Zur methodischen Vorbereitung von Data-Mining-Projekten unter Verwendung von CRISP-DM im Kontext diskreter Produktionsprozesse

Uwe Wieland, Marco Fischer

Technische Universität Dresden
Lehrstuhl für Wirtschaftsinformatik
Business Intelligence Research

Abstract

Die Analyse von Produktionsprozessen innerhalb von Data-Mining-Projekten stellt einen hohen Anspruch an die interdisziplinäre Zusammenarbeit zwischen Domänen- und Data-Mining-Experten. Die vorgeschlagene modellgestützte Methode offeriert dazu einen ersten Vorschlag, wie real-weltliche Produktionsprozesse in eine von analytischen Verfahren geprägte Welt überführt werden können. Dazu werden die Anforderungen aus beiden Bereichen sowie ein Standardvorgehen für diesen Diskursbereich in einem Modell integriert, dessen Erstellung nachfolgend erläutert wird.

1 Problemstellung und Motivation

Die industrielle Wertschöpfung erfolgt in fortwährend komplexer werdenden Produktionsprozessen, welche oft durch sehr vielschichtige Ursache-Wirkungsbeziehungen charakterisiert sind und hinsichtlich ihrer Planung, Durchführung, Steuerung und Kontrolle von umfangreichem Expertenwissen abhängig sind (Wiedenmann, 2001, S. 30 f.) Durch die zunehmende Verbesserung und Verbreitung von Sensorik und Aktuatorik steigert sich die Leistungsfähigkeit von autonomen Produktionssystemen, welche zukünftig selbständig miteinander durch Datennetze kommunizieren, entscheiden und agieren sollen.

Neben der geplanten technologischen Verbesserung im Prozessablauf stellt die enorme Menge an erzeugten Prozessdaten und das Management dieser Daten (Erhebung, Analyse und Verarbeitung) bereits heute eine zentrale Herausforderung, aber auch ein hohes Potential dar. Prozessdaten enthalten historisierte, wettbewerbsrelevante Informationen, welche die Charakteristik von Prozessen abbilden und daher neben dem menschlichen Expertenwissen einen gleichbedeutenden Stellenwert besitzen. Erklärungsmodelle sollen

dabei helfen, solche sozio-technische Systeme zu verstehen und anschließend zu verbessern (Kagermann, Wahlster & Helbig, 2013, S. 46 f.) Dieser Beitrag beschäftigt sich mit der methodischen Vorbereitung – Untersuchung der Geschäftsziele und Datenvorverarbeitung - von Data-Mining-Projekten zum Aufbau derartiger Erklärungsmodelle für diskrete Produktionsprozesse, welche über einen mächtigen Prozessdatenbestand verfügen.

Die Vorbereitung und Durchführung von Data-Mining-Projekten innerhalb von diskreten Produktionsprozessen stellt einen hohen Anspruch an die interdisziplinäre Zusammenarbeit zwischen Domänen- und Data-Mining-Experten. Dabei kommt es nicht allein auf die Daten an, sondern auch die Beschreibung der Daten und die untersuchte Domäne sind von großer Bedeutung für den Erfolg künftiger Projekte (Lukasz, Musilek, 2006, S. 19; Marban et al., 2007, S. 97 ff.; Mariscal, 2013, S. 160 ff.; Sharma & Osei-Bryson, 2009 S. 4114 ff.)

Mit CRISP-DM (Cross Industry Standard Process for Data Mining) basiert dieser Beitrag auf einem der meist verwendeten Vorgehen zur Durchführung von Data-Mining-Projekten, welches besonders im Anwendungsbereich der Industrie zu finden ist [Mariscal, 2010, pp. 139) Gemessen an der Evolution von Data-Mining-Vorgehensmodellen bildet CRISP-DM zum einen die Vereinigung von bereits sehr etablierten Vorgehen wie dem KDD-Prozess und industriellen Ansätzen (z.B. SEMMA) und zum anderen dient es als Ausgangspunkt für neue Ansätze (z.B. Cios et al. 2005, CRSIP-DM 2.0) (Azevedo, Santos, S. 185; Mariscal, 2010, S. 142). Das Referenzvorgehen CRISP-DM definiert und beschreibt pro Phase einzelne generische Aufgaben unabhängig vom Anwendungsbereich sowie den verwendeten Technologien, um Data-Mining-Projekte systematisiert durchführen zu können. Das Benutzerhandbuch als inhaltliche Erweiterung gibt ausführliche Tipps und Hinweise zu den einzelnen Phasen und deren Aufgaben (IBM, 2010, S. 3 f.; Lukasz & Musilek, S. 5). Die Lösung der Aufgaben ist jedoch von der jeweiligen Situation abhängig. Situationen werden durch einen Kontext definiert, welcher durch die Anwendungsdomäne und weitere Faktoren charakterisiert wird. CRISP-DM liefert ausschließlich ein sehr abstraktes Vorgehen für eine Zuordnung des generischen Modells auf konkrete Anwendungsbereiche (IBM, 2010, S. 4; Mariscal, 2010, S. 139). Zusammenfassend können daher folgende Begründungen für eine Konkretisierung (B) konstatiert werden:

B1: Eine situationsbezogene Konkretisierung ist methodisch nicht gewährleistet.

B2: Die Ermittlung der relevanten Datenquellen für eine Analyse bleibt sehr vage und isoliert vom eigentlichen Analyseobjekt.

B3: Die Integration von Rollen (z. B. Domänen- u. Data-Mining-Experte) in das Vorgehen wird nicht geregelt.

B4: Eindeutige Zusammenhänge und Abhängigkeiten zwischen den jeweiligen Ergebnissen der einzelnen Phasen werden außer Acht gelassen.

B5: Ergebnisse werden in ihrer Form nicht spezifiziert und sind damit nur schwer wiederverwendbar.

B6: Die Wiederverwendbarkeit von vorbereitenden Teilergebnissen bei unterschiedlichen Analyseverfahren ist nicht gegeben.

Bezugnehmend auf die Problemstellung und Motivation wird das Standardvorgehen CRISP-DM in ausgewählten Punkten der Analysevorbereitung spezialisiert und wiederverwendbar für diskrete Produktionsprozesse angepasst werden, um die interdisziplinäre Zusammenarbeit und damit den Aufbau von prozessspezifischen Erklärungsmodellen zu unterstützen.

2 Forschungsdesign

Das momentan etablierte Standardvorgehen für Data Mining CRISP-DM weist Konkretisierungsbedarf bezüglich einer Anwendung auf die Analyse von Prozessdaten diskreter Produktionsprozesse auf. Gerade in den Vorbereitungsphasen werden zahlreiche generische Aufgaben und Ergebnisse beschrieben, ohne eine konkrete Form zu definieren und eine Wiederverwendung zu adressieren. Entsprechend Design Science Research soll für die Lösung dieses Problems ein Artefakt in Form einer modellgestützten Methode konstruiert werden.

Aus der Problembeschreibung leitet sich die Notwendigkeit ab, die Vorbereitung von Data-Mining-Projekten im Kontext diskreter Produktionsprozesse methodisch zu unterstützen. Methoden werden dabei allgemein als Vorschriften bzw. Handlungsempfehlungen für Problemlösungen verstanden (Weller, 2010, S.36). Da sich Modelle als ein wichtiges Instrument der Erkenntnisgewinnung, Kommunikation und Dokumentation etabliert haben (Wand & Weber, 2002, S. 363), soll die methodische Unterstützung durch die systematische Verwendung von Modellen erfolgen. Damit lässt sich das zu erstellende Artefakt den modellgestützten Methoden zuordnen. Der Artikel verfolgt damit das Gestaltungsziel, eine modellgestützten Methode zur Vorbereitung von Data-Mining-Analysen im Kontext diskreter Produktionsprozesse zu entwickeln. Resultierend aus den vorangegangenen Betrachtungen ergeben sich für den Beitrag die folgenden Forschungsfragen:

- Welche Anforderungen bestehen an die Vorbereitungsphasen von Data-Mining-Projekten in diskreten Produktionsprozessen?
- Wie muss eine modellgestützte Methode gestaltet werden, um die Anforderungen zu erfüllen?

Zunächst werden die Anforderungen an diskrete Produktionsprozesse und die Datenvorverarbeitung in Data-Mining-Projekten beschrieben, die literaturgestützt erhoben worden und als Grundlage für die Konstruktion der modellgestützten Methode dienen.

Anschließend werden die Konstruktionsergebnisse präsentiert. Eine modellgestützte Methode besteht dabei stets aus einer Prozessbeschreibung, welche neben der Erstellung eines Modells, dessen Transformation und Nutzung beschreibt. Das konstruierte Modell muss konform zur verwendeten Modellierungssprache sein, die gemäß einer Sprachbeschreibung definiert ist [Weller, 2010, S. 42). Der Beitrag führt als Sprachbeschreibung ein Prozessmetamodell ein, das alle für die Prozessdatenanalyse relevanten Komponenten, deren Beziehungen und Regeln definiert. Mit Hilfe des Prozessmetamodells wird der Produktionsprozess im Ist-Zustand modelliert. Neben der Ablaufstruktur des Prozesses ist speziell die Strukturierung von Prozessdaten sowie die Abbildung existierender Abhängigkeiten und vermuteter Ursache-Wirkungsbeziehungen das Ziel der Modellierung, welche zunächst unabhängig von den Data-Mining-Zielen sowie den eingesetzten Verfahren ist. Der so modellierte Prozess kann mit Hilfe der Methode in ein prozessspezifisches Datenmodell transformiert werden, aus dem wiederum analysespezifische Falldatensätze abgeleitet werden können, welche die Grundlage für die Anwendung konkreter Data-Mining-Verfahren darstellen. Der Teil Modelltransformation und -nutzung ist in Kapitel 4.2 als Research in Progress gekennzeichnet.

Abschließend erfolgt eine merkmalsbasierte Evaluation, die nachweist, wie die konstruierten Artefakte der modellgestützte Methode die Vorbereitungsphasen des Data Mining nach CRISP-DM unterstützt.

3 Anforderungen

3.1 Anforderungen diskreter Produktionsprozesse

Innerhalb der betrieblichen Leistungserstellung bildet die Produktion einen Funktionsbereich, in dem Produktionsfaktoren miteinander kombiniert werden, so dass Endprodukte entstehen. Dabei wird die Produktion als Transformationsprozess angesehen. „Gegenstand der Produktion ist die Kombination und Transformation von Produktionsfaktoren (Input), so dass ein bestimmter Zweck (Output), das sogenannte Sachziel (z.B. die Herstellung von Automobilen), unter Beachtung des Formalziels (z.B. Gewinnmaximierung) bestmöglich erreicht wird. Die Ergebnisse des Transformationsprozesses sind die für den Absatzmarkt bzw. für weitere Transformationsprozesse bestimmten Güter oder Dienstleistungen, die nach ihrem Verwendungszweck als (End- oder Zwischen-) Produkte bezeichnet werden.“ (Kiener et al., 2012, S. 5). DANGELMAIER spricht davon, dass die

Inputs und Outputs eines Prozesses durch ihre relevanten Merkmalsausprägungen charakterisiert sind und sich darüber die Relation zwischen den einzelnen Inputs und Outputs herstellen lassen (Dangelmaier, 2009, S. 2ff.).

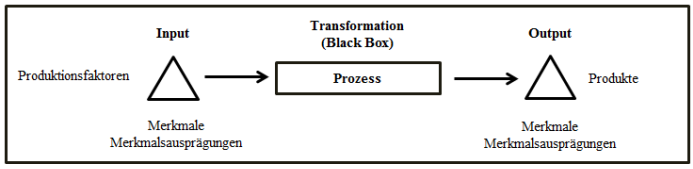

Abbildung 1: Input-Output-System von Produktionsprozessen [Da09], S. 3

Derartige Transformationsprozesse werden in der Regel als Input-Output-System beschrieben, welche durch eine Betrachtung der reinen Außensicht charakterisiert sind (Anforderung P1). Damit wird der Input-Output-Prozess wie in Abbildung 1 durch ein einziges Element mit dem nach außen wirksamen Objekten und deren Merkmalen beschrieben. Die Funktionalität des Transformationsschrittes bleibt dabei verborgen (Black-Box) und erfolgt entweder deterministisch oder stochastisch (Dangelmaier, 2009, S. 3). Produktionsprozesse können in einstufige und mehrstufige Prozesse unterschieden werden. Dabei bestehen die mehrstufigen Produktionsprozesse aus parallel oder sukzessiv ablaufenden einstufigen Produktionsprozessen. Somit stellen einstufige Produktionsprozesse stets eine Spezialisierung dar, welche innerhalb von komplexen Produktionsszenarien, zu mehrstufigen Produktionsprozessen kombiniert werden, was im Hinblick auf das Input-Output-System eine Komposition der einzelnen Elemente bedeutet. Innerhalb einer mehrstufigen Prozesskette wird auf die finalen Merkmale des Fertigungsobjektes hingearbeitet. Dazu müssen die Merkmalsänderungen an jedem Prozessschritt beherrschbar und transparent sein (Großmann & Wiemer, 2010, S. 856). Das Ziel eines jeden Produktionsprozesses ist ein reproduzierbarer Ablauf (Transformation), welcher stets zu einem konstanten Output führt (Weller, 2010, S. 70 f.; Wiedenmann, 2001, S. 27). Dafür sind gemäß GROSSMANN/WIEMER folgende Bedingungen zu erfüllen:

Tabelle 1: Bedingungen einer reproduzierbaren Fertigung nach Großmann & Wiemer, S. 855f.

Abk.	Bedingung	Beschreibung
P2	Definiertheit der finalen Produkteigenschaften	Festlegung des Fertigungsziels mit definierten Produkteigenschaften und dessen Toleranzgrenzen

P3	Durchgängigkeit der gesamten Prozesskette	Die Fertigung verläuft entlang einer durchgängigen Prozesskette und führt definiert und vollständig auf die finalen Eigenschaften des Produktes hin.
P4	Beherrschtheit der einzelnen Prozessschritte	Prozesse müssen bei einem gegebenen Input stets einen definierten Output liefern. (Definierte Merkmalsstruktur)
P5	Analysierbarkeit aller Prozessschritte	Zur Beherrschung eines Prozesses, muss jeder Prozessschritt beschrieben, analysiert und optimiert werden.
P6	Steuerbarkeit aller elementaren Zustandsänderungen	Jede elementare Zustandsänderung muss steuerbar sein.

In Bezug auf diese Bedingungen wird eine Einschränkung auf diskrete Produktionsprozesse eingeführt. Die Erweiterung „diskret“ erhalten Produktionsprozesse, in denen Produkte als abzählbare Einheiten hergestellt werden. Bei sogenannten Stückgutprozessen, können anders als bei kontinuierlichen Fertigungsprozessen, diskrete Schritte betrachtet, analysiert und gesteuert werden (Wiedenmann, 2001, S. 27f.).

3.2 Anforderungen an die Datenvorverarbeitung in Data-Mining-Projekten

Bezugnehmend auf die situative Anpassungsfähigkeit der zu konstruierenden modellgestützten Methode sind ausschließlich generelle Anforderungen zu ermitteln, welche die Phasen der Datenvorbereitung für eine Vielzahl an Data-Mining-Verfahren unterstützen. Die ermittelten Anforderungen sind mit den bereits in CRISP-DM implementierten Anforderungen für diese Phase abzugleichen, um die Nähe zum ausgewählten Standardprozess zu wahren. Die folgende Übersicht stellt die konsolidierten Anforderungen an die Datenvorverarbeitung dar und evaluiert diese durch weitere Quellen der Domäne:

Tabelle 2: Anforderungen Datenvorverarbeitung

Abk.	**Anforderungen**	**Quellen**
DM1	Umgang mit fehlenden Werten und deren Bedeutung (Missing Values) klären	(IBM, 2010), (Runkler, 2010), (Otte et al., 2004)
DM2	Skalenart ermitteln	(IBM, 2010) (Otte et al., 2004)
DM3	identische Formate/Schreibweisen pro Merkmal sicherstellen	(IBM, 2010), (Otte et al., 2004)

DM4	Wertebereiche von Merkmalen definieren	(IBM, 2010), (Otte et al., 2004)
DM5	Merkmalskorrelationen / Unabhängigkeiten entdecken	(IBM, 2010), (Otte et al., 2004)
DM6	statistische Lagewerte und Streuungsmaße zur Erkennung v. Datenrauschen/Ausreißern erheben	(IBM, 2010), (Runkler, 2010), (Otte et al., 2004)
DM7	Merkmalsbedeutungen / Relevanz ermitteln	(IBM, 2010), (Otte et al., 2004)
DM8	Balancierung der Daten untersuchen	(IBM, 2010), (Otte et al., 2004)
DM9	Schlüsselattribute/Schlüsselbeziehungen erkennen	(IBM, 2010), (Otte et al., 2004)
DM10	bekannte Ursachen und Wirkungen zwischen Merkmalen aufzeigen	(IBM, 2010)

Die Anforderungen der Datenvorbereitung fließen gemeinsam mit den Anforderungen diskreter Produktionsprozesse in die Konstruktion der modellgestützten Methode ein.

4 Konstruktionsergebnisse

4.1 Sprachbeschreibung

Ziel der modellgestützten Methode ist es, die Vorbereitung von Data-Mining-Analysen im Kontext diskreter Produktionsprozesse zu unterstützen. Dazu soll im ersten Schritt der zu analysierende Produktionsprozess im Ist-Zustand modelliert werden, um anschließend dieses Modell in analysespezifische Modelle zu transformieren, die für die Analyse der Prozess-Exemplardaten (Rohdaten des diskret gefertigten Produktes) genutzt werden können. Entsprechend dem Forschungsdesign erfordert eine modellgestützte Methode eine Sprachbeschreibung für eine Modellierungssprache, zu der die entstehenden Modelle konform sind. Jede Modellierungssprache verfügt dabei über eine festgelegte Syntax, die über eine Grammatik oder ein Metamodell beschrieben werden kann (Hesse & Mayr, 2008, S.389). Das vorgeschlagene Prozessmetamodell definiert dabei im Sinne einer Modellierungssprache alle möglichen Sprachkonzepte und Regeln zu deren Kombination (Wand & Weber, 2002, S. 364), um einen Produktionsprozess so zu modellieren, dass die Anforderungen diskreter Produktionsprozesse und an die Datenvorverarbeitung in Data-Mining-Projekten berücksichtigt werden. Als Ausgangspunkt für die Prozessmetamodellierung dient das Modell der Fertigungssteuerung (MFST) als etablierter Standard für diese Domäne (Dangelmaier und Felser, 1994, S. 35 f.; Großmann et al., S. 957). Durch

die Einschränkung des Anwendungsbereiches grenzt sich dieses Vorgehen von anderen Methoden der (Geschäfts-)prozessmodellierung ab (Dangelmaier & Felser, 1994, S. 36). Im MFST werden Produktionsprozesse als Input-Output-Systeme modelliert, in denen jeder Produktionsschritt als „Black-Box“ betrachtet wird, in den Güter hineinfließen (Inputs) und neue Güter hervorgebracht werden (Outputs) (Dangelmaier, 2009, S. 1 ff., 10 ff.). Der Produktionsprozess drückt sich dabei als Transformation von Input- und Output-Zuständen aus, die durch Merkmale und Merkmalsausprägungen charakterisiert sind. Mehrstufige Produktionsprozesse können in eine Kette einzelner Input-Output-Systeme zerlegt werden.

Abbildung 2 zeigt das entwickelte Prozessmetamodell, das in Anlehnung an MFST (Dangelmaier & Felser, 1994; Großmann et al., 2010) die Anforderungen diskreter Produktionsprozesse abdeckt und um die Anforderungen an die Datenvorverarbeitung in Data-Mining-Projekten erweitert wurde. Für die Darstellung des Prozessmetamodells wird ein Entity-Relationship-Modell gewählt und die Bedeutung der verwendeten Sprachkonzepte nachfolgend näher erläutert.

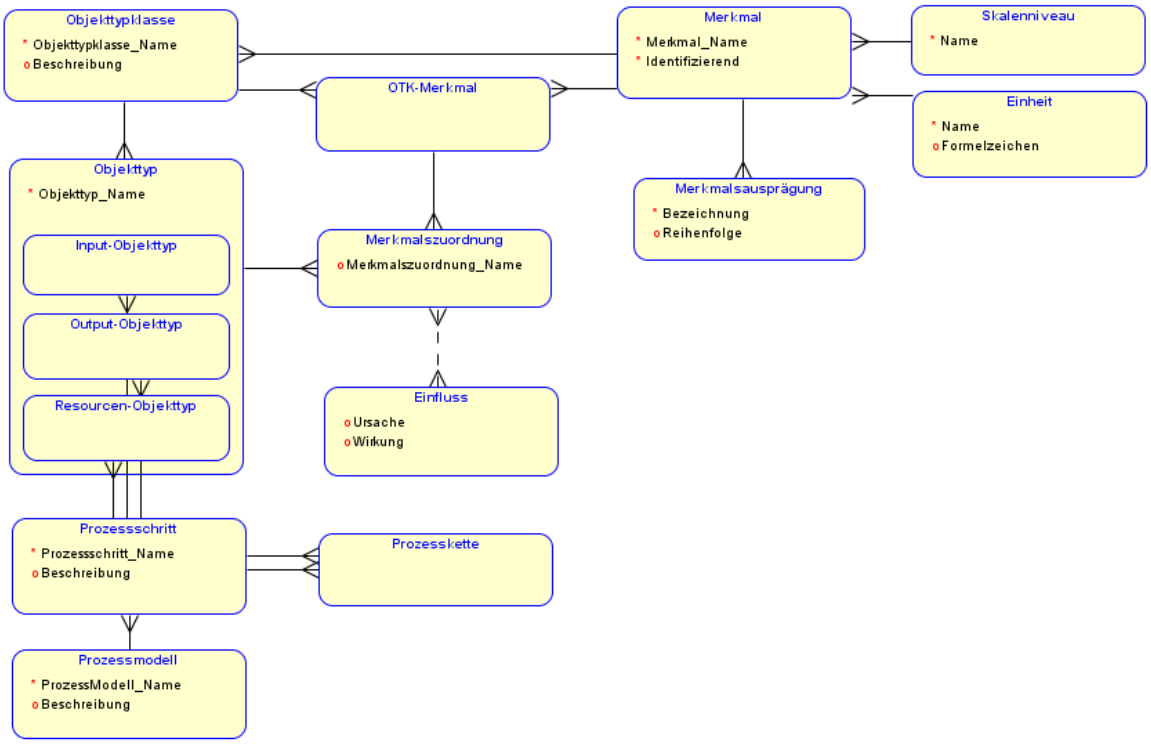

Abbildung 2: Prozessmetamodell – logische Sicht

Ein einstufiger Prozess bzw. ein ***Prozessschritt*** wird dabei grundsätzlich als eine Menge von ***Objekttypen*** repräsentiert, die unter Nutzung eines oder mehrerer ***Ressourcen-Objekttypen*** (z.B. Maschinen, Werkzeuge, u.a.) die Transformation eines oder mehrerer

Input-Objekttypen (z.B. Werkstücke, Material, u.a.) in ein oder mehrere ***Output-Objekttypen*** herbeiführen. Mehrstufige Prozesse werden über eine ***Prozesskette*** in eine chronologische Reihenfolge gebracht. Objekttypen abstrahieren individuelle Objekte (Exemplare) im Produktionsprozess und beschreiben immer genau einen Objektzustand. Sie können daher im Modell nur einmal verwendet werden. Selbst ein einziges zu bearbeitendes Produkt wird im Modell durch zwei Objekttypen abgebildet: einen für den Zustand vor der Bearbeitung (Input-Objekttyp) und einen für den Zustand nach der Bearbeitung (Output-Objekttyp) (Weller et al., 2010, S.76). ***Objekttypklassen*** (OTK) werden verwendet, um auf Modellebene dennoch den Zusammenhang zwischen gleichartigen Objekttypen hinsichtlich ihrer ***Merkmale*** abbilden zu können. Da die gleichen Merkmale von mehreren Objekttypklassen genutzt werden können, stellen die ***OTK-Merkmale*** diejenigen Merkmale dar, die aus der Gesamtheit aller Merkmale den Objekttypklassen zugeordnet sind. Für die Analyse mehrstufiger Prozesse ist zudem die Kenntnis von identifizierenden Schlüsselmerkmalen wesentlich, die direkt den Objekttypklassen zugewiesen werden. In der Regel wird dieses Merkmal ein Code oder eine Seriennummer sein mit der sich ein Exemplar eines Werkstückes im Produktionsprozess eindeutig bestimmen lässt.

Auf Modellebene können für Merkmale Sollwerte und Toleranzen für Merkmalsausprägungen vorgegeben werden, mit denen sich in der späteren Analyse der tatsächlichen Merkmalsausprägungen der Exemplardaten prozessuntypische Ausreißer erkennen lassen. Für Merkmale mit Nominal- oder Ordinalskala können im Prozessmetamodell gültige ***Merkmalsausprägungen*** definiert werden. Ein Vergleich dieses Werteumfangs mit dem tatsächlichen Vorkommens der Ausprägung in den Exemplardaten lässt auf die Balancierung dieses Merkmales schließen. Eine wichtige Anforderung aus dem Data Mining ist die Zuordnung des ***Skalenniveaus*** für jedes Merkmal. Daraus lässt sich später ableiten, welche Merkmale für welches Data-Mining-Verfahren in Frage kommen oder wie diese gegebenenfalls transformiert werden müssen. Für jedes Merkmal kann zusätzlich eine ***Einheit*** hinterlegt werden.

Jeder Objekttyp ist genau einer Objekttypklasse zugeordnet und besitzt daher alle Merkmale seiner Klasse, aber nicht alle Merkmale sind in jedem Prozessschritt von Interesse. Mit der ***Merkmalszuordnung*** können die für den entsprechenden Prozessschritt relevanten Zustände der verschiedenen Objekttypen einer Objekttypklasse beschrieben und für den jeweiligen Objekttyp gültige Sollwerte und Toleranzen lokal vorgegeben werden, welche die globalen Werte der Objekttypklasse überschreiben. Gleichzeitig lässt sich für den Objekttyp das Fehlen von Merkmalsausprägungen dokumentieren, was ein wesentliches Indiz für die Behandlung von Missing Values darstellt. Eine weitere wichtige im Prozessmetamodell abgebildete Anforderung, ist die Beschreibung von (vermuteten) Ursache-Wirkungsbeziehungen zwischen Merkmalen verschiedener Objekttypen. Ursache

und Wirkung lassen sich dabei als ***Einfluss*** in Form von Implikationen zwischen Aussagen, die mittels Variablen, Konstanten, Funktionen und (mathematischen u. logischen) Operatoren definiert werden, beschreiben. Gleichzeitig kann der Einfluss genutzt werden um Merkmalsabhängigkeiten innerhalb eines Objekttyps zu dokumentieren, was bei der anschließenden Auswahl von Eingangsmerkmalen für die Analyseverfahren unterstützt.

Aus dem Prozessmetamodell lässt sich ein relationales Datenmodell entwickeln, indem aus Entitäten Tabellen, aus Attributen Spalten und aus Relationen Schlüsselbeziehungen werden. Für die Implementierung des Prozessmetamodells und die Erfassung der analysespezifischen Prozessmetadaten und Objektzustände für die Modellerstellung eignet sich daher eine datenbankbasierte Anwendung.

4.2 Prozessbeschreibung

Modellerstellung

Gemäß Forschungsdesign benötigt die modellgestützte Methode neben der vorgestellten Sprachbeschreibung noch eine Prozessbeschreibung zur Erstellung eines Prozessmodells (Wiedenmann, 2001, S. 22; Weller, 2010, S. 42). Mit der Modellerstellung wird ein realer Produktionsprozess durch den Domänen-Experten, in ein Prozessmodell überführt, um anschließend für die Vorbereitung der spezifischen Prozessdaten eingesetzt zu werden.

In Anlehnung an die Vorgehensbeschreibung von CRISP-DM beginnt jedes Projekt mit der Formulierung der Geschäftsziele, welche den Anlass definieren und den Rahmen des Vorhabens bilden (Marban et al., 2007, S. 97 f.; Sharma & Osei-Bryson, 2009, S. 4116). Innerhalb der Analyse von Prozessen dienen die Geschäftsziele vordergründig der Beschreibung der geschäftlichen Erfolgsfaktoren und der Modellweite zur Fokussierung des zu untersuchenden Prozessausschnittes. Auf Grund der fehlenden Kenntnisse über die Input-Output-Relationen (Black-Box) zwischen den vorhandenen Prozessstrukturelementen werden für die Modellierung alle Objektzustände im ausgewählten Prozess betrachtet (Modellgranularität). Begrenzt durch die Modellweite erfolgt die Unterstützung der Situationsbeschreibung durch die Modellierung des konkreten Prozesses unter Verwendung der definierten Sprachbeschreibung.

Beginnend mit der Erfassung des Endproduktes und der beteiligten Fertigungsmittel als Objekttypklassen mit jeweils globalen Merkmalen und deren Ausprägungen, werden die einzelnen Prozessschritte mit ihrer Einordnung in der Prozesskette erfasst und einem Prozessmodell zugeordnet. Anschließend beginnt die Modellierung der einzelnen Prozessschritte in ihrer Außensicht, welche durch Input-, Output- und Ressourcen-Objekttypen sowie deren Merkmalszuordnungen definiert ist. Gemäß der globalen Merkmalszuord-

nung verfügt dabei jeder Objekttyp über die Merkmale seiner Objekttypklasse. Sollte innerhalb der Modellerstellung festgestellt werden, dass ein Merkmal fehlt, kann dies dem Merkmalskatalog hinzugefügt und über die Merkmalszuordnung verwendet werden. Im Anschluss an die Modellierung der Objekttypen können vermutete Ursache-Wirkungsbeziehungen innerhalb eines Objekttyps sowie zwischen Objekttypen in die Modellierung aufgenommen werden. Das beschriebene Vorgehen wird für jeden Prozessschritt innerhalb der Prozesskette durchgeführt. Das dazugehörige UML-Aktivitätsdiagramm des Ablaufes ist in Abbildung 3 dargestellt.

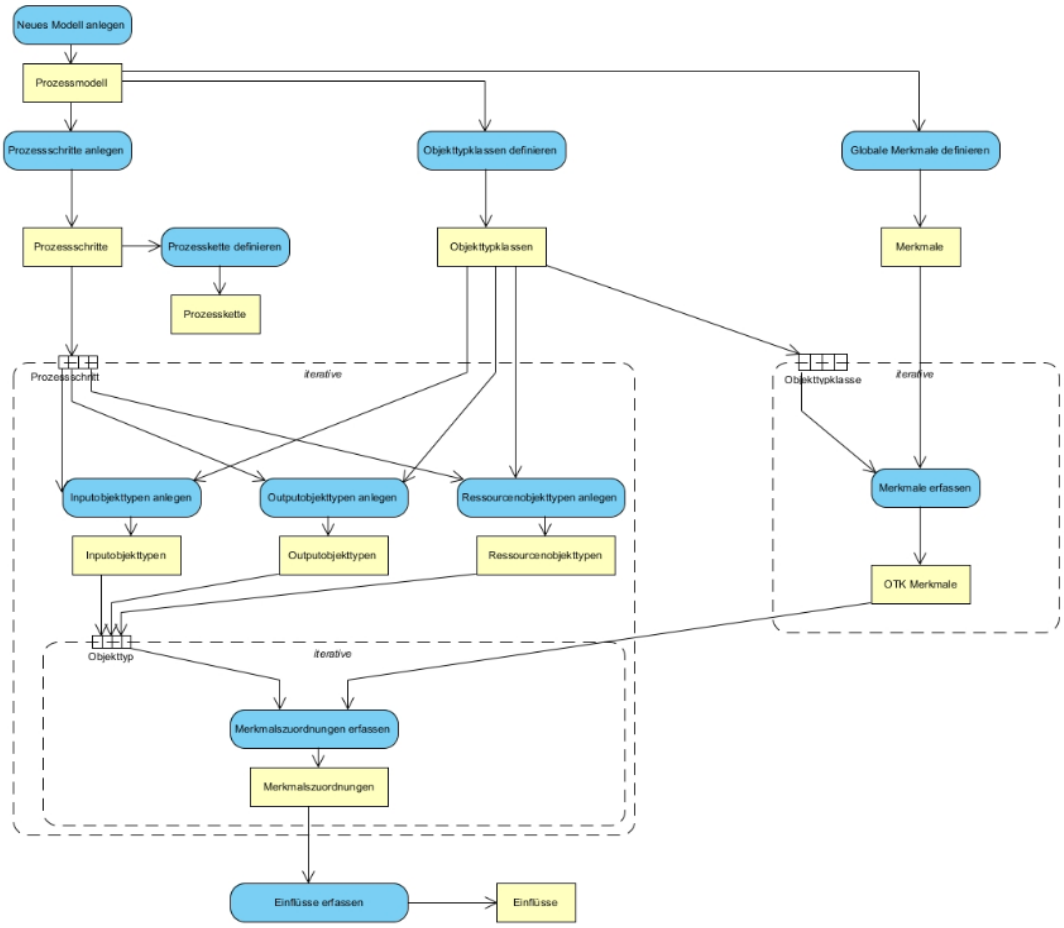

Abbildung 3: Prozessbeschreibung Modellerstellung

Reflektierend auf CRISP-DM wird innerhalb der Modellerstellung die Struktur des Prozesses erfasst und beschrieben. Anders als bei der darauffolgenden explorativen Datenuntersuchung und Überprüfung der Datenqualität, definiert die Modellerstellung Qualitätskriterien in Form von Metadaten und schafft Transparenz durch die Integration von Prozesswissen in Form von Ursache-Wirkungsbeziehungen. Durch die exakte Beschreibung

der Prozessstrukturen entsteht die Grundlage für eine semantisch korrekte Integration von relevanten Daten innerhalb der Datenaufbereitungsphase. Anhand dieser Aspekte wird sichtbar, dass durch das Prozessmodell eine Transformation des realen Prozesses in einen für Analysen aufbereiteten Untersuchungsbereich durchgeführt werden kann und somit die Überführung zwischen interdisziplinären Welten modellgestützt erfolgt.

Modelltransformation und Modellnutzung (Research in Progress)

Neben der Beschreibung der Modellierungsmethode zur Erstellung eines Prozessmodells geben modellgestützte Methoden konkrete Hinweise zur Transformation und Nutzung der erstellten Modelle. Durch eine Modelltransformation kann eine entsprechende Lösung zunächst wiederum als Modell dargestellt werden. Dazu wird das erzeugte Prozessmodell als Repräsentant der Domäne derart modifiziert bis eine Lösung des beschriebenen Problems auf Modellebene gefunden ist (Weller, 2010, S. 40). Sind sich alle Beteiligten einig eine Problemlösung im Modellraum gefunden zu haben, muss diese Lösung auf das fachliche Problem der Realität (Nicht-Modellraum) übertragen werden. Dies erfolgt mit einer Beschreibung der Modellnutzung.

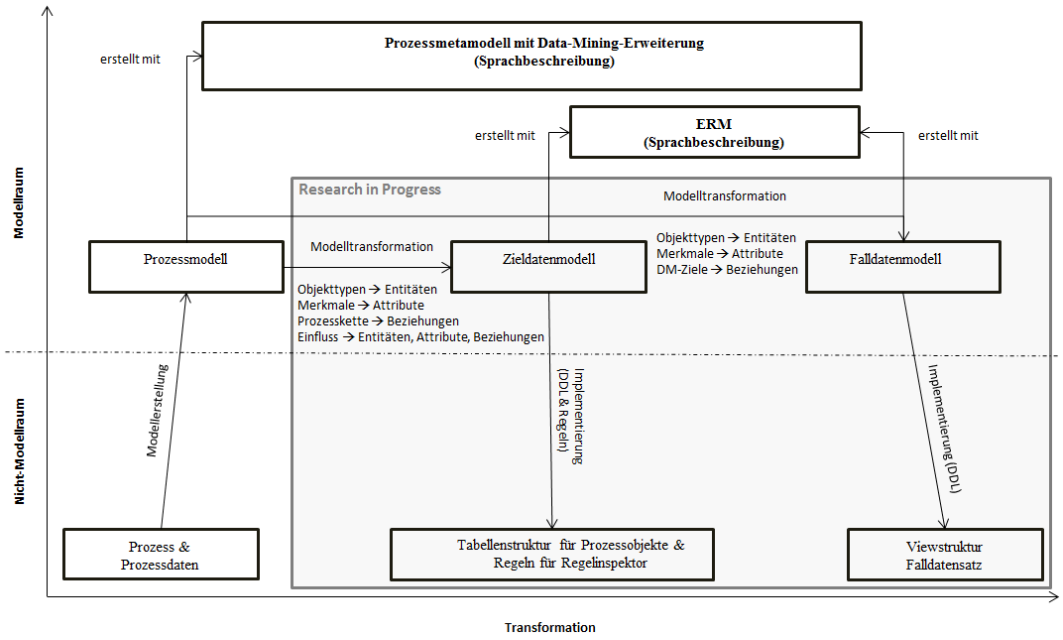

Abbildung 4: Research in Progress - Zerlegung des Modellierungsproblems

Bezogen auf die vorgestellte modellgestützte Methode soll das erzeugte Prozessmodell in ein Zieldatenmodell – in ERM-Sprachbeschreibung – überführt werden, um anschließend zur Generierung einer Tabellenstruktur für die konkreten Prozessdaten genutzt zu werden (siehe Abbildung 4). Diesbezüglich müssen Regelwerke definiert werden, welche den Übergang von einem Modell in das andere semiformal beschreiben. Die konkrete Nut-

zung des Zieldatenmodells erfolgt anhand eines generierten SQL-Skriptes, welches die relationale Tabellenstruktur anlegt und somit das Ziel für einen spezifischen Datenladeprozess aus den operativen Datenquellsystemen definiert. Weiterhin werden auf Basis der erfassten Prozessmetadaten (z.B. Sollwerte, Ursache-Wirkungsbeziehungen) Regeln generiert, welche zur Prüfung der Datenqualität auf die geladenen Prozessdaten angewendet werden.

Ausgehend von den modellierten Data-Mining-Zielen im Prozessmodell, werden alle vorgelagerten Objekttypen und deren Merkmale anhand eines identifizierenden Merkmals in einem analysegerechten Falldatenmodell – in ERM-Sprachbeschreibung – organisiert. Die konkrete Modellnutzung erfolgt über SQL-Skripte, welche je Data-Mining-Ziel eine View (Falldatensatz) auf die qualitätsgeprüfte Zieldatenstruktur definiert. Die erzeugten Falldatensätze werden anschließend als Ergebnis an die jeweiligen Data-Mining-Verfahren zur Analyse übergeben.

5 Evaluation und Ausblick

Die Systematisierung von Evaluationsmethoden innerhalb der Wirtschaftsinformatik (siehe Riege, Saat & Bucher, 2009, S. 75) bietet unterschiedliche Ansatzpunkte zur Evaluierung von entwickelten Artefakten. Im Beitrag wurde eine modellgestützte Methode als Lösungskandidat zur spezifischen Analysevorbereitung von Data-Mining-Projekten im Kontext von diskreten Produktionsprozessen vorgestellt. Anhand von abgeleiteten Konkretisierungsbedarfen und Anforderungen konnten die Sprachbeschreibung sowie die Prozessbeschreibung zur Modellerstellung fertig konstruiert und damit Teilaspekte der zweiten Forschungsfrage beantwortet werden. Die Prozessbeschreibung für die Modelltransformation und Modellnutzung wurden konzeptuell beschrieben und als Research in Progress gekennzeichnet. Für eine weiterführende Entwicklung ist es jedoch notwendig, die finalisierten Artefakte hinsichtlich ihrer korrekten Konstruktion auf Basis feststehender Anforderungen zu überprüfen (merkmalsbasierte Evaluation) (Riege, Saat & Bucher, 2009, S. 75). Die nachfolgende Übersicht macht die Ergebnisse der Evaluation ersichtlich und zeigt zugleich die konsolidierten Anforderungen zur Beantwortung der ersten Forschungsfrage auf.

Tabelle 3: Merkmalsbasierte Evaluation

Abk.	Anforderungen	Bewertung
Begründungen für eine Konkretisierung von CRISP-DM		
B1	Eine situationsbezogene Konkretisierung ist methodisch nicht gewährleistet.	Erfüllt – Input-Output-System ganzheitlich abgebildet

B2	Die Ermittlung der relevanten Datenquellen für eine Analyse bleibt sehr vage und isoliert vom eigentlichen Analyseobjekt.	Erfüllt – Objekttypen und deren Merkmalszuordnung beschreiben die exakte Zieldatenstruktur
B3	Die Integration von Rollen (z. B. Domänen- u. Data-Mining-Experte) in das Vorgehen wird nicht geregelt.	Teilweise – Domänen-Experte ist integriert, Data-Mining-Experte nach Modelltransformation
B4	Eindeutige Zusammenhänge und Abhängigkeiten zwischen den jeweiligen Ergebnissen der einzelnen Phasen werden außer Acht gelassen.	Teilweise – Prozessmodell als zentraler Ergebnis-speicher
B5	Ergebnisse werden in ihrer Form nicht spezifiziert und sind damit nur schwer wiederverwendbar.	Teilweise – in Sprachbeschreibung definiert
B6	Die Wiederverwendbarkeit von vorbereitenden Teilergebnissen bei unterschiedlichen Analyseverfahren ist nicht gegeben.	Teilweise – wird nach Modelltransformation vollständig erfüllt
Anforderungen diskreter Produktionsprozesse		
P1	Beschreibung des Prozesses als Input-Output-System	Erfüllt – Sprach- und Prozessbeschreibung
P2	Definiertheit der finalen Produkteigenschaften	Erfüllt - Sprachbeschreibung
P3	Durchgängigkeit der gesamten Prozesskette	Erfüllt - Sprachbeschreibung
P4	Beherrschtheit der einzelnen Prozessschritte	Erfüllt - Sprachbeschreibung
P5	Analysierbarkeit aller Prozessschritte	Erfüllt - Sprachbeschreibung
P6	Steuerbarkeit aller elementaren Zustandsänderungen	Erfüllt - Sprachbeschreibung
Anforderungen der Datenvorverarbeitung		
DM1	Umgang mit fehlenden Werten und deren Bedeutung (Missing Values) klären	Erfüllt – Sprachbeschreibung
DM2	Skalenart ermitteln	Erfüllt - Sprachbeschreibung
DM3	identische Formate/Schreibweisen pro Merkmal sicherstellen	Erfüllt - Sprachbeschreibung
DM4	Wertebereiche von Merkmalen definieren	Erfüllt - Sprachbeschreibung
DM5	Merkmalskorrelationen / Unabhängigkeiten entdecken	Erfüllt - Sprachbeschreibung
DM6	statistische Lagewerte und Streuungsmaße zur Erkennung v. Datenrauschen/Ausreißern erheben	Erfüllt - Sprachbeschreibung
DM7	Merkmalsbedeutungen / Relevanz ermitteln	Erfüllt - Sprachbeschreibung
DM8	Balancierung der Daten untersuchen	Erfüllt - Sprachbeschreibung

DM9	Schlüsselattribute/Schlüsselbeziehungen erkennen	Erfüllt - Sprachbeschreibung
DM10	bekannte Ursachen und Wirkungen zwischen Merkmalen aufzeigen	Erfüllt - Sprachbeschreibung

Zusammenfassend kann die Konstruktion der existenten Ergebnisse positiv evaluiert werden. Die Anforderungspunkte B3 bis B6 sind aufgrund der noch umzusetzenden Modelltransformation und -nutzung nur teilweise erfüllt.

Die modellgestützte Methode ermöglicht bereits mit der Modellerstellung eine gezielte Zusammenführung der Fach- und Data-Mining-Domäne und schafft damit ein Kommunikationsmittel für eine aufgabenteilige, interdisziplinäre Bearbeitung von analytischen Fragestellungen in diskreten Produktionsprozessen. Durch die integrierte, modellgestützte Beschreibung des betrachteten Produktionsprozesses entsteht eine einheitliche Wissensbasis über semantische Ablaufbeziehungen sowie über die zugrundliegende Datenstruktur und deren Besonderheiten. Mittels Zuführung von Metadaten erhält der Analyst notwendige Informationen über die Prozessdatenstruktur zur effizienten Konfiguration von Data-Mining-Verfahren.

Bezogen auf weitere Forschungstätigkeiten dient das Prozessmodell zur Generierung eines analysespezifischen Zieldatenmodells und spezifischer Falldatenmodelle für die vorhandenen Exemplardaten des jeweiligen Prozessschrittes sowie als Grundlage für eine strukturelle Überprüfung der Datenqualität und Verfahrensvoraussetzungen für die nachgelagerten Analysephasen. Die Konstruktion einer Prozessbeschreibung für die Modelltransformation und Modellnutzung sowie die Implementierung eines „Regelinspektors“ zur Realisierung der strukturellen Prüfungen auf Basis der integrierten Metadaten auf Merkmalsebene repräsentiert dazu den nächsten Forschungsschritt.

6 Literaturverzeichnis

Azevedo, A., Santos, M. F. (eds.): KDD, Semma and CRISP-DM: A parallel overview (2008)

Dangelmaier, W.: Theorie der Produktionsplanung und -steuerung Springer-Verlag, Berlin Heidelberg (2009)

Dangelmaier, W., Felser, W.: Ganzheitliche Modellierung von Fertigungsprozessen. Ein erster Schritt bei der Konstruktion unternehmensspezifischer Fertigungssteuerungssysteme. The Electronic Library of Mathematics 1994, 34–48 (1994)

Großmann, K., Wiemer, H., Weller, J., Großmann, K.K.: Reproduzierbare Fertigung in innovativen Prozessketten. Konzeption eines Beschreibungs- und Analysetools (Teil 2). ZWF - Zeitschrift für wirtschaftlichen Fabrikbetrieb 105, 954–958 (2010)

Großmann, K., Wiemer, H.: Reproduzierbare Fertigung in innovativen Prozessketten. Besonderheiten innovativer Prozessketten und methodische Ansätze für Ihre Beschreibung, Analyse und Führung (Teil 1). ZWF - Zeitschrift für wirtschaftlichen Fabrikbetrieb 105, 855–859 (2010)

Wolfgang Hesse, Heinrich C. Mayr: Modellierung in der Softwaretechnik: eine Bestandsaufnahme. Informatik-Spektrum, Vol. 31, No. 5., 377-393, (2008)

IBM: CRISP-DM 1.0 - Step-by-step data mining guide (2010)

Kagermann, H.; Wahlster, W.; Helbig, J.. (Hrsg.): Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0. Abschlussbericht des Arbeitskreises Industrie 4.0. Deutsche Akademie der Technikwissenschaften e.V., München, 2013.

Kiener, S., Maier-Scheubeck, N., Obermaier, R., Weiß, M.: Produktions-Management. Grundlagen der Produktionsplanung und -steuerung Oldenbourg Verlag, München (2012)

Knollmann, M., Meyer, M., Windt, K.: Data Mining-Methoden in der Produktionslogistik. Wissensgenerierung beim Umgang mit komplexen Daten und multikriterielen Entscheidungen. Industrie Management, 51–55 (2012)

Lukasz A., K., Musilek, P.: A survey of Knowledge Discovery and Data Mining process models. Cambridge University Press 2006, 1–24 (2006)

Marban, O., et al.: From the Business Decision Modeling to the Use Case Modeling in Data Mining Projects (2007)

Mariscal, G., Ó.M.C.F.: A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review 2010, 137–166 (2010)

Runkler, T.A.: Data Mining. Methoden und Algorithmen intelligenter Datenanalyse, VIEWEG+TEUBNER, Wiesbaden (2010)

Riege, C., Saat, J., Bucher, T.: Systematisierung von Evaluationsmethoden in der gestaltungsorientierten Wirtschaftsinformatik. in: Becker, J., Krcmar, H., Niehaves, B.: Wissenschaftstheorie und gestaltungsorientierte Wirtschaftsinformatik Physica-Verlag, Heidelberg (2009)

Sharma, S., Osei-Bryson, K.-M.: Framework for formal implementation of the business understanding phase of data mining projects. Expert Systems with Applications, 4114–4124 (2009)

Otte, R., Otte, V., Kaiser, V.: Data Mining für die industrielle Praxis Hanser Verlag, München, Wien (2004)

Weller, J.: Modellgestützte Prozessverbesserung. Entwicklung einer wiederverwendungsorientierten Methode zur durchgängigen Unterstützung der Modellerstellung, -transformation und -nutzung im Rahmen der Prozessverbesserung Dresden (2010)

Weller, J., et al.: Modellierung in der Produktionstechnik: Ein Ansatz zur effektiven Generierung von Technologie-Know-how für die Absicherung einer reproduzierbaren Fertigung. In: Esswein, W., Juhrisch, M., Turowski, K. (eds.): Modellierung betrieblicher Informationssysteme. Modellgestütztes Management, 69–86 (2010)

Wiedenmann, H.: Modellierung von Produktionsprozessen als Beitrag zur Generierung von Termin- und Kapazitätsplanungs-Systemen bei variantenreicher Serienfertigung Jost-Jetter Verlag, Heimsheim (2001)

Wand, Y.; Weber, R.: Research Commentary: Information Systems and Conceptual Modeling—A Research Agenda. In: Information Systems Research, 13, 363–377 (2002)