

Juli 2011

# Zwei vermeintlich gegenläufige Ansätze: Zur Integration von Geschäftsregeln und Data Mining

White Paper

Bosch Software Innovations



**BOSCH**

**Europa:**

Bosch Software Innovations GmbH  
Ziegelei 7  
88090 Immenstaad  
GERMANY  
Tel. +49 7545 202-300  
info-de@bosch-si.com  
www.bosch-si.de

**Amerika:**

Bosch Software Innovations Corp.  
161 N. Clark Street  
Suite 3500  
Chicago, Illinois 60601/USA  
Tel. +1 312 368-2500  
info@bosch-si.com  
www.bosch-si.com

**Asien:**

Bosch Software Innovations  
c/o Robert Bosch (SEA) Pte Ltd  
11 Bishan Street 21  
Singapore 573943  
Tel. +65 6571 2220  
info-sg@bosch-si.com  
www.bosch-si.sg

# Zwei vermeintlich gegenläufige Ansätze: Zur Integration von Geschäftsregeln und Data Mining

Irene M. Cramer

Häufig werden die Geschäftsregel- und Data Mining Ansätze als konkurrierend betrachtet, obwohl sie eigentlich ganz unterschiedliche Einsatzgebiete haben: Geschäftsregeln ermöglichen wertvolle Einblicke in die Berechnung von Geschäftsparametern und helfen dabei, implizites Wissen aufzuspüren, Data Mining bietet leistungsfähige Methoden für die Verarbeitung umfangreicher Datenbestände und für den Umgang mit komplexen, potenziell dynamischen Merkmalsräumen. In diesem Beitrag möchten wir an zwei Beispielen aus dem Bereich des Direktmarketings bzw. der Fertigungssteuerung zeigen, wie die beiden Konzepte kombiniert werden können.

## Inhalt

Einführung	2
Szenario 1: Direktmarketing	3
Szenario 2: Prozessleittechnik	6
Fazit	9
Danksagungen	10
Literaturhinweise und Software	10

## Einführung

Sehen Sie sich mit einer Flut an Daten konfrontiert? Sind Sie unschlüssig, mit welcher Methode Sie herausfinden können, ob sich in Ihren Daten etwas von Wert verbirgt? Suchen Sie nach relevanten Mustern? Hier empfiehlt sich eine Kombination von Geschäftsregeln und Data Mining, damit Sie die jeweiligen Stärken beider Ansätze bei der Datenverwaltung nutzen können. In diesem Beitrag zeigen wir anhand von zwei Anwendungsszenarien:

- Wie mithilfe eines Geschäftsregelsystems ein Data Mining-Toolkit mit Merkmalsvektoren gespeist werden kann.
- Wie sinnvolle Untergruppen ermittelt werden, wenn Sie die relevanten Merkmale bereits kennen.
- Wie markante Merkmale ermittelt werden, wenn Sie die relevanten Untergruppen bereits kennen.
- Wie die Ergebnisse der Data Mining-Schritte in Geschäftsregeln verwendet werden.

Betrachten Sie dieses Papier als eine Art Kochbuch: Wenn die Szenarien, d. h. die „Gerichte“ Direktmarketing oder Fertigungssteuerung, für Sie von geschäftlichem Interesse sind, lesen Sie sie wie ein Rezept. Wenn Sie etwas anderes kochen möchten, können Sie die Zutaten ändern und dieses Papier als eine Be-

schreibung der Methode verwenden. Wenn Ihr Anwendungsfall einen Unterbereich des Data Minings tangiert, der in diesem Papier nicht erörtert wird, nehmen Sie bitte Kontakt mit uns auf. Wir würden uns freuen, Ihnen ein anderes Rezept liefern zu können, das Ihrem speziellen Anwendungsfall gerecht wird.

Der Rest des Beitrags ist wie folgt strukturiert: Im Abschnitt „Szenario 1: Direktmarketing“ skizzieren wir das Direktmarketing-Szenario und erläutern, wie Ihre Geschäftsregeln mit automatisch konstruierten Klassifikatoren kombiniert werden können. Im Abschnitt „Szenario 2: Fertigungssteuerung“ skizzieren wir das Fertigungssteuerungsszenario und erläutern, wie Sie Ihre Daten analysieren können (z. B., um eine Hypothese zu überprüfen). Im Abschnitt „Fazit“ werden die wichtigsten Schritte zusammengefasst, die bei der Integration des Data Mining- und Geschäftsregelansatzes unternommen werden müssen.

## Einige wichtige Begriffe

Der Begriff *Merkmal* bezeichnet eine Variable (bzw. den Wert einer Variablen), die zum Charakterisieren eines Objekts verwendet wird. Ein Merkmal wird dann als unterscheidend betrachtet, wenn es dazu beiträgt, eine Gruppe von Objekten in Untergruppen zu zerlegen. Ein Merkmalsvektor ist ein n-dimensionaler Vektor aus Merkmalswerten, der ein Objekt repräsentiert.

Eine Untergruppe ist eine Sammlung von Objekten, die mindestens ein gleiches (oder ähnliches) Merkmal aufweisen. Eine Untergruppe wird als Klasse bezeichnet, wenn sie durch (intellektuelle) Abstraktion entstanden ist. Sie wird als Cluster bezeichnet, wenn sie auf Ähnlichkeit (der Merkmalsvektoren) basiert.

Beim **Data Mining** geht es um das Erkennen von Mustern und deren Extraktion aus Daten. Data Mining-Methoden werden für die (automatische) Verarbeitung großer Objektmengen und hochdimensionaler Merkmalsvektoren verwendet. Dies umfasst normalerweise vier Aufgabenbereiche: Clustering (Gruppenbildung), Klassifizierung, Regression und Erlernen von Assoziationsregeln. Es gibt zudem Data Mining-Methoden, die dazu beitragen, die relevanten Merkmale zu finden oder Daten zu analysieren und visualisieren; populäre Methoden hierfür sind: Korrelationsanalyse, Merkmalsauswahl oder Merkmalsgewichtung.

**Geschäftsregeln** sind Anweisungen, mit denen einige Aspekte des Unternehmens formalisiert werden. Sie können als Sätze in natürlicher Sprache, als grafisches Modell oder als mathematische Formel vorliegen. Gemäß der Business Rules Group (Hay, Healy, & Hall, 2000) können die Regeln Geschäftsbegriffe definieren oder in Beziehung setzen, und sie können Einschränkungen oder Ableitungen beschreiben. Geschäftsregeln werden von einem Geschäftsregel-Managementsystem wie Visual Rules (Informationen unter <http://www.visual-rules.de>) definiert, bereitgestellt, ausgeführt, überwacht und gepflegt.

## Szenario 1: Direktmarketing

Der Begriff **Direktmarketing** beschreibt eine Art der Werbung, die sich mit Techniken wie Katalogversand oder Werbeflehen (möglicherweise personalisiert) an die Mitglieder einer bestimmten Zielgruppe wendet. Im Gegensatz zur Fernseh- oder Radiowerbung oder zu Anzeigenkampagnen in Zeitschriften setzt das Direktmarketing voraus, dass die Reaktion der Konsumenten nachverfolgbar und messbar ist. Im B2C-Kontext (d. h. Business-to-Consumer, also Unternehmen-Verbraucher) wird das Direktmarketing in der Regel von kleinen und mittelständischen Unternehmen eingesetzt, die mit einem beschränkten Werbeetat eine maximale Rendite erzielen müssen.

Aus dem Blickwinkel des Vermarkters setzen Direktmarketingkampagnen das Sammeln geodemografischer Daten voraus sowie die Unterhaltung eines kundenorientierten Data Warehouses. Die direkten Reaktionen einer solchen Kampagne können wiederum verwendet werden, um die Marktsegmentierung und die Verbraucherprofile zu verfeinern. Aus dem Blickwinkel des Verbrauchers stellt Direktmarketing, sofern es sich um möglicherweise unerwünschte und irrelevante Werbung handelt, eine Verletzung der Privatsphäre dar.

Die Herausforderungen einer erfolgreichen Direktmarketingkampagne können daher wie folgt zusammengefasst werden:

- Zusammenstellen von qualitativ hochwertigen geo-demografischen Informationen und Verbraucherdaten aus unterschiedlichen Quellen
- Sorgfältiges Segmentieren des Marktes und Modellieren der Zielgruppenprofile
- Erstellen von gut abgestimmtem Werbematerial für unterschiedliche Mitgliedertypen in der Zielgruppe

**Szenario und Daten:** Da die Bindung vorhandener Kunden kostengünstiger ist als die Akquisition neuer Kunden, wird die Marketingabteilung eines mittelgroßen Online-Versandhändlers damit beauftragt, eine Direktmarketingkampagne für Kunden zu entwickeln, die dem Unternehmen in naher Zukunft verloren gehen könnten. In Verbindung mit dem Verlust von Kunden (auch als Kundenfluktuation bezeichnet) gibt es mehrere bekannte Faktoren, beispielsweise Unzufriedenheit mit dem Service, dem Preis oder der Kundenunterstützung, Differenzen bzgl. Rechnungsstellung, usw.

Die Marketing-Mitarbeiter wissen, dass mit Data Mining-Techniken Kundenfluktuationsmodelle berechnet werden können (Au, Li, & Ma, 2003). Die Kollegen in der Marketing- und Vertriebsabteilung sind jedoch Routiniers, was die Kundenbindung angeht, und dieser Erfahrungsschatz soll entsprechend berücksichtigt werden. Die Marketing-Mitarbeiter entscheiden sich daher, automatisch berechnete Modelle mit manuell erstellten Regeln zu kombinieren.

Die Kundenfluktuation soll auf der Basis von Daten modelliert werden, die im Data Warehouse des Unternehmens gespeichert sind. Diese Daten bilden Vektoren (siehe Abbildung 1), die die Kunden mithilfe verschiedener Merkmale beschreiben, u.a. Geschlecht, Kreditwürdigkeit, sozioökonomischer Status und Wohngegend des Kunden, Anzahl der Einkäufe in den letzten Monaten und Jahren, Anzahl Reklamationen und Retouren, usw.

Abbildung 1: Kunden als Merkmalsvektoren dargestellt (Auszug)

	B	C	D	E	F	J	AE	AF	PU
1	GENDER	CUSTOMER_RAT	NEIGHBORSES_OF_NEIG	AGE	CREDIT_RAT	PURCHASES	VALUE	YEA	
2	F		1 C	1	33	2	11	295,2	
3	M		1 U	1	37	2	3	447,69	
4	M		0 T	2	44	0	10	253,25	
5	F		0 T	4	54	2	5	704,76	
6	F		3 C	2	30	0	10	282,7	
7	M		1 R	1	46	4	10	767,02	
8	F		1 C	2	35	1	10	495,41	
9	F		2 R	1	51	3	9	409,55	
10	F		0 S	0	46	3	5	37,49	
11	M		3 S	0	54	2	11	1.659,96	
12	F		1 C	1	20	0	5	276,72	

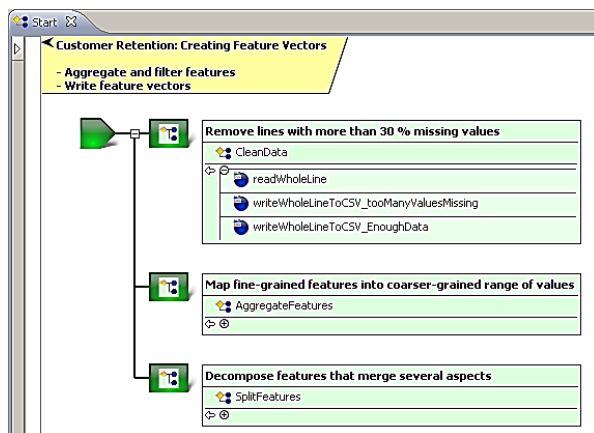
Jeder Kunde wird zudem mit einer von vier möglichen Kundenbewertungen gekennzeichnet: Neukunde (0), Gelegenheitskunde (1), Stammkunde (2) oder Großkunde (3). Die Bewertung des Kunden erfolgt unter anderem aufgrund der Häufigkeit der Einkäufe und der Höhe des Umsatzes.

Den Marketing-Mitarbeitern wird empfohlen, die Klassifikationsmodelle mithilfe des Data Mining-Tools KNIME (Einzelheiten unter <http://www.knime.org/>) auf der Basis von manuell ausgezeichneten, historischen Merkmalsvektoren zu berechnen. Diese Modelle klassifizieren Kunden als „Wechselkäufer“ oder „Stammkunden“ und prognostizieren, ob ein Kunde in naher Zukunft ggf. abwandern wird. Die Vermarkter bereiten daher einen Teil der Daten des vorherigen Jahres wie folgt vor: Kunden, die in den letzten Jahren als Stammkunden oder Großkunden geführt wurden, werden mit einer „Wechselkäufer“-Kennung markiert, wenn sie weder in den letzten sechs Monaten des vergangenen Jahres noch in diesem Jahr etwas bestellt haben. Zwei Vertriebsmitarbeiter prüfen anschließend die Daten und entfernen alle Zweifelsfälle.

**Vorgehen:** Da mehrere Merkmale kumuliert und gefiltert werden müssen, werden mit Visual Rules (siehe Abbildung 2) Geschäftsregeln modelliert, die

- Merkmalsvektoren mit mehr als 30 % fehlenden Werten entfernen.
- einige der zu fein aufgeschlüsselten Merkmale in einen weniger feingranularen Wertebereich überführen.
- einige der Merkmale zerlegen, die mehrere für das Szenario relevante Aspekte zusammenfassen.

Abbildung 2: Konstruktion von Merkmalsvektoren mit einem Regelmodell (Auszug)



Mithilfe von Geschäftsregeln werden die Daten zudem in das erforderliche Format konvertiert. Diese Daten werden dann genutzt, um die Klassifikationsmodelle zu trainieren und auszuwerten.

Abbildung 3: Beispielprozess für die Aufgabe „Klassifikation“ (mit KNIME)

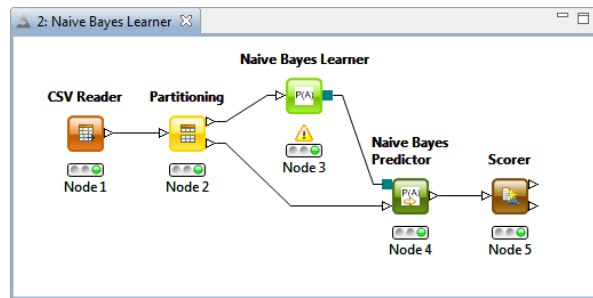


Abbildung 3 zeigt die wesentlichen Schritte im Prozess: Im ersten Schritt werden die Daten eingelesen (CSV Reader) und partitioniert (Partitioning), d. h. in Trainings- und Testdaten aufgeteilt. Im zweiten Schritt wird das Klassifikationsmodell basierend auf den Trainingsdaten berechnet (Naive Bayes Learner). Im dritten Schritt werden anhand des Klassifikationsmodells die Klassen der Testdaten bestimmt (Naive Bayes Predictor). Im letzten Schritt wird das Klassifikationsmodell ausgewertet (Scorer), d. h. für jedes Objekt in den Testdaten wird die prognostizierte Klasse mit der realen Klasse verglichen (Details im Feld „Populäre Auswertungskennzahlen“).

Angenommen, für das vorliegende Beispiel schneiden der Bayes-Klassifikator und der Entscheidungsbaum am besten ab. (Diese werden im vorliegenden Artikel der Einfachheit halber verwendet. Zweifellos gibt es ausgefeiltere Algorithmen; dennoch sind diese beiden für ihre gute Performanz bei bestimmten Aufgabenstellungen bekannt.)

### Populäre Auswertungskennzahlen

Die Leistungsfähigkeit eines Klassifikationsalgorithmus wird im Hinblick auf z. B. Präzision, Recall (Trefferquote) und Genauigkeit anhand einer spezifischen Wahrheitsmatrix gemessen.

$$precision = \frac{tp}{tp + fp} \quad recall = \frac{tp}{tp + fn}$$

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn}$$

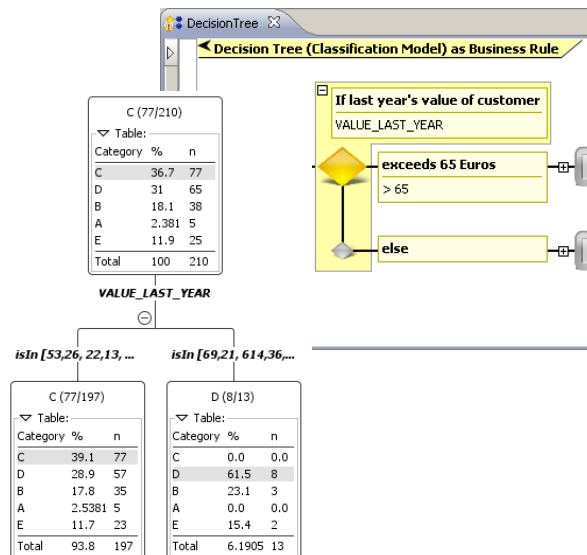
mit der Wahrheitsmatrix (binärer Fall):

		prognostizierter Wert	
		p'	n'
tatsächlicher Wert	p	richtig negativ	falsch positiv
	n	falsch negativ	richtig positiv

Der naive Bayes-Klassifikator ist als einfach und schnell bekannt, und trotz der Tatsache, dass sein simples Design grundlegende statistische Prämissen verletzt, funktioniert er häufig erstaunlich gut, wenn es um reale Probleme geht. Unglücklicherweise ist dieses Klassifikationsmodell nicht selbsterklärend, und seine Übersetzung in Geschäftsregeln ist kompliziert. Wenn die Marketing- und Vertriebsfachleute allerdings Geschäftsregeln aufgestellt haben, die ihrerseits einen Klassifikator darstellen, können die beiden Modelle kombiniert werden. Genauer gesagt müsste das Bayes-Modell über den Aufruf eines entsprechenden Dienstes im manuell erstellten Modell, d. h. in den Geschäftsregeln integriert werden. Mit der Kombination von Klassifikationsmodellen kann die Performanz des Klassifikationsensembles insgesamt signifikant erhöht werden. Eine umfassende Besprechung der möglichen Kombinationsmethoden für Klassifikatoren sowie der Vor- und Nachteile ist in (Tulyakov, Jaeger, Govindaraju, & Doermann, 2008) zu finden.

Für den Entscheidungsbaum ist nur wenig Datenvorbereitung erforderlich; es können daher äußerst schnell umfangreiche Datenbestände klassifiziert werden. Leider erzeugen Entscheidungsbäume manchmal Modelle, die keine Verallgemeinerung darstellen, d. h. sie erzeugen eine Überanpassung. Diesem Problem kann z. B. mit Pruning-Mechanismen begegnet werden. Da Entscheidungsbäume einfach zu verstehen und zu interpretieren sind (Abbildung 4 zeigt den Auszug eines Entscheidungsbaums, der auf der Basis der Direktmarketing-Merkmalvektoren berechnet wurde), können sie auf einfache Weise mit Geschäftsregeln kombiniert werden, die von Marketing- und Vertriebsfachleuten manuell aufgestellt werden.

Abbildung 4: Vereinfachter Beispiel-Entscheidungsbaum (mit KNIME berechnet) und zugehörige Geschäftsregel (mit Visual Rules modelliert)



Anhand der Klassifikation (die entweder mit einem Klassifikatorenensemble oder einem Geschäftsregelmodell berechnet wurde) können potenzielle Wechselkäufer in folgende Gruppen unterteilt und gezielt angesprochen werden:

1. Großkunden, mit denen es in den letzten Jahren keine rechnungsbezogenen Differenzen gegeben hat, erhalten einen Gutschein über 20 Euro sowie einen Fragebogen, in dem um Vorschläge, z. B. im Hinblick auf Serviceleistungen, die Produktpalette, usw. gebeten wird.
2. Großkunden, mit denen es kürzlich rechnungsbezogene Differenzen gegeben hat, erhalten einen Werbebrief mit einem Werbegeschenk und einem Gutschein über 15 Euro.
3. Stammkunden erhalten einen Werbebrief und einen Gutschein über 10 Euro.
4. Gelegenheitskunden erhalten einen Werbebrief und einen Gutschein über 5 Euro.

**Zusammenfassung:** Wie die beiden Klassifikationsbeispiele zeigen, können der Data Mining und Geschäftsregel-Ansatz auf unterschiedliche Weise kombiniert werden:

- Wenn das Klassifikationsmodell einfach zu interpretieren ist (wie ein Entscheidungsbaum, eine Entscheidungstabelle oder Entscheidungsregeln), kann es als Teil der Geschäftsregeln modelliert, d. h. reproduziert werden. In diesem Fall hilft das Data Mining-Toolkit beim Auffinden einiger geeigneter Klassifikationsregeln. Es vereinfacht zudem die Abschätzung der Mindestperformanz. Wenn ein Klassifikator in manuell erstellten Geschäftsregeln nachmodelliert wird, kann die Gesamtleistung schlechter ausfallen als die vom Data Mining-Toolkit abgeschätzte untere Schranke. Daher müssen die Geschäftsregeln in jedem Fall sorgfältig evaluiert werden (siehe Feld „Populäre Auswertungskennzahlen“), auch wenn der Ersteller des Modells ein Fachmann der Domäne ist.
- Wenn das Klassifizierungsmodell nicht einfach zu interpretieren ist (wie ein Bayes-Klassifikator, eine Support Vector Machine oder ein halbwegs komplexes neuronales Netzwerk), kann es als Dienstauftrag in die Geschäftsregeln integriert werden. In diesem Fall bilden die jeweils über das Data Mining-Toolkit bzw. die Geschäftsregeln konstruierten Klassifikatoren zusammen ein Klassifikationsensemble. Ein solches Ensemble muss ebenfalls sorgfältig evaluiert werden.

Welchen Nutzen hat diese Kombination der zwei Ansätze? Data Mining-Techniken haben sich in verschiedenen realen Anwendungen als sehr leistungsfähig erwiesen. Wie bereits erwähnt, können mit ihnen umfangreiche Datenbestände und komplexe, hochdimensionale Merkmalsräume verarbeitet werden. Automatisch berechnete Klassifikationsmodelle ermöglichen (zugegebenermaßen nicht unbedingt verständliche) Einblicke in die relevanten Klassifikationsparameter. Aus diesem Grund stehen einige Data Mining-Methoden skeptisch gegenüber. Geschäftsregeln sind (obwohl nicht in denselben Anwendungsbereichen) ähnlich leistungsfähig und in den meisten Fällen einfacher zu verstehen. Außerdem erweisen sich Ge-

schäftsregeln als extrem wertvoll, wenn sich die Unternehmensanforderungen schnell und häufig ändern: Sie werden nämlich von Fachleuten modelliert, die die Anforderungen am besten verstehen und wissen, wie auf implizites Wissen zugegriffen und dieses integriert werden kann. Die Kombination beider Ansätze bei der Klassifikation erzeugt daher einen außerordentlich willkommenen Synergieeffekt.

## Szenario 2: Prozessleittechnik

Die **Prozessleittechnik** erforscht und beschreibt Methoden bzw. Einrichtungen, die benötigt werden, um das Ergebnis eines Prozesses innerhalb eines gewünschten Bereichs zu halten. Damit verbunden ist zumeist die Steuerung von Sensoren und Aktoren. Beispielsweise ist der Produktfluss in einer Anlage ein solcher Prozess mit einem bestimmten, gewünschten Ergebnis: Hier wird der tatsächliche Wert mithilfe eines Durchflussmessers (Sensor) ermittelt, wohingegen der gewünschte Wert mit der Einstellung eines Ventils (Aktor) gesteuert werden kann.

In den meisten Anlagen werden Sensoren und Aktoren von einem System gesteuert, das als Prozessleitsystem bezeichnet wird. Dieses System besteht typischerweise aus einem Leitstand, d. h. einer Komponente, die alle Sensoren und Aktoren verwaltet, und einem Kontrollraum mit Überwachungs- und Anzeigeeinrichtungen (siehe Abbildung 5). Für Batch- bzw. kontinuierlichen Betrieb werden jeweils unterschiedliche Prozessleitsysteme eingesetzt. In diesem Beitrag konzentrieren wir uns auf die kontinuierliche Betriebsweise, die sich durch stetige Variablen und einen gleichmäßigen Produktfluss auszeichnet. Diese Betriebsweise wird bei der Herstellung sehr großer jährlicher Mengen von z. B. Chemikalien eingesetzt.

Abbildung 5: Beispiel für einen Kontrollraum (von RobertRED, <http://de.wikipedia.org/wiki/Leitstand>, abgerufen am 9. Februar 2011)



**Szenario und Daten:** Aufgrund der (zwar immer akzeptablen, jedoch) schwankenden Produktqualität, die vom Betriebsmeister und den Laborassistenten festgestellt wird, beschließt der Betriebsleiter eines Chemieunternehmens, Maßnahmen zur Qualitätssicherung einzuleiten. Zur Erklärung der beobachteten Schwankungen wurden zwei Hypothesen entwickelt:

1. Die Qualität des Endprodukts ist von der Qualität der Rohstoffe abhängig, deren Qualität wiederum von der Lieferanten- oder der Transportkette abhängig ist.
2. Die Prozesssteuerung von einer oder mehreren Anlagenteilen muss überarbeitet werden.

Die Produktion in der Anlage des Unternehmens wird mit einem Prozessleitsystem (hier Freelance von ABB, Details unter <http://www.abb.de/product/us/9AAC115759.aspx>) gesteuert, das 400 Sensoren überwacht und 200 Aktoren steuert. Diese Sensoren messen Temperatur, Druck, Durchflussrate, Füllstand und pH-Wert an unterschiedlichen Anlagenteilen. Die Einstellung der Aktoren, d. h. der Ventile, bestimmt die Durchflussrate und damit die Produktionsrate im Prozessschritt. Das Prozessleitsystem zeichnet alle fünf Sekunden die gemessenen Werte und die Einstellung der Ventile auf. Auf dieser Basis können Trendlinien berechnet und auf den Überwachungseinrichtungen im Kontrollraum angezeigt werden. Die Daten werden zudem fünf Jahre lang für den Fall archiviert, dass ein Audit durchgeführt werden muss. Alle vier Stunden prüfen die Laborassistenten die Qualität des Endprodukts und erfassen die Ergebnisse in einer Excel-Tabelle. In diesem Szenario gibt es drei Arten von Testwerten: Restfeuchte, Partikelgröße und Farbskala. Der Lieferant (und die Transportkette) jeder in der Produktion verwendeten Rohstoffcharge werden in einem SAP-System erfasst.

Der Betriebsleiter und die Laborassistenten beschließen, über einen Zeitraum von acht Wochen Daten zu sammeln. Sie vereinbaren, das Data Mining-Toolkit RapidMiner (Details unter <http://rapid-i.com>) zu verwenden und eine Korrelationsanalyse zu erstellen. (Sicherlich gibt es andere, ausgefeiltere Algorithmen für die Datenanalyse. Im vorliegenden Artikel wird die Methode gewählt, um die grundsätzliche Vorgehensweise zu demonstrieren.) Obwohl eine Korrelation nicht unbedingt eine kausale Beziehung impliziert, kann sie dennoch einen Hinweis darauf geben, welche der beiden oben genannten Hypothesen näher untersucht werden sollte.

**Vorgehen und Ergebnisse:** Aufgrund des zeitlichen Aspekts dieses Szenarios ist die Konstruktion der Merkmalsvektoren weniger offensichtlich als im Direktmarketingsszenario. Darüber hinaus müssen Informationen aus unterschiedlichen Quellen (d. h. dem Prozessleitsystem, einer separaten Excel-Tabelle und dem SAP-System) in geeigneter Weise zusammengeführt werden. Aus diesem Grund werden die Merkmalsvektoren mit Hilfe eines gleitenden Zeitfensters erstellt. Da die Laborassistenten alle vier Stunden die Qualität des Endprodukts prüfen, wird die Fenstergröße auf vier Stunden festgelegt. Die Sensorwerte und die Ventileinstellungen werden über diesen Zeitraum gemittelt, um die Datendimension zu verringern und Rauschen zu reduzieren. Mit Hilfe von Geschäftsregeln werden wie folgt Merkmalsvektoren konstruiert:

Pro Datensatz von Qualitätstestergebnissen in der Excel-Tabelle werden

- die Werte der 400 Sensoren und die Einstellungen der 200 Aktoren der letzten vier Stunden aus dem Prozessleitsystem ausgelesen.
- die Werte (Einstellungen) für jeden Sensor (Aktor) über diesen Zeitraum gemittelt, d. h. Mittelwert und Varianz berechnet (Details im Kasten „Warum Mittelwert und Varianz?“).
- auf Basis der Einstellung der Speichertankventile, der Produktionsrate und der im SAP-System erfassten Codenummern der Lieferant und die Transportkette ermittelt.

Abbildung 6: Prozess als Merkmalsvektoren dargestellt (Auszug)

	A	B	C	D	E	F
1	RESIDUAL_MOISTU	PARTICLE_SIZE	% COLOR_SCALE (KL	KFT10(4) MEAN	KFT10(4) VARIAN	KFT10(V) M
2	0,005	1,4	0,79	2990,2	720	45,1
3	0,002	2,9	0,78	3017,7	930	45,2
4	0,002	1,8	0,82	3036	960	45,2
5	0,002	2,2	0,81	3037,7	930	45,3
6	0,003	1,8	0,77	3000,2	900	45,0
7	0,005	2,2	0,83	2999,3	1230	45,0
8	0,003	2	0,77	2985,6	990	45,2
9	0,003	2,7	0,86	3046,7	1170	45,4
10	0,004	2,1	0,86	3009,2	1410	45,5
11	0,006	2,2	0,86	3005	1290	45,4
12	0,005	2,8	0,84	2959,3	660	44,9
13	0,003	1,4	0,87	2952,2	630	45,2

### Warum Mittelwert und Varianz?

Das arithmetische **Mittel** steht für die zentrale Tendenz (oder den Mittelwert) der 2.880 Werte (d. h. 12 Werte pro Minute x 240 Minuten = 2.880) pro Sensor bzw. Aktor. Der Mittelwert verdichtet die zeitliche Dimension der Merkmalswerte zu jeweils einem anstelle von ca. 3.000 Werten und erleichtert damit auch das Testen der oben angesprochenen Hypothesen.

Allerdings reagiert das arithmetische Mittel empfindlich auf Ausreißer und stellt daher ungleichmäßige bzw. verzerrte Verteilungen nicht immer adäquat dar. So können die Werte von zwei vierstündigen Zeiträumen mit dem gleichen Mittelwert erheblich abweichen: Nehmen wir beispielsweise an, dass viele der Werte im ersten Zeitraum sehr weit auseinander liegen, sich die meisten im zweiten Zeitraum jedoch in der Nähe des Mittelwerts befinden. Zur Darstellung dieses Unterschieds wird folglich eine weitere Kennzahl benötigt.

Diese liefert die **Varianz**. Sie gibt an, wie eine Reihe von Werten um den Mittelwert verteilt ist. Im Fall der oben erwähnten zwei vierstündigen Zeiträume bleibt der Mittelwert gleich, während die Varianz (vermutlich sogar große) Unterschiede aufweist. Dies lässt darauf schließen, dass im ersten Zeitraum der Produktdurchfluss in dem von diesem Sensor überwachten Prozessschritt geschwankt hat.

Abbildung 7 zeigt den Auszug einer Korrelationsanalyse, die mit RapidMiner berechnet wurde. Natürlich stehen in einem chemischen Herstellungsprozess viele der aufeinander folgenden Schritte bis zu einem gewissen Grad in Wechselbeziehung zueinander; in diesem Kontext sind jedoch die Korrelationen zwischen den Werten der Sensor-Aktor-Paare und den Qualitätskennzahlen von Interesse. Eine nähere Untersuchung dieser Korrelationen zeigt, dass eine der Qualitätskennzahlen (nämlich die Farbskala) in hohem Maße mit den Werten eines der Sensor-Aktor-Paare korreliert (Kennung des Sensors: KFP11(1), Kennung des Aktors: KFP11(V)). Eine Prüfung der Daten zeigt, dass die Werte des Sensors KFP11(1), eines Durchflussmessers, graduell ansteigen, und zwar von etwa 44,5 % Ventilöffnung zu Beginn der Testreihe bis etwa 45,5 % am Ende der acht Wochen. Gleichzeitig verschlechtern sich die Werte der Farbskala. KFP11(1) bestimmt die Menge des Kühlwassers, das für eine kontrollierte Kristallisation verwendet wird: Je höher die Temperatur des aus dem vorherigen Prozessschritt kommenden Produkts ist, desto mehr Kühlwasser wird benötigt, und je höher die Temperatur zu Beginn der Kristallisation ist, desto mehr unerwünschte Einschlüsse treten nach der Kristallisation im Produkt auf. Dies hat direkte Auswirkungen auf die Farbskala und somit auf die Qualität des Produkts. Obwohl der Sensor (mit der Kennung KFT10(4)) des vorherigen Prozessschritts, ein Thermometer, Normalwerte registriert, steigt die Menge des Kühlwassers. Der Betriebsleiter und die Meister beschließen daher, die Sensoren der gesamten Einheit zu prüfen und finden dabei heraus, dass das Thermometer im Prozessschritt vor KFP11(1) tatsächlich beschädigt ist. Siehe Abbildung 8 zur Illustration.

Die paarweise Korrelationstabelle (siehe Abbildung 7) zeigt zudem, dass keine der Qualitätskennzahlen mehr als moderat mit den Merkmalen Lieferant oder Transportkette korreliert. Das heißt, es gibt keinen Beleg für die zuvor erwähnte Hypothese 1: Die Qualität des Produkts scheint also nicht von der Qualität der Rohstoffe abhängig zu sein.

Insgesamt wurden 336 Merkmalsvektoren (d. h. 6 pro Tag x 7 Tage pro Woche x 8 Wochen = 336) mit den folgenden Merkmalen konstruiert (siehe Abbildung 6):

- 3 Qualitätstestwerte (Restfeuchte, Partikelgröße und Farbskala)
- 800 Sensorwerte (Mittelwert und Abweichung für 400 Sensoren)
- 400 Aktoreinstellungen (Mittelwert und Abweichung für 200 Ventile)
- 6 Datensätze für Lieferant und Transportkette (jeweils für 3 Rohstoffe)

Abbildung 7: Ergebnisse der Korrelationsanalyse, berechnet mit RapidMiner (Auszug)

First Attribute	Second Attribute	Correlation
COLOR_SCALE (KLETT)	KFP10(V) MEAN	0.129
COLOR_SCALE (KLETT)	KFP10(V) VARIANCE	-0.028
COLOR_SCALE (KLETT)	KFP11(I) MEAN	0.987
COLOR_SCALE (KLETT)	KFP11(I) VARIANCE	0.033
COLOR_SCALE (KLETT)	KFP11(V) MEAN	0.991
COLOR_SCALE (KLETT)	KFP11(V) VARIANCE	-0.078
COLOR_SCALE (KLETT)	SUPPLIER	0.008
COLOR_SCALE (KLETT)	TRANSPORTATION_CHAIN	0.062

Abbildung 8: Korrelierte Werte, die auf einen beschädigten Sensor hinweisen

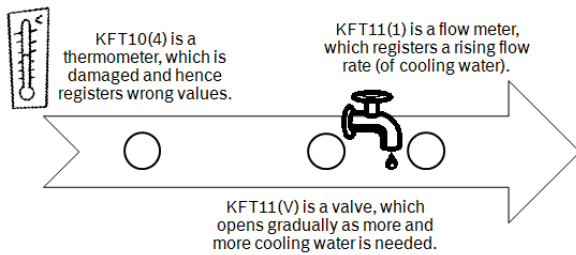


Abbildung 9 enthält eine Übersicht der Merkmale, deren Wertebereiche und weitere statistische Angaben. Sie zeigt, dass die Werte des Sensors KFT17(3) weit mehr Streuung aufweisen, also durch eine deutlich größere Varianz gekennzeichnet sind als bei allen anderen Sensoren. Eine eingehendere Analyse der Daten ergibt im Hinblick auf das Sensor-Aktor-Paar KFT17 ein interessantes Muster. (Das Sensor-Aktor-Paar besteht aus einem Sensor mit der Kennung KFT17(3) und einem Aktor mit der Kennung KFT17(V).) Wann immer die Abweichung von KFT17(3) eine hohe Streuung aufweist, steigt die Partikelgröße, d. h. die Qualität des Produkts wird beeinträchtigt. Bei näherer Untersuchung stellt sich heraus, dass KFT17(V) bei jedem Neustart des Prozesses zunächst stark oszilliert, bis die optimale Einstellung erreicht ist (siehe Abbildung 10). Dies ist der Grund für das beobachtete Abweichungsmuster, das zudem die oben angesprochene Hypothese 2 erhärtet, d. h., die Prozesssteuerung dieser Einheit muss verbessert werden.

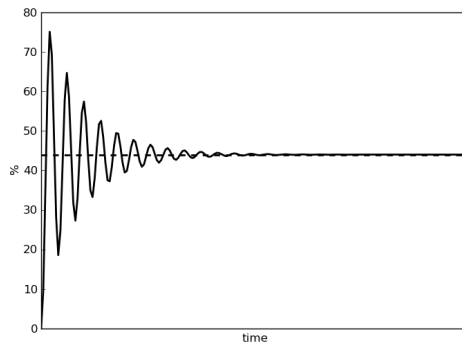
Abbildung 9: Übersicht der Merkmale einschließlich Statistiken (Auszug)

Role	Name	Type	Statistics	Range	Mi...
regular	KFP11(V) VARIANCE	integer	avg = 755 +/- 88.033	[600.000 ; 900.000]	0
regular	KFP11(I) VARIANCE	integer	avg = 740.714 +/- 89.203	[600.000 ; 900.000]	0
regular	KFT17(3) VARIANCE	integer	avg = 695.089 +/- 728.504	[150.000 ; 3600.000]	0
regular	KFP6(V) VARIANCE	integer	avg = 1082.679 +/- 258.950	[600.000 ; 1500.000]	0
regular	KFP3x4(2) VARIANCE	integer	avg = 1081.964 +/- 258.250	[600.000 ; 1500.000]	0
regular	KFT13x1(2) VARIANCE	integer	avg = 1075.268 +/- 272.947	[600.000 ; 1500.000]	0
regular	KFT15(4) VARIANCE	integer	avg = 1075.179 +/- 262.607	[600.000 ; 1500.000]	0

Nach diesem achtwöchigen Experiment beschließen der Betriebsleiter, die Meister und die Laborassistenten, Geschäftsregeln zu modellieren, mit denen die

Sensor- und Aktorwerte kontinuierlich geprüft werden. Mit diesen Regeln wird beispielsweise eine Plausibilitätsprüfung der von den Sensoren in aufeinander folgenden Einheiten registrierten Werte durchgeführt. Darüber hinaus suchen die Regeln automatisch nach ungewöhnlichen Mustern wie der oszillierenden Ventilöffnung in Abbildung 10.

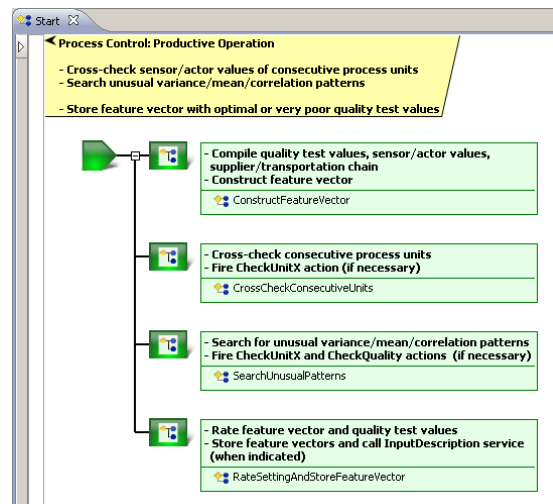
Abbildung 10: Öffnung des Ventils KFT17(V) nach einem Neustart des Prozesses (geglättet)



Zur weiteren Verbesserung des Prozesses und zum Überwachen von Qualitätsproblemen protokollieren die Regeln außerdem kontinuierlich Merkmalsvektoren mit (a) optimalen/beinahe optimalen und (b) schlechten Qualitätstestwerten. Diese können dann z. B. alle vier bis sechs Monate von einem Team von Qualitätsexperten inspiziert werden. Siehe Abbildung 11 zur Veranschaulichung.

Obwohl das Prozessleitsystem diese Konsistenzprüfungen im Prinzip (zumindest teilweise) selbst durchführen und die Merkmalsvektoren beproben könnte, wird es mit der Verwendung von Geschäftsregeln möglich, unmittelbar angepasste Tests laufen zu lassen: Wenn die Betriebsmeister beispielsweise entscheiden, eine bestimmte Einheit oder einen Prozessschritt zu prüfen, können sie die Geschäftsregeln unabhängig von IT-Beratern anpassen und das Experiment in kürzester Zeit selbst anstoßen.

Abbildung 11: Regelmäßig ausgeführtes Geschäftsregelmodell zur Qualitätssicherung (Auszug)



**Zusammenfassung:** Sie können sich des Eindrucks nicht erwehren, dass die Analyse dem Stochern im Nebel gleichkommt? Irgendwie haben Sie Recht: Wie das Beispiel zur Prozesssteuerung zeigt, kann die Datenanalyse wesentlich komplexer ausfallen als die Berechnung eines Klassifikators (vgl. Direktmarketing-Beispiel). Wenn Sie allerdings im Vorfeld Hypothesen entwickeln, die sie mit der Analyse letztlich untersuchen wollen, dann lichtet sich der Nebel meist schnell. Auf der Basis dieser Hypothesen ist es dann auch möglich zu entscheiden, welche Data Mining-Methoden verwendet werden müssen, welche Modelle mit Geschäftsregeln erstellt werden können und wie beide Konzepte möglichst vorteilhaft miteinander kombiniert werden.

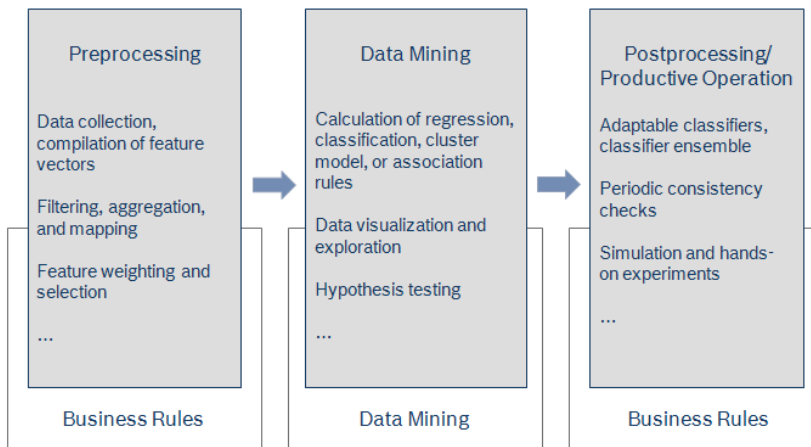
- entscheiden, wie die Geschäftsregeln und die Ergebnisse der Data Mining-Schritte
- integriert werden sollen.

Die Erfahrung zeigt, dass Sie Ihren **speziellen Anwendungsfall** sowohl mit einem Geschäftsregelexperten als auch mit einem Data Mining-Experten diskutieren sollten. Vermutlich haben Sie erst danach ein Gefühl dafür, was mithilfe von Geschäftsregeln modelliert werden kann und welche Data Mining-Methoden für die Datenanalyse verwendet werden sollten.

## Fazit

Wie die beiden Szenarien zeigen, befassen sich Geschäftsregel- und Data Mining Ansätze mit jeweils unterschiedlichen Aufgabenstellungen: Während Data Mining dabei hilft, Daten zu durchforsten und noch unbekannte Muster zu ermitteln, helfen Geschäftsregeln dabei, die notwendigen Datenverarbeitungsschritte (Vor- und Nachverarbeitung) durchzuführen und bekannte Parameter zu berechnen. Viele reale Anwendungen können, wie in diesem Beitrag dargelegt, von einer Kombination beider Konzepte profitieren. Eine Darstellung der hierfür notwendigen Schritte finden Sie in Abbildung 12.

Abbildung 12: Integration des Geschäftsregel und Data Mining-Ansatzes (Prozess)



Wenn Sie eine solche Kombination entwerfen möchten, müssen Sie

- ermitteln, was Sie über Ihre Daten wissen und was Sie nicht wissen:
  - Ersteres können Sie mithilfe von Geschäftsregeln abbilden.
  - Zweiteres können Sie mit Data Mining-Methoden berechnen.
- definieren, wie die Daten untersucht werden sollen, d. h., welche Data Mining-Methoden Ihnen helfen können, die relevanten Muster zu finden.

## Danksagungen

Wir bedanken uns bei Andreas Müller, Jan Trnka, Dr. Jürgen Cramer und Stefan Schacht für ihre aufschlussreichen Kommentare zu den beiden in diesem Beitrag diskutierten Szenarien. Darüber hinaus danken wir Eric Düll für das Modellieren der Geschäftsregeln.

## Literaturhinweise und Software

Au, T., Li, S., & Ma, G. (2003). Applying and Evaluating Models to Predict Customer Attrition Using Data Mining Techniques. *Journal of Comparative International Management*, 6(1).

Hay, D., Healy, K. A., & Hall, J. (2000). *Defining Business Rules - What Are They Really? The Business Rules Group. Final Report Version 1.3*. Abgerufen am 11. Januar 2001 von [http://www.businessrulesgroup.org/first\\_paper/br01c0.htm](http://www.businessrulesgroup.org/first_paper/br01c0.htm).

Tulyakov, S., Jaeger, S., Govindaraju, V., & Doermann, D. (2008). Review of Classifier Combination Methods. In: S. Marinai & H. Fujisawa (Hrsg.), *Studies in Computational Intelligence: Machine Learning in Document Analysis and Recognition*, S. 361-386.

Alle Geschäftsregeln wurden mit **Visual Rules** modelliert. Einzelheiten zum reichhaltigen Funktionsumfang und zu den intuitiven Nutzungsmöglichkeiten von Visual Rules finden Sie unter <http://www.visual-rules.de>. Wenn Sie mehr über das Modellieren von Geschäftsregeln und die Integration des Geschäftsregelkonzepts in Ihre Architektur erfahren möchten, empfehlen sich die folgenden Publikationen: *Voll durchstarten mit dem Regelprojekt* (Javamagazin 5/2010) und *Gut geregelt: Neue und bewährte Einsatzorte für Regeln in Softwarearchitekturen* (JavaSPEKTRUM 5/2010). Beide Beiträge sind im Bosch SI-Mediencenter erhältlich (siehe <http://www.bosch-si.de/medien-download.html>). Neuigkeiten und Wissenswertes finden Sie auch in unserem Technologie-Blog (siehe <http://blog.bosch-si.com/>).

Alle Experimente zum maschinellen Lernen wurden mit den Toolkits für maschinelles Lernen **KNIME** und **RapidMiner** durchgeführt. Details zum Funktionsumfang der beiden Toolkits und mehrere sehr anschauliche Nutzungsbeispiele finden Sie unter <http://www.knime.org/> und <http://rapid-i.com>.

## Informationen zur Autorin



Dr. Irene M. Cramer ([irene.cramer@bosch-si.com](mailto:irene.cramer@bosch-si.com)) ist promovierte Computerlinguistin. Bevor sie 2009 ihre Arbeit bei Bosch Software Innovations aufnahm, war sie Mitglied des International Post-Graduate College „Language Technology and Cognitive Systems“ (Saarbrücken/Edinburgh) und als Forscherin an den KI-Projekten SmartWeb (Universität des Saarlandes) und HyTex (Technische Universität Dortmund) beteiligt. Sie hat zahlreiche Beiträge und Monografien in den Bereichen Sprachtechnologie, Data Mining und Business Rules Management veröffentlicht.

© Bosch Software Innovations GmbH, 2011. Alle Rechte vorbehalten. Die Verbreitung oder Reproduktion dieser Publikation oder von Teilen hieraus für einen beliebigen Zweck oder in einer beliebigen Form bedarf der vorherigen ausdrücklichen schriftlichen Zustimmung der Bosch Software Innovations GmbH. Die in dieser Publikation enthaltenen Informationen können ohne vorherige Ankündigung überarbeitet werden. MLDS, Visual Rules und Work Frame Relations sind eingetragene Marken der Bosch Software Innovations GmbH. BOSCH und das Symbol sind eingetragene Marken der Robert Bosch GmbH, Deutschland. Einige der in diesem Beitrag verwendeten Produkt- und Firmennamen sind Marken und/oder eingetragene Marken. Sie werden ausdrücklich nur zu Referenzzwecken verwendet und sind, ungeachtet einer Kennzeichnung, Eigentum der jeweiligen Markeninhaber.