RESEARCH

Kai Jannaschk

Infrastruktur für ein Data Mining Design Framework

Eine Untersuchung mit Fallbeispielen

Springer Vieweg

Infrastruktur für ein Data Mining Design Framework

Kai Jannaschk

Infrastruktur für ein Data Mining Design Framework

Eine Untersuchung mit Fallbeispielen

Springer Vieweg

Kai Jannaschk
Kiel, Deutschland

Zugl.: Dissertation, Christian-Albrechts-Universität Kiel, 2017

ISBN 978-3-658-22039-6 ISBN 978-3-658-22040-2 (eBook)
https://doi.org/10.1007/978-3-658-22040-2

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

Springer Vieweg

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Akronyme

AMBI AZTI's Marine Biotic Index. 152

BQI Benthic Quality Index. 152, 153, 155, 185

CRISP-DM Cross Indutries Standard Process in Data Mining. 14–17, 19, 20

DAG Direkter azyklischer Graph. 130

DDP Data Driven Perspective. 54, 67, 68, 73

DM Data Mining. 1–9, 11, 12, 14, 17–21, 30–32, 38, 39, 41, 45, 49–56, 58, 59, 65–67, 70–75, 77, 80–85, 95, 98, 99, 101, 122, 137, 146, 149, 152, 153, 181–185, 200

DMD Data Mining Design. 52, 54, 99, 154, 166, 183, 185

DSF Design-Science-Forschung. 47, 48

DWH Data Warehouse. 24, 28

KDD Knowledge Discovery in Databases. 6, 12, 13, 18–20

MAR Missing at Random. 96, 97

MCAR Missing Completly at Random. 96, 97

MDP Model Driven Perspective. 54, 59, 67, 68

NMAR Not Missing at Random. 96, 97

UDP User Driven Perspective. 54, 56–59, 62, 67

WRRL Wasserrahmenrichtlinie. 151, 152, 155, 156

1 Einleitung

„Lieber Geld verlieren als Vertrauen."

Ein Satz, der dem Unternehmer Robert Bosch zugeschrieben wird. Er spiegelt im Kern die drei Grundsätze Glaubwürdigkeit, Zuverlässigkeit, und Verantwortung in seiner Unternehmensführung wider.

Im Kontrast hierzu stehen Programme, wie z. B. „PRISM" und „TEMPORA" (Macaskill u. a. (2013), Gellman (2013)). Diese Namen stehen synonym für Programme, welche ungerichtet und wahllos Kommunikationsdaten und deren Metainformationen sammeln, auf diesen Datensammlungen durch Anwendung von Algorithmen Zusammenhänge erkennen (wollen) und entsprechende Erkenntnisse den Nutzern der Systeme aufzeigen. Vordergründiges Entwicklungsziel solcher Programme besteht in der Abwehr von terroristischen Gefahren. Um diese in Datensammlungen aufzuzeigen, werden extensiv Verfahren angewendet, die unter dem Schlagwort „Data Mining (DM)" geführt werden.

Der Begriff „Data Mining" ist ein Sammelbegriff für eine Menge von Theorien und Techniken im Bereich der Datenanalyse. In der englischsprachigen Literatur finden sich weitere ähnlich klingende Bezeichnungen, so z. B. „knowledge mining from data", „knowledge extraction" oder auch „data dredging" (Han und Kamber (2006)). Diese Begriffe werfen Fragen auf:

- Was versteht man unter dem Wort „knowledge"?
- Was heißt Wissen?
- Welche Eigenschaften besitzen die Daten?
- Wie schaut der Prozess des „Mining" aus?

K. Jannaschk, *Infrastruktur für ein Data Mining Design Framework*,
https://doi.org/10.1007/978-3-658-22040-2_1

- Ist ein solcher Prozess tatsächlich unstrukturiert, wie es der Begriff des „dredging“ suggeriert?

In der Arbeit von Yang und Wu (2006) ist eines der zu lösenden Probleme im Bereich des DM die fehlende Systematik in der Herangehensweise. Ein Hauptkritikpunkt ist, dass das Vorgehen bei DM-Projekten entweder zu „ad-hoc“ oder aber zu speziell auf das zugrunde liegende Problem zugeschnitten ist.

In DM-Tools wie WEKA[1] oder RapidMiner[2] bekommt ein Anwender eine Palette von möglichen Verfahren an die Hand gegeben, um einen Datensatz zu analysieren. Die Auswahl und Eignung der Verfahren zur Analyse seines Datensatzes obliegt dem Anwender.

Vergleichbar ist dies mit einem Werkzeugkasten. Der Nutzer des Kastens hat die Wahl zwischen zahlreichen Werkzeugen. Diese Werkzeuge wurden entwickelt, um bestimmte Problemen mit ihnen bewältigen zu können. Manche Werkzeuge lassen sich auf verschiedenste Problemstellungen anwenden. So kann man mit einem Hammer durchaus Nägel als auch Schrauben in einer Wand versenken. Aber ist der Hammer tatsächlich immer die richtige Wahl? Auch wenn ein Anwender weiß, wie ein Hammer funktioniert, muss man nicht alles als Nagel betrachten.

Bei Wu u. a. (2007) werden eine Reihe der am häufigsten eingesetzten Algorithmen zur Datenanalyse bis zum Jahre 2006 aufgelistet. Die Algorithmen entstammen den Teilbereichen Klassifikation, Clustering, Statistical Learning, Association Analysis und Link Mining. Enthalten sind allein 4 Verfahren für die Klassifikation. Aber worin unterscheiden sie sich, und können sie ggf. je alternativ angewendet werden? Unter welchen Umständen sollten vor der Anwendung eines dieser Klassifikationsverfahren auf einem Datensatz zunächst andere Verfahren wie beispielsweise des Clustering erfolgen? Welche Auswirkung hat eine vorherige Clusterung auf das Ergebnis einer Klassifikation? Lassen sich Resultate des einen Verfahrens als qualitätsfördernde Maßnahmen eines anderen Verfahrens verwenden?

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://www.rapid-i.com/

Jedes Analyseverfahren setzt zunächst einen Datensatz voraus. Die Struktur eines Datensatzes lässt sich z. B. über ein geeignetes Relationsmodell beschreiben. Funktionale Abhängigkeiten spiegeln Abhängigkeiten zwischen den Attributen eines Datensatzes wider. Weiterhin existieren zahlreiche Qualitätsmaße zur Beschreibung eines Datensatzes (Dasu und T. Johnson (2003), Olson (2003), Batini und Scannapieca (2006)), um z. B. fehlende Werte oder Ausreißer innerhalb der Attributwerte anzugeben. In Abhängigkeit von den Skalen der Attribute existieren Verfahren, um solche Qualitätsmaße zu beeinflussen. Aber ist die Anwendung in jedem Fall sinnvoll? Werden generell alle vorhandenen Attribute eines Datensatzes für die Anwendung eines Analyseverfahrens benötigt, oder besteht die Möglichkeit, eine ausreichende Auswahl zu treffen, um eine aufgeworfene Hypothese als Fragestellung einer Datenanalyse anzunehmen oder abzulehnen?

Weiterhin stellt sich die Frage: Wie entstehen eigentlich die Datensätze, die für eine Analyse herangezogen werden? Aus welchen Quellen entstammen sie? Welchen Einfluss auf das Analyseergebnis besitzt die Historie der analysierten Daten?

Die Glaubwürdigkeit eines Analyseergebnisses hängt letztlich von der Nachvollziehbarkeit und der Systematik des gesamten Analyseprozesses ab. Dies fängt mit der Datenerhebung an, reicht über die richtige Aufbereitung der Daten für die Analyse, der Auswahl und Verkettung der Analyseverfahren, bis zur Beurteilung und Interpretation des Ergebnisses. Die Zuverlässigkeit eines Analyseergebnisses gibt darüber Auskunft, inwieweit es sich unter Verwendung analoger Verfahren reproduzieren lässt. Es liegt in der Verantwortung eines Analysten, wie die erzielten Resultate interpretiert und in bestehendes Wissen eingebettet werden.

Struktur der Arbeit

Mit der vorliegenden Arbeit wird das Ziel verfolgt, aufzuzeigen, wie DM glaubwürdig, zuverlässig und verantwortlich funktionieren kann. Daher erfolgt in Kapitel 2 nach einer Begriffsdefinition des DM, ein

Vergleich bestehender Prozessmodelle im DM, wie systematisches Data Mining nach Ansicht von Fayyad bzw. der Industrie im Falle des CRISP-Modelles erfolgen kann. Hierbei werden die entsprechenden Schwächen als auch Stärken dargestellt. Weiterhin werden veröffentlichte Fallbeispiele betrachtet und deren Problemstellungen und Defizite bei der Lösung aufgezeigt.

Im Anschluss wird in Kapitel 3 ein DM-Framework entwickelt, in welches die im Vorwort aufgezeigten Bestandteile und Begriffe wie „Algorithmus“, „Datenmodell“, „Historie“, „Qualität“ einsortiert und erweitert werden. Das Framework gliedert sich hierbei in drei Schichten, und zeigt auf, welche Teilfragen auf dem Weg von einer Problemstellung zu einer Lösung beantwortet, und wie erhaltene Lösungen in den Kontext eines Fragestellers integriert werden können.

Anschließend in Kapitel 4 werden die für dieses Framework benötigten Bausteine für einen erfolgreichen DM-Prozess definiert, und Lösungsmöglichkeiten für einige offene Fragestellungen im Bereich der Datenerfassung und -analyse mit Fokus auf der Durchführung naturwissenschaftlicher Experimente präsentiert. Dazu zählt die Betrachtung mit dem Begriff Datensatz selbst sowie der Qualität eines Datensatzes, die Entwicklung eines Systems zur Unterstützung bei der Erstellung und der Geschichte eines zu analysierenden Datensatzes, sowie die Betrachtung, wie Änderungsanforderungen in den laufenden Prozess integriert werden können. Außerdem wird ein Ansatz zur Systematisierung bestehender Datenanalysealgorithmen aufgezeigt.

Letztlich erfolgt in Kapitel 5 die Darstellung einiger durchgeführter Fallbeispiele, welche in Zusammenarbeit mit dem Helmholtz-Zentrum für Ozeanforschung Kiel entstanden.

2 Systematisches Data Mining - State of the Art

Um über DM zu sprechen, muss zunächst geklärt werden, was unter DM zu verstehen ist, mit welchen Prozessen DM-Resultate erzielt werden, welche Stärken und Schwächen diese Prozesse aufweisen, und wie diese Prozesse Anwendung in der Realität finden. Mit diesem Kapitel und den jeweiligen Abschnitten werden diese Fragen beantwortet.

2.1 Data Mining Definition

Eine Definition des Begriffes Data Mining bzw. Knowledge Discovery von Frawley, Gregory Piatetsky-Shapiro und Matheus (1992) aus dem Jahre 1992 lautet:

> Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Given a set of facts (data) F, a language L, and some measure of certainty C, we define a pattern as a statement S in L that describes relationships among a subset F_S of F with a certainty c, such that S is simpler (in some sense) than the enumeration of all facts in F_S. A pattern that is interesting (according to a user-imposed interest measure) and certain enough (again according to the user's criteria) is called knowledge.[1]

[1] Wissensentdeckung ist eine nicht triviale Extraktion von impliziten, ggf. unbekannten, und potentiell nützlichen Informationen aus Daten. Gegeben sei eine Menge von Fakten (Daten) F, eine Sprache L, und ein gewisses Maß an

K. Jannaschk, *Infrastruktur für ein Data Mining Design Framework*,
https://doi.org/10.1007/978-3-658-22040-2_2

Gerade die Verwendung des Begriffes „Wissen“ ist verwirrend. Wissen ist das Resultat einer Bewertung von Fakten durch eine Person oder einer Gruppe von Personen. Eine Bewertung ist somit subjektiv, und kann nicht als das alleinige Ergebnis einer Datenanalyse betrachtet werden.

Für die weitere Arbeit wird der Begriff „Anwender“ synonym für eine Einzelperson als auch für eine Gruppe von Personen verwendet.

Fayyad, G. Piatetsky-Shapiro und P. Smyth (1996a) kommen zu folgender Definition von Knowledge Discovery in Databases (KDD):

> KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[2]

Offen bleibt bei Fayyad, Shapiro und Smyth, was die Eigenschaften eines Musters *valide*, *neu*, *potentiell nützlich* und *verständlich* bedeuten. Ziel des DM-Prozesses ist lt. Definition, irgendein Muster zu finden, welches die genannten Eigenschaften erfüllt.

Edelstein (1999) definiert DM als:

> Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.[3]

Diese Definition des DM beschränkt sich auf die Vorhersage. Die Erklärung von Daten durch das Aufzeigen von Zusammenhängen ist lediglich ein Baustein auf dem Weg zu einer Vorhersage. Worin der

Sicherheit C, so definieren wir ein Muster als eine Aussage S in L, welche Beziehungen in einer Teilmenge F_S von F mit einer Sicherheit c beschreibt, so dass S einfacher (in gewissem Sinne) ist, als eine Aufzählung aller Fakten in F_S. Ein Muster, welches interessant (in Abhängigkeit von der Erwartungshaltung eines Anwenders) und sicher genug (entsprechend eines Qualitätskriteriums eines Anwenders) ist, wird Wissen genannt.

[2] Wissensentdeckung in Datenbanken ist ein nicht trivialer Prozess valide, neue, potentiell nützliche und letztlich verständliche Muster in Daten zu finden.

[3] Data Mining ist ein Prozess, der eine Vielzahl von Datenanalysewerkzeugen nutzt, um Muster und Zusammenhänge in Daten zu finden, die für valide Vorhersagen genutzt werden können.

Unterschied zwischen Beziehungen und Mustern besteht, wird nicht geklärt.

Edelstein (1999) formuliert weiterhin zwei Schlüsselaspekte für den Erfolg eines DM-Projektes:

1. eine präzise Formulierung des Problems,

2. das Nutzen der richtigen Daten zur Analyse.

Durch die Formulierung des Problems wird verdeutlicht, dass eine Zielsetzung für den DM-Prozess essentiell ist. Eine Zielsetzung ist bei den vorherigen Definitionen nicht zu finden. Die Notwendigkeit eines Zieles bestätigt ebenfalls Noonan (2000) mit der Aufforderung:

Begin with the end in mind.

Erst im zweiten Aspekt bei Edelstein (1999) sind die „richtigen“ Daten zur Analyse heranzuziehen, wobei zu klären ist, wie die „richtigen“ Daten gefunden werden können. Die Gefahr besteht, dass zur Bestätigung einer Hypothese zielgerichtet aufbereitete Daten herangezogen werden, und somit ein Analyseergebnis negativ beeinflusst wird.

Deutlich wird, dass Erkenntnisse, die mit DM-Methoden gewonnen werden, von einem Vorwissen eines Anwenders, und dem Bewusstsein über sein Wissensbedürfnis abhängen. Es ergeben sich die somit vier Szenarien (Tabelle 2.1) für die Anwendung der Methoden des DM.

- Ein Anwender besitzt Wissen, und ist sich der Anforderung an der Validierung seines Wissens bewusst.

- Ein Anwender ist sich bewusst, welches Wissen ihm fehlt. Er versucht somit, seinen subjektiven Wissensstand auszubauen.

- Ein Anwender verfolgt nicht bewusst das Ziel, sein Wissen zu bestätigen bzw. zu widerlegen.

- Einem Anwender fehlt sowohl das Wissen, als auch das Bewusstsein über das fehlende Wissen selbst.

Tabelle 2.1: DM-Szenarien unter Berücksichtigung von Vorwissen und Bedürfnis eines Anwenders

	Bewusstsein	kein Bewusstsein
Vorwissen	**Szenario 1**: Validierung zwischen vorhandenem Wissen und DM-Ergebnis	**Szenario 2**: DM-Ergebnis als überraschende Möglichkeit, Wissen zu validieren
kein Vorwissen	**Szenario 3**: gezielter Einsatz von DM-Ergebnissen zur Wissenserweiterung	**Szenario 4**: DM-Ergebnis wird als gegeben betrachtet

Werden DM-Algorithmen angewendet, so muss letztlich das daraus resultierende Muster durch den Anwender interpretiert und beurteilt werden. Die entsprechenden Beurteilungsmöglichkeiten sind jedoch von der Zielsetzung des Anwenders abhängig. Der grundlegende Unterschied zwischen den Szenarien besteht darin, dass ein Anwender sein Bedürfnis an Wissen kennt und formulieren kann. Sofern ein Bedürfnis klar formuliert ist, kann an einer Befriedigung dieses Bedürfnisses gezielt gearbeitet werden. Eine Beurteilung eines DM-Ergebnisses ist schwierig, sofern weder das Ergebnis in ein Vorwissen eingebettet werden kann, noch ein Anwender sich der bestehenden Wissenslücken bewusst ist.

DM ist als eine induktive Methode aufzufassen, bei der ein Anwender durch die Analyse von Daten zielgerichtet Informationen zu gewinnen sucht, die sich generalisieren bzw. mindestens auf andere unbekannte Daten übertragen lassen. Die jeweilige Zielstellung wird durch die subjektiven Interessen eines Anwenders, basierend auf seinen Instinkten, Gefühlen, Erfahrungen, seiner Intuition, seinen Werten, persönlichen Überzeugungen, seinem gesunden Menschenverstand, seiner durch kognitive und mentale Prozesse gegebenen Erkenntnis definiert, und die gewonnenen Informationen werden in sein abrufbares Wissen integriert.

Die durch eine Analyse zu erwartenden Informationen sind Beschreibungen über den Aufbau und die Beziehungen in analysierten Daten und reichen bis zu Prognosemodellen, mit welcher Wahrscheinlichkeit unbekannte Daten entsprechende identische Eigenschaften aufweisen. Zusammenfassend bezeichnet man die zu gewinnenden Informationen daher als Muster.

In Daten gefundene Muster unterliegen der Bewertung eines Anwenders. Ein kausaler Zusammenhang innerhalb der Bestandteile eines Musters wird durch eine Datenanalyse nicht gegeben. Ob ein gefundenes Muster tatsächlich der Realität entspricht und sich entsprechend einordnen lässt, muss durch den Anwender validiert werden. Die Frage, wie dieser Vergleich durchgeführt werden kann, um den tatsächlichen Wert und die Gültigkeit eines Musters zu prüfen, wird durch einen Analyseprozess nicht geklärt. Der Anwender bringt sowohl einen fachlichen Hintergrund für den Kontext des DM-Projektes mit, als auch den technischen Hintergrund für die Anwendung von verwendeten Analyseverfahren.

Um die Methode DM einzuordnen, wird im folgenden der Begriff des „Modelles" verwendet. Hierzu ist es erforderlich, diesen Begriff mit einer Semantik und zugehörigen Eigenschaften zu versehen, wie es Thalheim (2013) zeigt.

Ein *Modell* ist eine Abstraktion. Ein Modell unterliegt bei Konstruktion und Nutzung einem *Kontext*, in dem es eingebettet ist, unterliegt den Regeln und Richtlinien der Konstrukteure bzw. Anwender, der *Community of Practice*, verfolgt ein *Ziel bzw. Zweck*, wofür es geschaffen/genutzt wird, und wird durch die Auswahl, der mit dem Modell *repräsentierten Artefakte*, beschränkt. Konstruktion und Verwendung eines Modells unterliegt entsprechenden *Entwicklungs- und Verwendungsmethoden*. Paradigmen oder Restriktionen als *Grundlagen* der bei Konstruktion/Anwendung eingesetzten Methoden werden als bekannt vorausgesetzt. Diese Grundlagen wirken ebenfalls auf die Sprache, Konzepte, Muster o.ä., welche die *Basis* für ein Modell bilden. Die Zusammenhänge zwischen den Bestandteilen der Modelldefinition sind in Abbildung 2.1 illustriert.

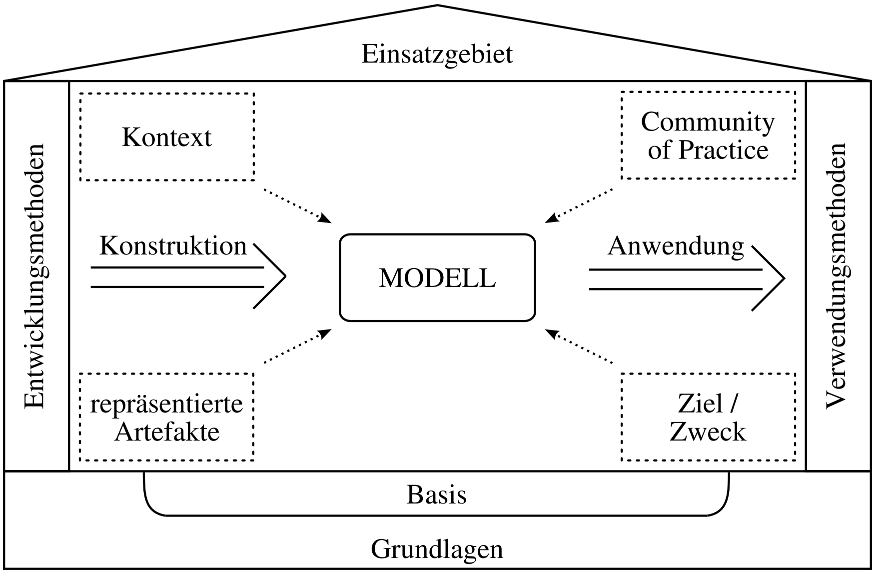

Abbildung 2.1: Definition des Begriffes „Modell“

Modelle werden als ein vereinfachtes Abbild der Realität genutzt, und ergeben sich aus verschiedensten Beobachtungen ebenjener Realität durch einen Anwender. Daraus resultierende subjektive Feststellungen werden mit Hilfe von Modellen beschrieben. Jedes Modell wird mit Konzepten und Theorien aus dem entsprechend betrachtetem Bereich angereichert. Ein Modell selbst kann sowohl ein Zusammenspiel von Modellen widerspiegeln, als auch mehrere Modelle in einer Modellfamilie zusammenfassen. Eine Modellfamilie ist hierbei eine Menge von Modellen mit ähnlichen Zielen und Hintergrund, die sich in spezifischer Parametrierung unterscheiden. Eine Modellfamilie ermöglicht es einem Anwender, einen Sachverhalt aus verschiedenen Blickwinkeln und mit unterschiedlichen Abstraktionsstufen zu betrachten. Ein System ist als eine Realisierung eines Modells zu verstehen. Bei der Realisierung werden entsprechende Annahmen und Rahmenbedingungen für das System festgelegt. Ein System kann mit der Realität interagieren, oder es kann Tests an Modellen durchführen.