in Datenmodellen, jedoch ohne die Rohdaten zu verlieren; diese werden zusätzlich gespeichert. Somit können Data-Mining-Verfahren auf den Rohdaten oder/und auf mehreren abgeleiteten Datenmodellen, Datenausschnitten oder -sichten agieren. Mit entsprechend skalierbarer, parallelisierbarer Software und Hardware können fachliche Zusammenhänge mittels statistischer Modelle zur Laufzeit ermittelt werden. Der Begriff Big Data ist dabei wie bereits besprochen sehr vielschichtig und reicht von der allgemeinen Charakteristik über Datenvolumina, Verarbeitungsgeschwindigkeit und Formatdiversität (drei „Vs“) bis hin zu vollumfänglichen Big-Data-Plattformanbietern in der Cloud, welche Dienste zur Datenaufnahme, -speicherung und -analyse (einfache und erweiterte) bis hin zur Datenvisualisierung zur Verfügung stellen.

Unter Data Mining – der erweiterten Datenanalyse (advanced analytics) – wird die Weiterentwicklung von der beschreibenden Analyse (descriptive analytics) über die vorhersagende Analyse (predictive analytics) bis hin zur empfehlenden Analyse (prescriptive analytics) verstanden, welche auf Grund der Vorhersage maßgeschneiderte Handlungsempfehlungen zur Verbesserung der Situation gibt.

Big Data stellt – umfassend betrachtet – Technologien zur effizienten Nutzung von Data-Mining-Algorithmen und die erforderliche Datenbasis zur Verfügung, eine Datenbasis, oft Data Lake oder Data Hub genannt, die im Vergleich zu vorhergehenden Jahrzehnten größer ist, sich schneller bewegt und wächst, aber auch vielfältiger in Strukturen und Formaten ist.

2.2 Spezielle Anforderungen der Industrie an die Datenanalyse

Während in den vorhergehenden Abschnitten wichtige Begriffe und Einordnungen gegeben wurden, wird in diesem Abschnitt stellvertretend am Beispiel der Prozessindustrie dargestellt, welche Anforderungen die Industrie an die Datenanalyse stellt. Für Fertigungsindustrie und Produktmodellierung gelten die Aussagen analog.

Aus zahlreichen Data-Mining- und Big-Data-Projekten und Gesprächen mit Ingenieuren, Betriebsleitern und Werksleitern wurden folgende „zeitlose“ Forderungen deutlich, die die Verantwortlichen an die Datenanalyse stellen:

- Datenreduktion und Handlungsempfehlung im Fehlerfall
- Prozessvisualisierung
- Prozessmodellierung
- Prozessprognose
- Prozessanalyse und Prozessoptimierung
- Reduzierung von Engineeringaufwänden

Gewünschte Unterstützungen in den oben genannten Bereichen sind deshalb als „zeitlos“ zu betrachten, da sich die Wünsche der Manager in den 1990er, den 2000er und den heutigen Jahren nicht grundlegend unterscheiden. Datenanalysen sind und bleiben nur Mittel zum Zweck, den Verantwortlichen bei der Umsetzung ihrer Ziele zu helfen.

Die nachfolgenden Erläuterungen der einzelnen Punkte beziehen sich zwar schwerpunktmäßig auf die technische Industrie, können aber leicht verallgemeinert werden.

Datenreduktion und Handlungsempfehlung im Fehlerfall

Während im normalen Betrieb einer technischen Anlage die Anforderungen an den Operator aufgrund des hohen Automatisierungsgrades relativ gering sind, wird er im Fehlerfall oftmals überfordert. Da ein Fehler oft weitere Fehler verursacht, kommt es im Störfall zu einer Kettenreaktion und zu einem Meldeschwall in den Meldefolgenanzeigen. Dem Operator ist es oftmals nicht möglich, aus der Vielzahl von Fehlermeldungen die tatsächliche Ursache zu erkennen. Damit kann er gerade in den kritischen Situationen keine Gegenmaßnahmen ergreifen.

Das Ziel einer gleichzeitigen und parallelen Auswertung Hunderter Messgrößen und die Auswertung ihrer Wechselbeziehungen untereinander ist gegenwärtig zwar möglich, aber mit sehr hohem Engineeringaufwand verbunden. Insbesondere dann, wenn sich abhängige und unabhängige Variablen nicht identifizieren lassen und auch unabhängige Variablen untereinander in Wechselwirkung stehen, wird die Auswertung problematisch.

Prozessvisualisierung am Beispiel einer Industrieanlage

Die Prozessvisualisierung besitzt eine besondere Bedeutung, denn sie ist die Schnittstelle zum Prozess. Alle Informationen über den Prozess, alle Störungen, alle Bedienungen erfolgen über die Schnittstelle des sogenannten Human Machine Interface, kurz HMI. Hierbei wird weiterhin zwischen HMI für die Prozessbedienung und HMI für die Prozessanalyse unterschieden. Das HMI der Prozessbedienung dient dem unmittelbaren Beobachten und Fahren des Prozesses. Die dafür notwendigen Prozesseingriffe werden gewöhnlich über dieses HMI durchgeführt.

Für das Interface zur Prozessbedienung gibt es mehrere Varianten, so z. B. Kombinationen von Monitoren direkt vor dem Bedienpersonal und Großbildschirmen im Hintergrund. Die Anlage wird dabei in Kreise, Funktionsgruppen, Bereiche und Blöcke eingeteilt. Verschiedene Anwahlmechanismen lassen ein schnelles Navigieren durch die verschiedenen Anlagenteile zu und ermöglichen dem Prozessbediener einen schnellen Überblick über den laufenden Prozess. Spezielle Übersichtsbilder, sogenannte Anlagenschemata oder Fließbilder, zeigen die wichtigsten Prozessgrößen in einem einzigen Fließbild. Im Normalfall sind damit alle Anforderungen für eine einfache Prozessbedienung abgedeckt.

Die Prozessfehler werden in Alarmhierarchien verschiedener Prioritäten zusammengefasst und in Meldefolgenanzeigen visualisiert. Im Allgemeinen ist es über Online-Suchfunktionen möglich, direkt von einer solchen Fehlermeldung zu den jeweiligen Kreis- oder Funktionsgruppenbildern zu gelangen, um die Ursache für einen etwaigen Fehler schnell zu identifizieren. Problematisch ist allerdings, dass gerade während einer tatsächlichen Störung im Prozess eine Vielzahl von Meldungen generiert wird und der Operator im Fall eines solchen Meldungsschwalles oftmals nicht mehr die Zeit hat, die tatsächlichen Ursachen für den aufgetretenen Fehler zu ermitteln. Deshalb werden heute große Anstrengungen unternommen, um Fehlerfrüherkennungssysteme zu entwickeln, die die Reaktionszeit zwischen Fehlererkennung und Auswirkung im Prozess verlängern. Andere Konzepte versuchen, eine intelligentere Verdichtung der Alarmmeldungen vorzunehmen, um schneller Ursachen ermitteln zu können.

Eine weitere Verbesserung in der Prozessbeobachtung wurde schon vor 25 Jahren durch sogenannte Massendatendisplays vorgeschlagen, **[ZEH94]** und **[ZIN93]**, sie haben sich aber nicht durchgesetzt. Diese Displays stellen schematisch den Prozess als Ganzes dar. Ihre Besonderheit liegt darin, dass nicht die einzelnen numerischen Werte der Prozessgrößen visualisiert werden, sondern stets die Differenz Δy_i zwischen dem Soll- und dem Istwert $\Delta y_i = y_i^{SOLL} - y_i^{IST}$ einer jeden Prozessgröße y_i. Diese Differenz Δy_i wird z. B. als Auslenkung einer „Waage“ α_i kodiert dargestellt. Ist die Differenz null, befindet sich die Waage im Gleichgewicht; sind Sollwert und Istwert unterschiedlich, schlägt die Waage aus.

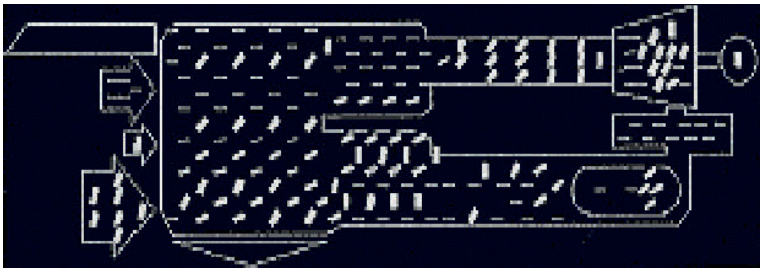


Bild 2.9 Massendatendisplay aus **[ZIN93]** zur Darstellung eines aktuellen Kraftwerkszustandes

In Bild 2.9 befinden sich keine numerischen Werte, sondern viele kleine Waagen α_i, die den jeweiligen Prozessabschnitten zugeordnet wurden. Damit sieht ein Bediener auf einen Blick, ob sein Gesamtprozess im Gleichgewicht ist, denn sind alle Waagen waagerecht, stimmen alle Istwerte mit ihren Sollwerten überein. Ein derartiges Massendatendisplay wäre eine der modernsten Visualisierungsformen. Allerdings hat es zwei entscheidende Nachteile:

- erstens gibt es keine Auskunft über die Vergangenheit des Prozesses, sondern immer nur über den aktuellen Prozesszustand, und
- zweitens bedeutet die Vorgabe von Sollwerten einen erhöhten Engineeringaufwand, denn es muss für jeden möglichen Prozesszustand der dazugehörige Sollwert y^{SOLL} jeder zu beobachtenden Prozessvariablen vorab ermittelt werden.

Der Aufbau der für diese Anzeige notwendigen Referenzmodelle ist deshalb sehr aufwendig. Insbesondere in den transienten Übergangsphasen, d. h. in den Phasen, in denen die Anlage von einem Zustand in einen anderen Zustand wechselt (zum Beispiel bei Produktänderungen), sind oftmals keine Referenzmodelle verfügbar. Deshalb haben Massendatendisplays keine große Verbreitung in der Industrie gefunden.

Auch andere Verfahren wie Vektordiagramme, Trenddisplays und Fisheyes geben nur einen Gesamtüberblick über den aktuellen Prozesszustand und dessen augenblickliche Änderungen. Verfahren, die einen Gesamtüberblick über den aktuellen *und* vergangenen Prozesszustand visualisieren können, gibt es gegenwärtig nicht. Eine Analyse komplexer Vorgänge (Transienten) kann damit nicht durchgeführt werden.

Prozessmodellierung

Das Ziel jeder Modellierung ist es, den Zusammenhang f zwischen n Eingangsgrößen x_i und k Ausgangsgrößen y_j zu ermitteln. Dieser Zusammenhang kann linear oder nichtlinear, statisch oder dynamisch sein.

Im Bereich der Prozessmodellierung bzw. der Modellierung von Prozessstrecken sind gegenwärtig verschiedene Verfahren im Einsatz:

- Theoretische Modellbildung (z. B. Verwendung von Differentialgleichungen)
- Experimentelle Modellbildung (z. B. Experimentelle Aufnahme von Kennlinien, Sprungfunktionen, Stoßfunktionen), halb-empirische Modellbildung
- Empirische Modellbildung (z. B. Verwendung von neuronalen Netzen)

Es ist gegenwärtig üblich, im Bereich Prozessmodellierung arbeitspunktabhängige, lineare Regressionen auf Signalebene durchzuführen. Bei ihrer Anwendung gibt es aber in den transienten Übergangsphasen Genauigkeitsprobleme, da die Signale in diesen Phasen gegenwärtig nur mit ca. 80 % Genauigkeit modelliert werden können.

Obwohl die klassischen Methoden der theoretischen Modellierung einige Vorteile haben (Verstehbarkeit, Erklärbarkeit, generelle Gültigkeit), geht der Trend hier seit längerem zu einer verstärkten Nutzung von induktiven Techniken. Die Modellierung komplexer, dynamischer Zustände ist allerdings noch nicht befriedigend gelöst. Bedeutende Ansätze waren vor einigen Jahren Expertensysteme, die sich aber aufgrund eines hohen Engineeringaufwandes nicht übermäßig stark verbreitet haben. Allerdings kann man mittels symbolischer Lerntechniken, wie Entscheidungsbäumen oder Assoziationsregeln (Kapitel 3), derartige Aufwände auf ein Minimum begrenzen. Auch neuronale Netze als Vertreter sub-symbolischer Lerntechniken werden zunehmend eingesetzt.

Die Modellierung stellt die Basis für zahlreiche nachgeschaltete Auswerteverfahren dar. Insbesondere wird sie gegenwärtig für folgende Aufgaben im industriellen Bereich genutzt:

- Auslegungsrechnung zur Anlagenplanung
- Aufbau von Referenzmodellen für Komponenten
- Aufbau von Referenzmodellen für Prozesssignale
- Fehlerdetektion im Zusammenhang mit Diagnoseapplikationen
- Datenvalidierung
- Simulation des Prozesses (i. Allg. nur stationär möglich)
- What-if-Analyse (i. Allg. nur stationär möglich)
- Sensitivitätsanalyse für wichtige Eingangsgrößen

Prozessanalyse und Prozessüberwachung

Im Unterschied zur Prozessüberwachung, die online erfolgt, werden bei der Prozessanalyse die Daten auch offline ausgewertet. Das Ziel beider Methoden besteht in der Detektion und Analyse von Fehlern und in der Optimierung der Fahrweise einer Anlage. Dabei kann zwischen signalorientierten und zustandsorientierten Ansätzen unterschieden werden.

Beim *signalorientierten Ansatz* werden für jedes wichtige Prozesssignal Grenzwerte festgelegt und der aktuelle Wert mit dem Grenzwert verglichen. Verletzt der Istwert y^{IST} einen dieser vorgegebenen Grenzwerte y^{u} oder y^{o}, wird durch das Leitsystem eine Fehlermeldung generiert, die in die Alarmhierarchien eingeht und in der Meldefolgenanzeige visualisiert werden kann.

Problematisch bei diesem Ansatz ist, dass ein zu enges Einstellen der Alarmgrenzen (y^{u}, y^{o}) dazu führen kann, dass eine Vielzahl von Fehlalarmen ausgelöst wird. Denn gerade in den transienten Phasen – während des Einschwingens auf einen neuen stationären Prozesszustand – kann es zu einem Überschwingen von einzelnen Prozesssignalen kommen. Dieses Verhalten ist aber ein normales Prozessverhalten und kein Systemfehler. Sollen diese Fehlalarme verhindert werden, lässt

sich das mit statischen Alarmgrenzen nur dann realisieren, wenn die Grenzen sehr weit eingestellt sind. Bei diesem Vorgehen können allerdings tatsächliche Alarme übersehen werden.

Um nur im Fehlerfalle eine Meldung zu generieren, werden in modernen Beobachtungssystemen Toleranzbänder mit dynamischer Breite ΔTH verwendet, die um den Gutverlauf (Sollverlauf) des Referenzsignales gelegt werden, siehe Bild 2.10. In den stationären Phasen ist dieses Toleranzband eng (ΔTH ist klein), um schon kleinste Abweichungen des Signals von seinem Sollverlauf zu detektieren. In den transienten Übergängen wird das Toleranzband breit um den Sollverlauf eingestellt (ΔTH ist groß), um ein normales Flattern der Signale nicht schon als Grenzwertverletzung zu melden.

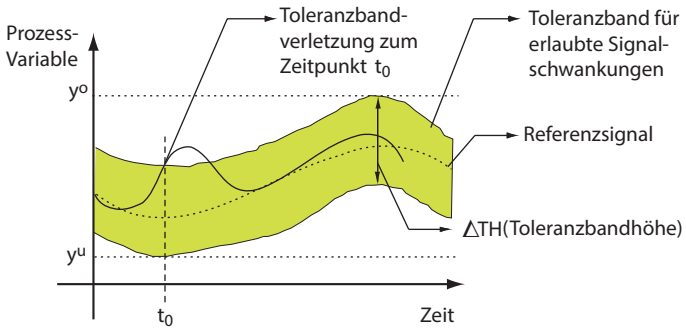


Bild 2.10 Verwendung von Toleranzbändern für Fehlererkennung

Beim zustandsorientierten Ansatz werden alle relevanten Signale gleichzeitig betrachtet und ausgehend von diesen Signalen ein Prozesszustand Z^{IST} ermittelt. Dieser Ist-Prozesszustand wird dann mit einem Soll-Prozesszustand Z^{SOLL} verglichen und Abweichungen werden einem Diagnosesystem gemeldet. Dabei wird ein Prozesszustand Z durch die aktuellen Werte seiner k Prozessgrößen beschrieben.

Der Unterschied zwischen den Ansätzen besteht darin, dass beim zustandsorientierten Ansatz die Wechselwirkung der Signale mitberücksichtigt wird, während bei den signalorientierten Methoden jedes Signal für sich allein beobachtet werden muss. Wird ein Prozesszustand Z' betrachtet, der sich vom Prozesszustand Z dadurch unterscheidet, dass sich eine Vielzahl von Prozessvariablen von ihrem Sollwert entfernt hat, so kann mit einem signalorientierten Verfahren der Zustand Z' nicht von Z unterschieden werden, wenn alle Prozesssignale noch innerhalb ihrer statischen oder dynamischen Grenzwerte liegen. Mit einem zustandsorientierten Ansatz würde sich diese Unterscheidung erkennen lassen.

Eine zusammenfassende Bewertung der verwendeten Prozessanalysetechniken zeigt, dass die gängigen Werkzeuge ein großes Leistungsspektrum hinsichtlich der Analyse und Überwachung eines technischen Prozesses besitzen. Allerdings sind

die Engineeringaufwände für die meisten Applikationen sehr hoch. Auch der Einsatz moderner Verfahren, wie z. B. neuronaler Netze zur Erkennung von Fehlern, hat sich als problematisch herausgestellt, da es in vielen Anwendungsfällen einen permanenten Mangel an Fehlerfällen gibt. Aus diesem Grund lässt sich kein Klassifikator für die Fehlererkennung realisieren, da dieser nur unter hohem Aufwand trainiert werden könnte. Diese Aufwände lassen sich aber kommerziell nicht rechtfertigen.

Detaillierte Analysen von komplexen transienten Prozessvorgängen (Anfahren, Abfahren, Lastwechsel) werden praktisch nicht durchgeführt. Alle komplexen Prozessauswertungen beziehen sich in der Regel auf stationäre Prozesszustände einer Anlage. Des Weiteren werden keine systematischen Sensitivitätsanalysen und Varianzanalysen durchgeführt.

Prozessprognose

Unter einer Prognose wird hier die zeitliche Extrapolation von Signalen oder Zuständen verstanden, die das Ziel hat, frühzeitig Änderungen im Prozess zu erkennen, um durch geeignete Gegenmaßnahmen darauf reagieren zu können. Die im Industriebereich eingesetzten Extrapolationsverfahren basieren meistens auf linearen Modellen oder Ansätzen mit neuronalen Netzen. Eine Extrapolation des gesamten Prozesses einer Anlage wird wegen hoher Engineeringaufwände und schlechter Prognosequalität noch zu selten durchgeführt.

Prozessoptimierung

Im Bereich der Prozessoptimierung muss zwischen den verschiedenen Prozessebenen unterschieden werden: Optimierung auf unterer oder oberer Prozessebene.

Für die Optimierung auf einer prozessnahen, unteren Ebene wird eine Vielzahl moderner Regelungsverfahren angewendet, um bessere und robustere Regelstrategien aufbauen zu können. Auf oberster Prozessebene sind zum Beispiel sogenannte *Plant-Management-Systeme* im Einsatz.

Es ist gegenwärtig jedoch nicht üblich, aus bekannten und gewünschten Ausgangsgrößen eines Prozesses auf die verursachenden Eingangsgrößen zu schließen. Insbesondere wird aus den zulässigen Schwankungen (Varianzen) einer Zielgröße noch nicht auf die zulässigen Schwankungen (Varianzen) aller Einflussgrößen geschlossen. Prozessoptimierungen dieser Art sind sehr selten anzutreffen.

Engineeringaufwände für Analysesysteme

Eines der größten Probleme vieler Analyseverfahren im Prozessbereich ist der hohe Engineeringaufwand aller zugrundeliegenden Anwendungen. Diese Aufwände bestehen im Anpassen der Softwareprodukte an die konkreten Erfordernisse einer jeden Anlage. Sie können die Lizenzkosten der Software bei weitem übersteigen.

Durch diese hohen Engineeringaufwände, die schwerpunktmäßig im Einstellen (Tuning) des Modells an die konkrete Anlage liegen, sind viele der o. g. Anwendungen wirtschaftlich nicht immer vertretbar. Eine Reduzierung der Engineeringaufwände für bestehende oder für neu zu entwickelnde Verfahren ist deshalb seit Jahren unumgänglich.

2.3 Gibt es einen Handlungsbedarf für die Industrie?

Die Untersuchung der gegenwärtig verwendeten Datenanalyseverfahren im Bereich der technischen Industrie kann trotz vieler positiver Ergebnisse in folgender Situationsbeschreibung zusammengefasst werden:

- Es herrscht eine signalorientierte Prozessanalyse (Alarmgrenzen, Toleranzbänder) in den eingeschwungenen Prozesszuständen vor.
- Komplexere Überwachungssysteme (Massendatendisplays, Expertensysteme, Performancetools) sind meistens nur für stationäre Zustände im Einsatz.
- Die Überwachungsverfahren von transienten Prozesszuständen sind unzureichend.
- Es gibt keine befriedigende Überwachung, Auswertung und Gegenüberstellung komplexer Prozessvorgänge.
- Es werden selten Sensitivitätsanalysen durchgeführt.
- Es werden selten ganzheitliche Prozessoptimierungen durchgeführt.
- Fast alle verwendeten Analyseverfahren haben hohe bis sehr hohe Engineeringaufwände.

Diese Einschätzung erhebt nicht den Anspruch auf Vollständigkeit, im Einzelfall können permanente Weiterentwicklungen zu einer Aufhebung der aufgelisteten Einschränkungen führen. Dennoch zeigt sie, wo im gegenwärtigen industriellen Umfeld noch Potenziale für die Datenanalyse zu finden sind.

Werden die aktuellen Hauptaufgaben im Prozessbereich, wie Erhöhung der Anlagenverfügbarkeit, Verringerung der Störungen, Erhöhung des Anlagenwirkungsgrades, Verringerung der Produktionskosten und Reduzierung der Engineeringkosten den gegenwärtigen Möglichkeiten von Data Mining gegenübergestellt, so ergeben sich neben zahleichen Einzellösungen folgende systematische Aufgaben für die Data-Mining-Anwendungen in der Industrie, siehe Tabelle 2.2:

Tabelle 2.2 Aufgaben für Data Mining in der Prozessindustrie

	Querschnittsaufgaben für Data-Mining-Verfahren in der Industrie (insbesondere Prozessindustrie)
1	Ganzheitliche Prozessüberwachung
2	Diagnosen (Detektionen, Lokalisation, Fehler-Ursachenbestimmung)
3	Analyse und Vergleich komplexer Vorgänge und Transienten
4	Automatische Generierung von Regeln und Hypothesen aus Prozessdaten
5	Prognose von komplexen Prozesszuständen
6	Vektorielle Optimierung von Prozessen
7	Reduzierung der Engineeringkosten

Zukünftige Data-Mining-Verfahren sollten damit in den genannten Bereichen der Tabelle 2.2 eingesetzt werden.

Bild 2.11 verdeutlicht noch einmal, dass es eine Reihe von Aufgaben im Industriebereich gibt, die mit klassischen (a) oder mit modernen Analysetechniken zumindest theoretisch gelöst werden können (b). Allerdings gibt es auch zahlreiche Problemstellungen, die mit diesen Techniken aktuell nur unzureichend lösbar sind (c) und (d).

Data Mining und Big Data unterstützen gerade die Lösung dieser Probleme und werden deshalb zwangsläufig zu einer noch breiteren Anwendung in Industrie und Gesellschaft führen. Es lohnt daher, sich mit den Grundlagen der Data-Mining-Technologien vertraut zu machen.

Verfolgt man die Geschichte der Innovationen, so erkennt man, dass die Gesellschaft immer dann Lösungen schafft, wenn sie dringend gebraucht werden. Die aktuell beobachtbare Verstärkung von Data-Mining- und Big-Data-Anwendungen in allen Bereichen von Industrie und Gesellschaft ist damit ein gesetzmäßiger, weiterer Schritt auf dem Wege der „Digitalisierung“ unserer Lebenswelt.

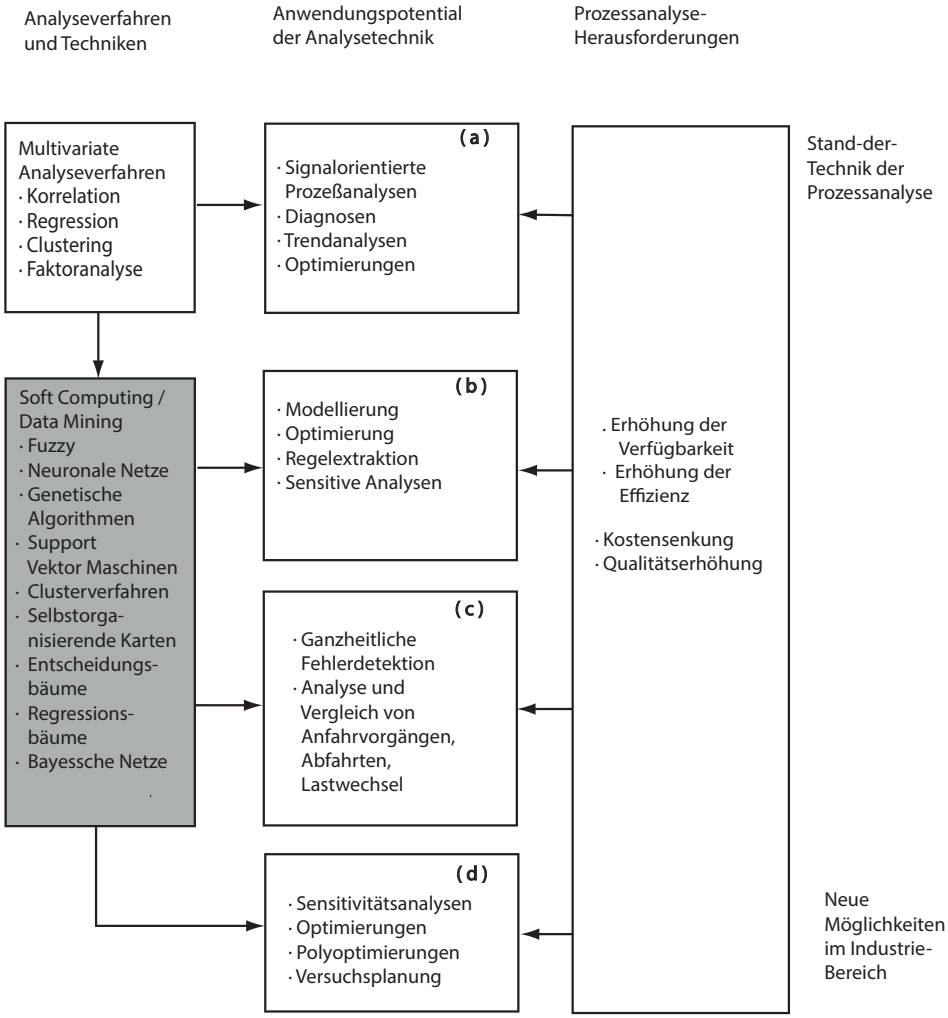


Bild 2.11 Gegenüberstellung von Analyseverfahren und Herausforderungen in der Industrie

3 Das theoretische und mathematische Konzept der technischen Datenauswertung

3.1 Einführung

Obwohl der Data-Mining-Prozess als Kern jeder Big-Data-Auswertung im Detail sehr komplex ist, war man von Anfang an bestrebt, die Arbeitsschritte, die in jedem Data-Mining-Projekt durchgeführt werden müssen, zu verallgemeinern und zu standardisieren.

Bereits vor über 20 Jahren wurde von den Firmen NCR *Systems Engineering* (USA, Dänemark), *DaimlerChrysler AG* (Deutschland), *SPSS* Inc. (USA) und der *OHRA Verzekeringen en Bank Groep B.V.* (Niederlande) ein einheitlicher Standard für Data-Mining-Prozesse, der sogenannte CRISP-Standard, geschaffen, siehe [CRI00]. Dieser CRISP-Standard, an dem seit 1996 gearbeitet wurde und der für *Cross Industry Standard Process for Data Mining* steht, hat sich mittlerweile als der industrielle Standard durchgesetzt, wenngleich nicht jeder der großen Data-Mining- und Big-Data-Anbieter diesen Namen benutzt.

Die CRISP-Methodologie beschreibt einen hierarchischen Prozess. Auf dem ersten bzw. obersten Level ist der Data-Mining-Prozess in abstrakten Phasen organisiert (z. B. Datenvorverarbeitung). Der zweite Level ist der generische Level, um alle möglichen Data-Mining-Situationen und Applikationen berücksichtigen zu können (z. B. Lückenbefüllung in Daten). Der dritte Level beschreibt die speziellen Aufgaben in speziellen Situationen (z. B. Lückenbefüllung in numerischen Datentabellen). Der vierte Level definiert die Aktionen, Entscheidungen und möglichen Resultate.

Insbesondere ist in [CRI00] ein Referenzmodell eingeführt, das die zeitliche Abfolge eines Data-Mining-Projektes spezifiziert. Dieses Referenzmodell besteht im Wesentlichen aus den folgenden zeitlichen Phasen:

1. Business Understanding (Aufgaben- und Prozessverständnis)
2. Data Understanding (Datensichtung und Datenverständnis)
3. Data Preparation (Datenvorverarbeitung und Transformation)