Datenkompetenzen für die Massen - Muss Self-Service Data Mining scheitern?

Daniel Badura[1] und Michael Schulz[1,2]

[1] valantic Business Analytics GmbH, Deutschland
[2] NORDAKADEMIE Hochschule der Wirtschaft, Deutschland

Abstract. Data Mining ist ein Prozess, bei dem mittels statistischer Verfahren komplexe Muster in meist großen Mengen von Daten gesucht werden. Damit dieser von Organisationen verstärkt zur Entscheidungsunterstützung eingesetzt werden kann, wäre es hilfreich, wenn Domänenexperten durch Self-Service-Anwendungen in die Lage versetzt würden, diese Form der Analysen eigenständig durchzuführen, damit sie nicht mehr auf Datenwissenschaftler und IT-Fachkräfte angewiesen sind. In diesem Artikel soll eine Versuchsreihe vorgestellt werden, die eine Bewertung darüber ermöglicht, wie geeignet etablierte Data-Mining-Softwareplattformen (IBM SPSS Modeler, KNIME, RapidMiner und WEKA) sind, um sie Gelegenheitsanwendern zur Verfügung zu stellen. In den vorgestellten Versuchen sollen Entscheidungsbäume im Fokus stehen, eine besonders einfache Form von Algorithmen, die der Literatur und unserer Erfahrung nach am ehesten für die Nutzung in Self-Service-Data-Mining-Anwendungen geeignet sind. Dabei werden mithilfe eines einheitlichen Datensets auf den verschiedenen Plattformen Entscheidungsbäume für identische Zielvariablen konstruiert. Die Ergebnisse sind im Hinblick auf die Klassifikationsgenauigkeit zwar relativ ähnlich, die Komplexität der Modelle variiert jedoch. Aktuelle grafische Benutzeroberflächen lassen sich zwar auch ohne tiefgehende Kompetenzen in den Bereichen Informatik und Statistik bedienen, sie ersetzen aber nicht den Bedarf an datenwissenschaftlichen Kompetenzen, die besonders beim Schritt der Datenvorbereitung zum Einsatz kommen, welcher den größten Teil des Data-Mining-Prozesses ausmacht.

Keywords: Data Mining, Self-Service Analytics, Entscheidungsbäume, Datenkompetenzen.

1 Einleitung

In der Tradition von End User Computing (EUC) [41] existiert Self-Service Business Intelligence (SSBI) seit einigen Jahren als Ansatz zur Befähigung von Domänenexperten zur eigenständigen Durchführung auch komplexer Analysen [3, 10]. Diese Eigenständigkeit wird von einigen Praktikern und Forschern als die Emanzipation von IT-Fachkräften verstanden [49]. So wird es Anwendern beispielsweise ermöglicht, spezielle Visualisierungsformen anzuwenden oder sogar individuell Datenquellen in die

Analyseumgebung zu integrieren [27]. Andere gehen noch weiter und sehen in SSBI zusätzlich auch die Emanzipation von Datenwissenschaftlern[1] [22, 42].
Ein Grund für den Bedarf an mehr Eigenständigkeit ist der Umstand, dass immer mehr Daten verfügbar sind und als Entscheidungsgrundlage dienen können. Domänenexperten wollen diese Daten nutzen, ohne auf Spezialisten angewiesen zu sein. Deshalb werden einfach verständliche und zumindest teilweise automatisierte Möglichkeiten auch in komplexen Bereichen wie dem Data Mining[2] immer wichtiger [16, 49]. Viele Softwarehersteller haben dies erkannt und bieten Produkte mit sehr einfachen Benutzeroberflächen an, die die Erstellung von Analysen intuitiv zu machen versuchen [6, 14, 25, 26, 32, 40]. Andere bieten sogar die Möglichkeit der automatisierten Bildung von Modellen [22, 29, 35]. Wir bezeichnen all diese Ansätze als Self-Service Data Mining (SSDM). Es scheint auf der einen Seite unstrittig, dass Datenwissenschaftler nicht unabhängig von Domänenexperten arbeiten können, da es ihnen an einem ausreichenden Verständnis des Untersuchungsgegenstandes mangelt [44]. Dass Domänenexperten auf der anderen Seite Aufgaben von Datenwissenschaftlern übernehmen können, wird jedoch vermehrt angenommen [3, 4]. Ein Argument, das für diesen SSDM-Ansatz spricht, ist, dass fundierte Datenkompetenz zwar hilfreich, aber nicht für jede Form der Mustererkennung zwingend erforderlich ist [21]. Datenkompetenz (englisch: Data Literacy) wird als "die Fähigkeit des planvollen Umgangs mit Daten" definiert und beinhaltet die Kompetenzen, Daten erfassen, erkunden, managen, kuratieren, analysieren, visualisieren, interpretieren, kontextualisieren, beurteilen und anwenden zu können [18]. Inwieweit von Domänenexperten erwartet werden kann, dass sie mit Hilfe von SSDM den kompletten Analyseprozess übernehmen, soll mit der in diesem Artikel vorgestellten Versuchsreihe untersucht werden.

2 Aktueller Stand der Forschung

Abbildung 1 stellt SSDM als Teilgebiet von SSBI, bzw. EUC dar. EUC wird bereits seit den 1980er-Jahren erforscht [1, 30, 41]. Die Idee, Domänenexperten zu eigenständigen computergesteuerten Analysen zu befähigen, kam etwa zehn Jahre später auf [11], auch wenn der Begriff SSBI erst 2008 in der wissenschaftlichen Literatur erschien [43]. Dagegen gibt es bezüglich der aktuellen Evolutionsstufe von SSDM recht unterschiedliche Positionen. So waren Kriegel et al. [33] auf der einen Seite bereits im Jahr 2007 der Meinung, dass die Datenvorbereitung in künftigen Data-Mining-Plattformen automatisiert werden würde, ohne den Nutzern die Kontrolle über die einzelnen Schritte

[1] Datenwissenschaftler sind mit der Erkenntnisgewinnung aus Daten beschäftigt und benötigen neben analytischen Fähigkeiten vor allem Kenntnisse im Umgang mit großen Datenmengen [17].

[2] Data Mining ist die Gewinnung von Erkenntnissen und Bildung von Vorhersagemodellen auf Basis großer Datenmengen [14]. Im Kontext dieses Artikels verstehen wir unter dem Begriff den vollständigen Prozess der Informationsgewinnung, von der Datenvorbereitung bis zur Evaluierung der Modelle.

zu nehmen. Verschiedene Konzepte, die dank umfassender Automatisierung auch Gelegenheitsanwendern Zugang zu fortgeschrittenen Analysekapazitäten ermöglichen sollen, existieren auch bereits [37, 50].

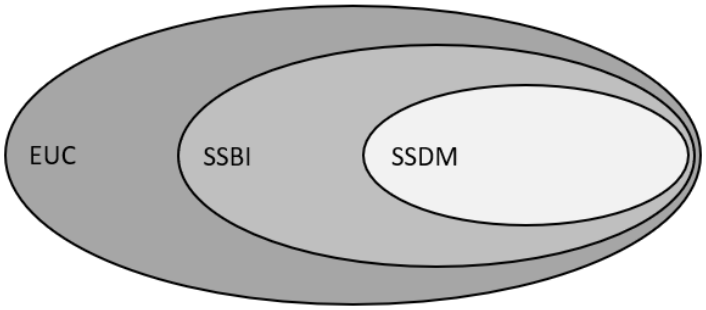

Fig. 1. Self-Service Data Mining als Teil von Self-Service Business Intelligence und End User Computing [46].

Auf der anderen Seite erachten Autoren wie Goyal und Vohra [20] die üblichen Data-Mining-Plattformen als wenig geeignet für Laien und schlagen die Entwicklung von speziell auf diese Zielgruppe ausgerichteter Software vor, die Benutzerfreundlichkeit über Genauigkeit und Flexibilität stellt. Andere Experten mahnen, dass die Arbeit mit Daten zu komplex ist, um alleine von Domänenexperten durchgeführt werden zu können [34, 49]. Brown [8] wies darauf hin, dass die Nachfrage nach SSA zu sinken scheint und es unwahrscheinlich ist, dass Experten in näherer Zukunft durch Software ersetzt werden können. Sie identifizierte jedoch Embedded Analytics, also die Integration von Analysewerkzeugen in Nicht-Analytics-Software, als Gebiet, auf dem es noch viel Potential gibt, da die meisten Nutzer ungerne zwischen verschiedenen Softwareplattformen wechseln.

Insgesamt gibt es hinsichtlich der Entwicklung von SSDM noch sehr unterschiedliche Meinungen und Argumente, weshalb weitere Betrachtungen erforderlich sind.

3 Methodik und Datengrundlage

In diesem Artikel soll eine Versuchsreihe vorgestellt werden, mit der überprüft werden kann, ob aktuelle Data-Mining-Software geeignet ist, sie Domänenexperten ohne datenwissenschaftlichen Hintergrund zur Verfügung zu stellen. Als Ansatz zur Beantwortung dieser Frage haben wir uns auf eine Gruppe von Algorithmen konzentriert, die der Literatur und unserer Erfahrung nach besonders gut für SSDM geeignet zu sein scheint [15, 28].

Während viele Data-Mining-Methoden hauptsächlich darauf ausgelegt sind, die Ergebnisgenauigkeit zu maximieren, ermöglichen Entscheidungsbäume zusätzlich eine verhältnismäßig leichte Interpretation ihrer Herleitung, da die einzelnen Schritte des Algorithmus grafisch dargestellt werden können und die wichtigsten Attribute dort in

einer hierarchischen Reihenfolge zu finden sind [7, 39, 47]. Entscheidungsbäume können somit als eine Art Untergrenze für SSDM gesehen werden. Wenn es für Domänenexperten nicht möglich ist, mit ihnen brauchbare Modelle zu erstellen, scheint dies mit komplexeren Algorithmen noch unwahrscheinlicher. Die Verständlichkeit der Modelle hat eine große Relevanz für ihren praktischen Einsatz, da Wissen über die Struktur von Modellen oft als ebenso wichtig empfunden wird wie genaue Vorhersagen von zukünftigem Verhalten [14]. Nutzer sind häufiger bereit, Analyseergebnissen zu vertrauen, deren Herleitung sie verstehen können [1, 15].

Für diese Untersuchung haben wir auf Grundlage eines ausgewählten Datensets Entscheidungsbäume in verschiedenen Softwareprodukten konstruiert und diese miteinander verglichen, um so Erkenntnisse über die Qualität der gebildeten Modelle und die Benutzerfreundlichkeit der Workflows zu gewinnen. Dabei wurde jeder Baum zweimal konstruiert. In einer Variante wurde versucht, das bestmögliche Modell zu bilden. In der anderen wurden nur ein Mindestmaß an Datenvorbereitung und die Standardkonfigurationen in den Programmen verwendet. Auf diese Weise sollte überprüft werden, welchen Unterschied Datenkompetenzen wirklich machen.

Eine Herausforderung ist die Entscheidung, welche Kennzahlen für eine Aussage über die Qualität der Modelle geeignet sind. In diesem Artikel wird neben der Genauigkeit der untersuchten Modelle auch die Komplexität der erzeugten Bäume, ermittelt durch die Anzahl der Ebenen und Blätter, als Kennzahl verwendet [12]. Dass eine weniger komplexe Struktur einem ansonsten gleichwertigen, aber komplexeren Modell vorzuziehen ist, wurde bereits in verschiedenen Kontexten nachgewiesen (vgl. z.B. Millersche Zahl [36], Ockhams Rasiermesser [19] und das Minimum Description Length Principle [5].

Die Untersuchungen in diesem Artikel wurden mit aktuellen Versionen der Plattformen IBM SPSS Modeler, RapidMiner, WEKA und KNIME durchgeführt. Sie wurden ausgewählt, da sie laut ihrer Anbieter allesamt Self-Service-Funktionalitäten besitzen, beziehungsweise sehr einfach und intuitiv zu benutzen sind [6, 14, 25, 26, 32, 40]. Um den Trade-off zwischen der erhöhten Benutzerfreundlichkeit und einer möglicherweise eingeschränkten Flexibilität der Produkte mit grafischen Benutzeroberflächen zu untersuchen, wurden zusätzlich R und Python berücksichtigt. Sie sind seit einigen Jahren die verbreitetsten Programmiersprachen in der Durchführung von Data-Mining-Analysen und bieten viele Funktionen, welche die Arbeit mit statistischen Modellen erleichtern [14, 45]. Beide setzen ein gewisses Maß an Statistik- und Informatikkenntnissen voraus und eignen sich somit weniger für SSDM.

Für die Vergleiche wurde das Datenset aus [13] genutzt.[3] Es bietet viele Möglichkeiten für Klassifikationen und ist Teil einer ausführlichen Studie, die für die in diesem Artikel gebildeten Modelle als Orientierung dient. Darin wurde eine Reihe von Algorithmen für die Klassifikationen getestet und jeweils der beste ausgewählt. Auf Heuristiken und andere Verfahren zur Minderung der erforderlichen Rechenleistung wurde dabei größtenteils verzichtet, um möglichst optimale Modelle zu finden. Unter anderem wurden für jedes Ziel 160 Millionen Entscheidungsbäume erstellt, die sich durch die

[3] Das Datenset enthält 1.885 Datensätze, denen 33 Attribute zugeordnet werden.

Konfiguration ihrer Parameter und die verwendeten Kombinationen von Attributen unterschieden. Damit eignet sich die Studie gut als Benchmark für unsere eigenen Versuche, die im Folgenden beschrieben werden.

Um die verschiedenen Softwareprodukte zu vergleichen, wurden die Datensätze als binäre Ziele klassifiziert. Die Daten wurden zuerst in Python vorbereitet und in eine Trainings- und eine Testpartition aufgeteilt. So wurde sichergestellt, dass die Versuche auf einer einheitlichen Grundlage stattfinden konnten. Der Schritt der Datenvorbereitung erfordert das höchste Maß an Datenkompetenzen und ist kaum von Domänenexperten durchzuführen. Mithilfe der Trainingspartition wurden dann Entscheidungsbäume der am weitesten verbreiteten Typen *C4.5*, *C5.0*, *CHAID* und *CART* konstruiert [7, 31, 38, 39]. Dabei wurde Cross-Validation genutzt, um die jeweils besten Konfigurationen zu finden. Im zweiten Teil der Versuchsreihe wurden die gleichen Entscheidungsbäume dann ohne Veränderung der Standardparameter erstellt, um die Herangehensweise eines Anwenders ohne fortgeschrittene Datenkompetenzen zu simulieren.

4 Erste Ergebnisse und weiteres Vorgehen

Insgesamt erzielten die Modelle relativ ähnliche Genauigkeiten, auch wenn die Baumstrukturen teilweise sehr unterschiedlich waren. Ein möglicher Grund dafür ist, dass recht viel Arbeit in die Datenvorbereitung gesteckt wurde (vgl. Kapitel 3), sodass die Modelle in den verschiedenen Tools auf einer gemeinsamen Grundlage aufgebaut werden konnten. Dieser Schritt dient unter anderem der Qualitätssicherung und erfordert ein gewisses Maß an Datenkompetenz und Domänenwissen. Die Modellierung selbst kann mithilfe der grafischen Benutzeroberflächen der verschiedenen Plattformen auch von Nutzern ohne tiefgehende Statistik- und Programmierkenntnisse vorgenommen werden. Allerdings führen diese Fähigkeiten meistens zu besseren Modellen, da die einzelnen Parameter der Algorithmen besser an die Daten angepasst werden können. Die Modelle aus dem zweiten Teil der Versuchsreihe wiesen in den beiden Kennzahlen *Genauigkeit* und *Komplexität* Verschlechterungen gegenüber den Modellen aus dem ersten Teil auf. Dies deutet darauf hin, dass auch Entscheidungsbäume ein gewisses Maß an Datenkompetenzen erfordern und nicht ohne Einschränkung für SSDM geeignet sind. Wie solche Datenkompetenzen in den nächsten Jahren weitere Verbreitung finden können, ist ebenfalls ein wichtiges Thema, das Aufmerksamkeit bedarf. Ohne sie wird ein großer Teil des Potentials in den stetig wachsenden Datenmengen ungenutzt bleiben.

Eine genauere Betrachtung der Ergebnisse und das Ziehen geeigneter Rückschlüsse sollen in den nächsten Schritten unserer laufenden Arbeit erfolgen. In dieser Untersuchung stehen Entscheidungsbäume im Fokus, weil sie leicht zu interpretieren sind. Diese Betrachtungen könnten in der Zukunft auf andere Datensets und weitere verhältnismäßig simple Repräsentationsformen wie Regelwerke, k-Nearest Neighbors oder Naive Bayes ausgeweitet werden. Ein alternativer Ansatz wäre eine geringere Gewichtung der Nachvollziehbarkeit der Modelle und stattdessen eine Konzentration auf die Automatisierung des gesamten Prozesses, wie er beispielsweise in [29] verfolgt wurde.

References

1. Alavi, M., Weiss, I.: Managing the Risks Associated with End-User Computing. Journal of Management Information Systems, 5-20 (1985).
2. Allahyari, H., and Lavesson, N.: User-oriented assessment of classification model understandability. 11th scandinavian conference on Artificial intelligence. IOS Press, (2011).
3. Alpar, P., Schulz, M.: Self-Service Business Intelligence. Business & Information Systems Engineering. 58.2, 151-155 (2016).
4. Banker, S., https://www.forbes.com/sites/stevebanker/2018/01/19/the-citizen-data-scientist, last accessed 2018/31/05.
5. Barron, A., Rissanen, J., Yu, B.: The Minimum Description Length Principle in Coding and Modeling. Information Theory 50 Years of Discovery, IEEE Press (1998).
6. Berthold, M., Cebron, N., Dill, F., Gabriel, T., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization. Springer (2007).
7. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression trees. Wadsworth & Brooks, Monterey, CA (1984).
8. Brown, M., https://www.forbes.com/sites/metabrown/2016/12/30/why-self-service-analytics-wont-replace-data-analytics-professionals-may-help-them, last accessed 2018/30/05.
9. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0. CRISP-DM consortium (1999).
10. Chiang, K., Wells, A., https://tdwi.org/articles/2017/03/21/5-rules-for-successful-self-service-analytics.aspx, last accessed 2018/30/05.
11. Codd, E., Codd, S., Salley, T.: Providing OLAP to user-analysts: An IT mandate. E.F. Codd & Associates (1993).
12. Fayyad, U., Irani, K.: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. IJCAI (1993).
13. Fehrman, E., Muhammad, A., Mirkes, E., Egan, V., Gorban, A.: The Five Factor Model of personality and evaluation of drug consumption risk. Data Science, 231-242 (2017).
14. Frank, E., Hall, M., Witten, I.: Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2016).
15. Freitas, A.: Comprehensible classification models: a position paper. ACM SIGKDD explorations newsletter 15.1, 1-10 (2014).
16. Gartner, https://www.gartner.com/newsroom/id/3570917, last accessed 2018/30/05.
17. Gartner, https://www.gartner.com/it-glossary/data-scientist, last accessed 2018/30/05.
18. Gesellschaft für Informatik: Data Literacy und Data Science Education: Digitale Kompetenzen in der Hochschulausbildung. Berlin (2018).
19. Gibbs, P., Hiroshi, S., http://math.ucr.edu/home/baez/physics/General/occam.html, last accessed 2018/30/05.
20. Goyal, M., Vohra, R.: Applications of data mining in higher education. International journal of computer science 9.2, 113-120 (2012).
21. Gualtieri, M.: The Forrester Wave: Predictive Analytics And Machine Learning Solutions, Q1 2017. Forrester Research (2017).
22. Halper: TDWI Self-Service Analytics Maturity Model Guide – Interpreting Your Assessment Score, TDWI (2017).
23. Huang, T C.-K., Liu, C.-C., Chang, D.-C.: An Empirical Investigation of Factors Influencing the Adoption of Data Mining Tools. International Journal of Information Management 32(3): 257-270 (2012).

24. Hyafil, L., Rivest, R.: Constructing optimal binary decision trees is NP-complete. Information Processing Letters, 15-17 (1976).
25. IBM, https://www.ibm.com/products/spss-modeler, last accessed 2018/30/05.
26. IBM, https://www-01.ibm.com/common/ssi/ShowDoc.wss?docURL=/common/ssi/rep_ca/2/897/ENUS217-442/index.html&request_locale=en, last accessed 2018/30/05.
27. Imhoff, C., White, C.: Self-Service Business Intelligence – Empowering Users to Generate Insights. TDWI Best Practices Report (2011).
28. Johansson, U., Niklasson, L.: Evolving decision trees using oracle guides. In: IEEE Symposium on Computational Intelligence and Data Mining, 238-244 (2009).
29. Kanter, M., Veeramachaneni, K.: Deep Feature Synthesis: Towards Automating Data Science Endeavors. In: 2015 IEEE International Conference on Data Science and Advanced Analytics (2015).
30. Kasper, G., Cerveny, R.: A laboratory study of user characteristics and decision-making performance in end-user computing. Information and Management, 87-96 (1985).
31. Kass, G.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, 29.2, 119–127 (1980).
32. KNIME, https://www.knime.com/about, last accessed 2018/30/05
33. Kriegel, H., Borgwardt, K., Kröger, P., Pryakhin, A., Schubert, M., Zimek, A.: Future trends in data mining. Data Mining and Knowledge Discovery, 87-97 (2007).
34. Mazón, J., Zubcoff, J., Garrigos, I., Ortega, R.: Open Business Intelligence: on the importance of dataquality awareness in user-friendly data mining. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops (2012).
35. Mierswa, I., https://rapidminer.com/blog/rapidminer-makes-self-service-advanced-analytics-available-hadoop/, last last accessed 2018/30/05.
36. Miller, G.: The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. The Psychological Review (63), 81–97 (1956).
37. Ogbuokiri, B., Udanor, C., Agu, N.: Implementing bigdata analytics for small and medium enterprise (SME) regional growth. Department of Computer science, University of Nigeria, Nsukka, Enugu state (2015).
38. Quinlan, J., https://www.rulequest.com/see5-unix.html, last accessed 2018/30/05.
39. Quinlan, J.: Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers (1986).
40. RapidMiner, https://rapidminer.com, last accessed 2018/30/05.
41. Rockart, J., Flannery, L.: The Management of End User Computing. In: Communications of the ACM (1983).
42. Schuff, D. Corral, K.: Enabling Self-Service BI: A Methodology and a Case Study for a Model Management Warehouse. Information Systems Frontiers 20(2), 275-288 (2018).
43. Spahn, M., Kleb, J., Grimm, S., Scheidl, S.: Supporting business intelligence by providing ontology-based end-user information self-service. In: Proceedings of the first international workshop on Ontology-supported business intelligence (2008).
44. Viaene, S.: Data Scientists aren't Domain Experts. IEEE IT Professional 15(6) 12-17 (2013).
45. Wallace, B., Dahabreh, I., Trikalinos, T., Lau, J., Trow, P., Schmid, C.: Closing the Gap between Methodologists and End-Users: R as a Computational Back-End. Journal of Statistical Software 49.5 (2011).
46. Watson, H.: Tutorial: Business Intelligence – Past, Present and Future. In: Communications of the Association for Information Systems (2009).
47. Witten, H., Eibe, F., Hall, M.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Burlington (2011).

48. Wu, X.: Top 10 algorithms in data mining. Knowledge and Information Systems 14.1., 1-37 (2008).
49. Zaghloul, M., Ali-Eldin, A., Salem, M.: "Towards a Self-service Data Analytics Framework." International Journal of Computer Applications 80.9, 41-48 (2013).
50. Zorrilla, M., García-Saiz, D.: A service oriented architecture to provide data mining services for non-expert data miners. Decision Support Systems 55.1., 399-411 (2013).