Originalarbeit

BHM (2019) Vol. 164 (1): 26–30
https://doi.org/10.1007/s00501-018-0803-z

BHM Berg- und Hüttenmännische Monatshefte

Herausforderungen beim Einsatz von Datenanalytik für eine ressourceneffiziente Produktion

Katharina Mertens

Lehrstuhl für Wirtschafts- und Betriebswissenschaften, Montanuniversität Leoben, Leoben, Österreich

Eingegangen 6. November 2018; angenommen 20. November 2018; online publiziert 10. Januar 2019

Zusammenfassung: Durch Industrie 4.0 wächst die Möglichkeit der Vernetzung und der dadurch erzeugten und speicherbaren Datenmengen stetig an. Datenanalytik als neues Trend-Thema soll aus diesen Datenmengen Informationen bzw. Wissen generieren und im besten Falle zu einer Optimierung von Produktionsprozessen und Materialeinsatz führen. In diesem Artikel werden die Grundlagen der Datenanalytik erläutert sowie zwei Case Studies, die im Umfeld der Chargenfertigung bzw. Unikatfertigung durchgeführt wurden, vorgestellt. Abschließend werden die Herausforderungen in den Projekten und Lösungsansätze präsentiert.

Schlüsselwörter: Datenanalytik, Case Studies, Chargenfertigung, Unikatfertigung, Datenqualität

Challenges in Using Data Analytics for Resource-efficient Production

Abstract: The amount of data produces as well as the possibility for connecting multiple data sources constantly increase in the context of Industry 4.0. Data analysis as a new trend topic should generate information or knowledge from these data sets and, in the best case, lead to an optimization of production processes and material usage. In this article, the basics of data analysis are explained and two case studies carried out in the field of batch production or unique production are shown. Finally, the challenges encountered in the projects and approaches to solutions are presented.

Keywords: Data analytics, Case studies, Batch production, One-of-a-kind production, Data quality

DI K. Mertens, BSc (✉)
Lehrstuhl für Wirtschafts- und Betriebswissenschaften,
Montanuniversität Leoben,
Franz Josef-Straße 18,
8700 Leoben, Österreich
katharina.mertens@unileoben.ac.at

1. Einleitung

Die datengestützte Produktion ist in der 4. industriellen Revolution auf eine hohe Vernetzung und zielorientierte Auswertung von Daten zurückzuführen [1]. Dies eröffnet neue Möglichkeiten zur effektiven Datenverarbeitung und -analyse, um Optimierungspotenziale aufzuzeigen. Dadurch erhält die Datenanalytik einen immer höheren Stellenwert im Bereich der Produktion [2].

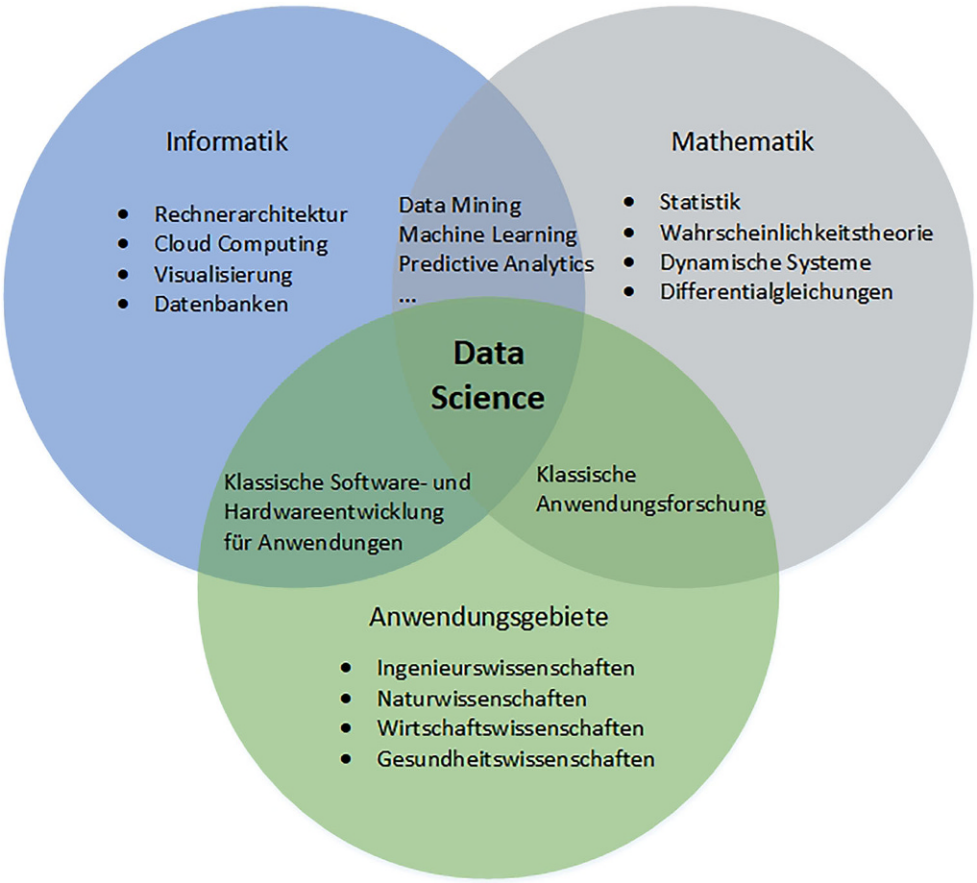
Um einen Einblick in die Thematik Datenanalytik zu erhalten, wird ein Überblick über die Grundlagen gegeben und zwei Case Studies werden kurz vorgestellt. Anschließend werden die Herausforderungen, die in diesen Projekten zu Tage traten, thematisiert und eine Handlungsempfehlung abgegeben.

2. Grundlagen der Datenanalytik

Seit 40 Jahren existiert der Begriff Data Science. Es gibt der Kombination von Mathematik, Informatik und der jeweiligen Anwendungsdomäne einen Überbegriff (siehe Abb. 1; [2, 3]). Durch eben diese Anwendungsgebiete werden neue Potenziale für die Datenanalytik erschlossen und bestehende Ansätze verbessert. Es wird untersucht, in wie weit die Transparenz von Prozessen durch Daten erhöht werden kann und diese als Basis für Entscheidungen zu Planung und Steuerung der Prozesse dienen können [2].

Die vielfältigen Methoden der Datenanalytik können unterschiedlich klassifiziert werden. Das Unternehmen IBM zieht dazu die Eigenschaften der Daten heran – Geschwindigkeit (Velocity), Menge (Volume), Vielfalt (Variety) und Richtigkeit der Daten (Veracity) [4] –, wohingegen in Abb. 2 die Herangehensweise an die Problemstellung und das Ziel der Anwendung zur Unterscheidung dienen. Die Herangehensweise kann in modellgetrieben, d. h. Einsatz der klassischen Statistik – mit einem zugrundeliegenden Modell, und datengetrieben, d. h. Data Mining und maschinelles Lernen – mit einem unbekannten Prozess im Hintergrund, aufgeteilt werden [5]. Beschreibung (Deskription), Erforschung (Exploration), Erklärung (Diagnose) und Prognose bilden

Abb. 1: Komponenten von Data Science [2, 3]



die Ziele der Anwendung, die durch die klassische Statistik, das Data Mining und das maschinellen Lernen erreicht werden können [6].

Auf sämtliche Daten, die in Produktionssystemen anfallen, können diese Ansätze angewendet werden. Durch die Wahl der geeigneten Methoden, wie Ausreißererkennung, Assoziation, Clustering (Segmentierung), Klassifikation, Prognose und Regression, ist es möglich, ein klares Verständnis des Systems und des Optimierungspotenzials zu schaffen [2].

Um dieses Verständnis zu erreichen und die geeignete Methode zu finden, ist ein strukturiertes Vorgehen notwendig. Projekte im Bereich der Datenanalytik orientieren sich an verschiedenen Rahmenwerken [8]. Eines davon ist der CRISP-DM, der Cross Industry Standard Process for Data Mining. Das umfassende Modell besteht aus den sechs Phasen:

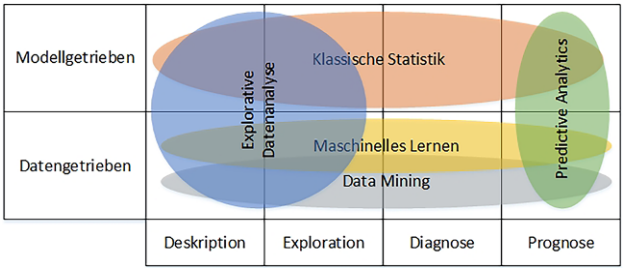


Abb. 2: Qualitative Unterscheidung verschiedener Ansätze zur Datenanalyse [2]

1. Geschäftsmodell verstehen,
2. Daten verstehen, Daten aufbereiten,
3. Modellierung,
4. Evaluation und
5. Einsatz.

Abb. 3 stellt diese und die Abhängigkeiten zu einander dar [7]. Dieses Rahmenwerk bildet die Grundlage für die beiden vorgestellten Case Studies, die mit Hilfe von datenanalytischen Methoden bearbeitet worden sind.

3. Case Studies

Bei beiden Case Studies wurde die Zielsetzung zu Beginn vom verantwortlichen Projektteam seitens der Industrie festgelegt. Es wurde ein rein datengetriebener Ansatz verfolgt, da entweder ein modellgetriebener Ansatz aufgrund zu vieler Einflussparameter nicht möglich war (siehe 3.1) oder die Ergebnisse des Modells durch die Daten validiert werden sollten (siehe 3.2).

Der CRISP-DM wurde bei beiden Case Studies als Grundlage verwendet, wobei beide nur bis zum Schritt Evaluation durchgeführt wurden. Es sind verschiedene Hindernis-

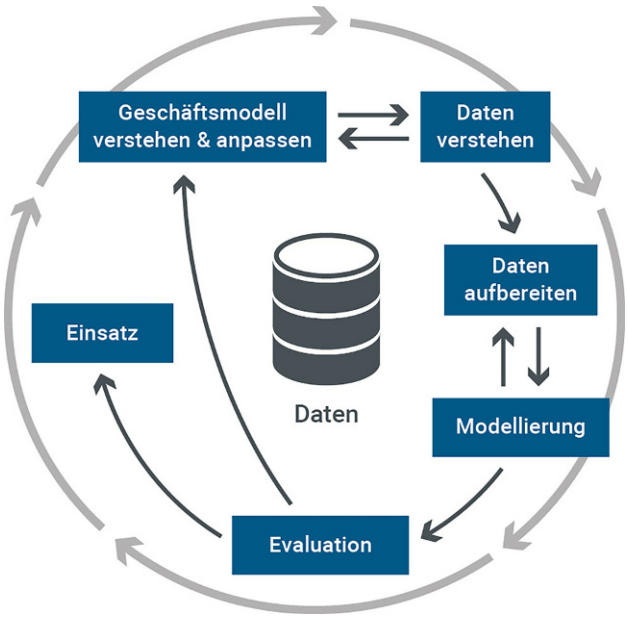


Abb. 3: CRISP-DM [7]

se bzw. Herausforderungen aufgetreten, die in Abschn. 4 erläutert werden.

3.1 Chargenfertigung in der Chemischen Industrie

Im Rahmen einer Chargenfertigung in der chemischen Industrie gibt es Prozesse, die unter sich verändernden Bedingungen ablaufen. Sie benötigen oftmals Referenzwerte zur Adjustierung der Prozessbedingungen. Diese Referenzwerte werden unter Realbedingungen auf der jeweiligen Anlage ermittelt. Allerdings führen diese Testläufe zu einer schlechten Auslastung der Anlage und – bei zeitaufwendigen Prozessen – zu einer Verlängerung der Durchlaufzeit sowie zu höheren Kosten. Zusätzlich kann es sich um zerstörende Prüfungen handeln und es kann zu einem Ausschuss der entstehenden Produkte kommen, die nicht innerhalb der Toleranzgrenzen liegen.

Daher war das Ziel dieses Projekts, diesen Referenzprozess durch Methoden der Datenanalytik zu umgehen. Da keine klassische Modellierung des Systems möglich war, wurde ein datengetriebener Ansatz mit einem Klassifikationsmodell, das auf Daten aus den vorher stattfindenden Prozessen basiert, verfolgt. Dieser Ansatz ist in Abb. 4 zu sehen, wobei die Phasen des CRISP-DM ebenfalls eingetragen wurden.

So wurden im ersten Schritt die Rahmenbedingungen der Produktion betrachtet und Daten erhoben. Mehrere Datensets unterschiedlicher Herkunft und unterschiedlicher Formate lagen am Ende vor. Die Datentypen – wie metrische und nicht-metrische Daten unterschiedlicher Skalenniveaus –, die verschiedene Beschränkungen hinsichtlich statistischen aber auch mathematischen Operationen haben, unterscheiden sich auch in ihrer Aussagekraft.

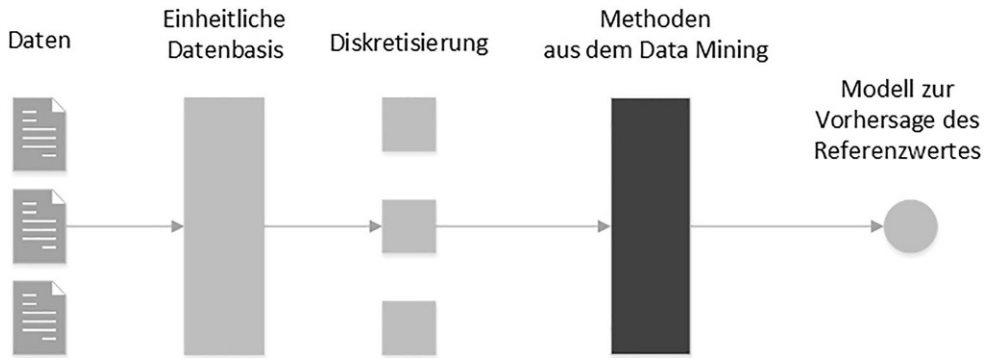
Das Verständnis dieser Daten war essentiell, um diese anschließend aufbereiten und eine einheitliche Datenbasis schaffen zu können. So waren Skalenniveaus vor allem für die Zusammenführung der Datensets wichtig, um die Bedeutung nicht zu verfälschen. Als Grundlage für die Zusammenführung wurde abhängig von den Skalenniveaus und den Aussagen der Daten Regeln definiert. Zusätzlich mussten unvollständige Datensätze ausgeschlossen oder aufgefüllt werden. Innerhalb eines Attributs, als ein Merkmal des Prozesses, wurden Ausreißeranalysen durchgeführt, die zu einer Veränderung der Datenbasis führen, aber auch eine Verbesserung zur Folge haben können. Das ursprüngliche und das von Ausreißern bereinigte Datenset dienten als Grundlage für die Diskretisierung, eine Einteilung der Daten anhand der jeweiligen Referenzwerte – dem Label. Dies war notwendig, dass die Klassifikationsalgorithmen, wie u. a. Decision Tree (Entscheidungsbaum), Support Vector Machines und Esemblemethoden, auf beide Datensets angewendet werden konnten.

Die unterschiedlichen Methoden lieferten unterschiedliche Ergebnisse, auch die unterschiedlichen Datensets beeinflussten die Vorhersagegenauigkeit der Modelle, die nicht zufriedenstellend war. Ein abschließender Test mit bis dato unbekannten Daten hat das Ergebnis erneut verschlechtert.

3.2 Unikatfertigung in der Schwerindustrie

Die Unikatfertigung birgt viele Herausforderungen, da jedes hergestellte Produkt eine komplexe Einzelanfertigung mit hohen Auftragszeiten und einer Unsicherheit aufgrund der Einmaligkeit ist. So ist manchmal eine Reproduktion

Abb. 4: Modell zur Vorhersage des Referenzwertes



nicht oder nur eingeschränkt möglich [9]. Daher sind datenanalytische Verfahren nur eingeschränkt sinnvoll oder nutzbar, da diese auf großen Datenmengen, die unter gleichen Bedingungen erfasst werden, basieren. In dieser Case Study lag daher der Fokus auf einer datengetriebenen Validierung des modellgetriebenen Ansatzes. Welche Einflussfaktoren (Attribute) haben den größten Einfluss auf den Zielwert? Hierbei handelte es sich ebenfalls um eine Qualitätskennzahl, die den Umfang der Nacharbeiten bestimmt. Eine bessere Kontrolle der Einflüsse könnte den Umfang der Nacharbeiten und somit Kosten minimieren.

Die Datenbasis wurde bereitgestellt. Durch mehrere Schleifen durch den CRISP-DM in Absprache mit dem Projektteam wurde eine neue Datenbasis geschaffen. So wurde u. a. nach dem Aspekt Produktart unterschieden, um eine homogenere Datenbasis zu erhalten und Ausreißer zwischen Produktarten ausschließen zu können. Zusätzlich wurden durch die subjektive Einschätzung des Projektteams redundante oder voneinander abhängige Informationen ausgeschlossen. Eine Schleife lief vom Datenverständnis hin über die Datenaufbereitung hin zur Modellierung und Evaluation im CRISP-DM. Nach einer Diskretisierung des Labels wurde jeweils eine Kombination von Verfahren zur Gewichtung der einzelnen Attribute eingesetzt. Dazu zählen unter anderem die Gewichtung durch die Korrelation und den Information Gain. Das jeweilige Ergebnis wurde vom Projektteam bewertet. Um die Methodik bzw. die Erkenntnisse zu validieren, wurden zusätzlich innerhalb einer Produktart zwei Datensets durch Trennung erzeugt, deren Ergebnisse miteinander verglichen wurden. Hier kam es nur zu wenigen Übereinstimmungen. Daher kann nicht gesagt werden, ob Ergebnisse, die mit der Expertise bzw. dem modellgetriebenen Ansatz übereinstimmen bzw. zufällig auftreten oder nicht.

4. Herausforderungen

Die erste Case Study, deren Umfang vor allem durch den Aufwand für das Verständnis und der Aufbereitung der Daten enorm war, kämpfte neben den Problemen der Datenbeschaffung bzw. der Datenvielfalt (unterschiedliche Datenquellen und -formate) auch mit der Datenqualität. Die Daten waren teilweise nicht konsistent oder nicht vorhanden, da keine Prüfung bei der Datenaufnahme durchgeführt wurde oder Messpunkte nicht eindeutig definiert waren. U. a. wurden zerstörende Prüfungen durchgeführt, was die Durchgängigkeit der Datenaufnahme gestört haben könnte. Innerhalb des Betrachtungszeitraums kam es zu einem Technologiesprung, was die Vergleichbarkeit erschwert hat.

Bei beiden Projekten – trotz unterschiedlicher Produktionsart und Industrie – wurde eine maßgebliche Ursache für ein enttäuschendes Ergebnis identifiziert: das Verhältnis zwischen Attributen und Datensätzen. In der Literatur werden Empfehlungen von 1:10 angegeben, wobei dies von der Homogenität der Stichprobe abhängig ist [10]. Bei beiden Case Studies wurde eine Diskretisierung vorgenommen, bei der eine zu große oder zu kleine Klassenanzahl das Ergebnis beeinflussen kann. Ein größerer Stichprobenumfang kann das Ergebnis ebenfalls beeinflussen. In der zweiten Case Study wurde stark auf die Expertise des Projektteams gesetzt, was das Ergebnis verfälscht haben könnte, u. U. da hier auch bereits mit einem aufbereiteten Datenset gearbeitet wurde.

5. Zusammenfassung und Ausblick

Werden die Herausforderungen zusammengefasst, so ist ein Erfolg oder Misserfolg vor allem von der Datenbasis abhängig, die auch die Methodenwahl bestimmt und somit die Möglichkeit, die Zielsetzung zu erreichen, stark beeinflusst. In Zukunft gilt es daher, ein vollständiges Assessment der Datenreife, wie es Bernerstätter [11] vorschlägt, durchzuführen, damit die oben genannten Herausforderungen im Bereich der Datenquellen, -formate und -qualität frühzeitig erkannt werden und die Datenbasis realistisch eingeschätzt werden kann. Dieses Assessment, das im zweiten Schritt des CRISP-DM stattfinden sollte, wird so zum Grundstein für die weiteren Möglichkeiten im Projekt.

Die Größe des Stichprobenumfangs hängt ebenso von der Datenqualität ab, wobei ein Mindestmaß der in der Literatur angegebenen Empfehlung des Verhältnisses von Attribut zu Datensätzen von 1:10 [10] eingehalten werden sollte.

Funding. Open access funding provided by Montanuniversität Leoben.

Literatur

1. Bauernhansl, T.: Die Vierte Industrielle Revolution – Der Weg in ein wertschaffendes Produktionsparadigma, in: Bauernhansl, T.; ten Hompel, M.; Vogel-Heuser, B. (Hrsg.): Industrie 4.0 in Produktion, Automatisierung und Logistik, Wiesbaden: Springer Fachmedien Wiesbaden, 2014
2. Freitag, M.; Kück, M.; Ait Alla, A.; Lütjen, M.: Potentiale von Data Science in Produktion und Logistik: Teil 1 – Eine Einführung in aktuelle Ansätze der Data Science, Industrie 4.0 Management, 31 (2015), Nr. 5, S. 22–26
3. Schutt, R.; O'Neil, C.: Doing Data Science, 1. Auflage, Beijing (u. a.): O'Reilly Media, 2013
4. IBM Big Data & Analytic Hub, http://www.ibmbigdatahub.com/infographic/four-vs-big-data (31.10.2018)
5. Breiman, L.: Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author), Statistical science, 16 (2001), Nr. 3, S. 199–231
6. Fahrmeir, L.; Künstler, R.; Pigeot, I.; Tutz, G.: Statistik: Der Weg zur Datenanalyse. 5., Aufl, Berlin: Springer (Springer-Lehrbuch), 2004
7. Shearer, C.: The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5 (2000), Nr. 4, S. 13–22
8. Azevedo, A.; Santos, M. F: KDD, SEMMA and CRISP-DM: A Parallel Overview, In: Proceedings of IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, 2008, S. 182–185

9. Haux, M.; Friedewald, A.; Lödding, H.: Unsichere Auftragszeiten in der Unikatfertigung, ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb, 112.3 (2017), S. 129–132
10. Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R.: Multivariate Analysemethoden: Eine Anwendungsorientierte Einführung, 13., überarbeitete Auflage, Berlin (u. a.): Springer., (Springer-Lehrbuch), 2011
11. Bernerstätter, R.: Data Maturity for Smart Factory Applications – An Assessment Model, Acta Technica Corviniensis – Bulletin of Engineering, 1 (2018), S. 31–35

