Data Mining – das etwas andere Eldorado

Funktionsweise und Einsatzmöglichkeiten

Erhält man ein persönliches Anschreiben, das die eigenen Wünsche genau trifft, kann es zwar purer Zufall sein, aber auch ein Ergebnis von «Data Mining» und «Big Data». Data Mining ist für Grosskonzerne eigentlich nichts Neues. Heute eröffnet die Computerisierung zusammen mit dem Internet jedoch auch kleinen Organisationen ungeahnte Möglichkeiten der Optimierung im umkämpften Weltmarkt. Aber was ist Data Mining und wo kann es eingesetzt werden?

Rudolf Tanner

Grosskonzerne nutzen von globalen IT-Firmen angebotene Data-Mining-Technologien. Autofirmen suchen zum Beispiel im Internet und auf sozialen Netzwerken nach Hinweisen, die auf Defekte an ihren Wagen schliessen lassen, bevor der Schaden epidemische Ausmasse annimmt. Data Mining (DM) ist besonders von Interesse für Firmen wie Telekomanbieter, Versicherungen und Krankenkassen, die ihre Kunden binden wollen. Aber auch für Mikrochiphersteller, die damit ihre Chipausbeute optimieren, oder andere Hersteller, die damit Prozessabweichungen vorhersagen und dadurch Ausschuss vermeiden. Die Pharmaindustrie und Medizinforschung ihrerseits nutzen diese Technologie bei der Risikobeurteilung, und Banken und Stromkonzerne für die Vorhersage von Aktienkursen bzw. des Stromverbrauchs. Viele Branchen könnten von DM profitieren.

Für KMU oder Zulieferbetriebe haben die Informationen in sozialen Netzwerken oder im Internet nicht unbedingt die gleiche Bedeutung wie für Grosskonzerne. DM kommt z.B. in KMUs zum Einsatz, um Nutzen aus den eigenen Daten z.B. für Qualitätssicherung oder «Business Intelligence» zu ziehen.

Zusammenhänge entdecken

Data Mining befasst sich mit der Identifikation von neuen, möglicherweise wertvollen und verständlichen Zusammenhängen und Mustern in bestehenden Daten. Man unterscheidet zwischen daten- und hypothesengetriebenen Ansätzen.

Bei Ersteren ist das Ziel, Muster von bisher unbekannten oder ungewöhnlichen Informationen (d.h. Ereignissen), wie beim Kreditkartenbetrug, in den Daten zu finden. Beim zweiten Ansatz baut man ein Modell, das die Daten beschreibt und Vorhersagen z.B. über den Verlauf des Aktienkurses oder eines Kundenverhaltens erlaubt. Ein Modell ist eine vereinfachte mathematische Funktion, die die realen Zusammenhänge eines Umstandes kennt und deshalb darüber Voraussagen machen kann. Diese Kenntnis wird dem Modell mittels Training an Beispieldaten beigebracht. Dazu bedient man sich verschiedener Methoden und Algorithmen, auf welche noch eingegangen wird.

Die dazu benötigten Trainingsdaten werden heute mit Gold aufgewogen und die Datenspeicher, die sogenannten «Data Warehouses», werden immer grösser, weil die Online-Firmen wie eShops und Suchmaschinen fleissig Daten sammeln. Bei Kleinfirmen begnügt man sich meist mit den Daten auf dem eigenen Server. Um die Statistik zu verbessern, ist es sinnvoll, bei der Auswertung auch archivierte Daten zu berücksichtigen.

Ein grösseres DM-Projekt hat verschiedene Akteure: den Kunden, der die Resultate nutzen will; den DM-Projektleiter, der das Projekt betreut; den DM-Analysten, der die in Prosa verfassten Kundenwünsche in technische Anforderungen, die später der Entwicklung von DM-Modellen dienen, übersetzt und den DM-Ingenieur, der das Modell entwickelt und zur Funktionsreife bringt. Die IT-Abteilung einer Firma ist ebenfalls involviert. Sie unterstützt den DM-Ingenieur, erstellt die nötigen Zusatz- und Hilfsprogramme, stellt den Datenzugang sicher, integriert, wenn notwendig, neue Hardwarekomponenten oder erweitert eine bestehende Applikation, um die DM-Resultate den Mitarbeitern zur Verfügung zu stellen.

Es gibt auch standardisierte DM-Methoden bzw. Industriestandards. Erwähnt sei hier nur der populärste und

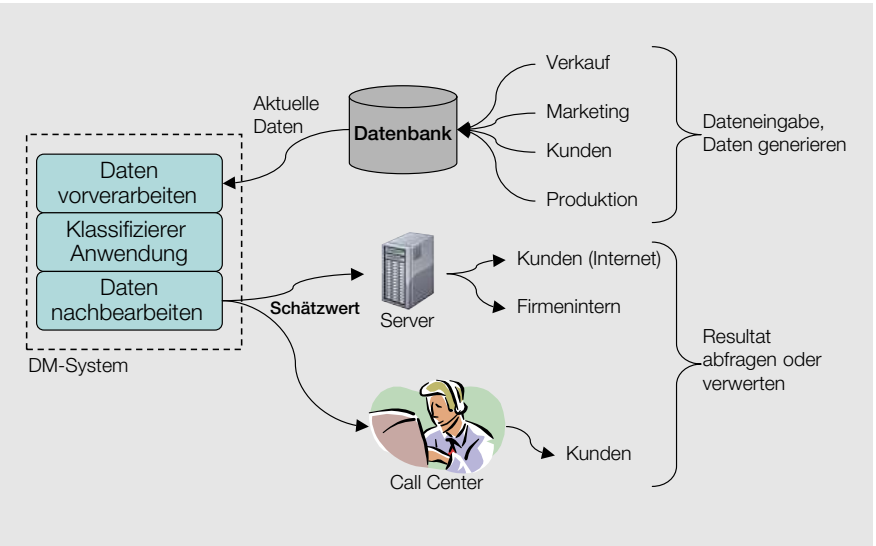

Bild 1 Ein Data-Mining-System im Einsatz.

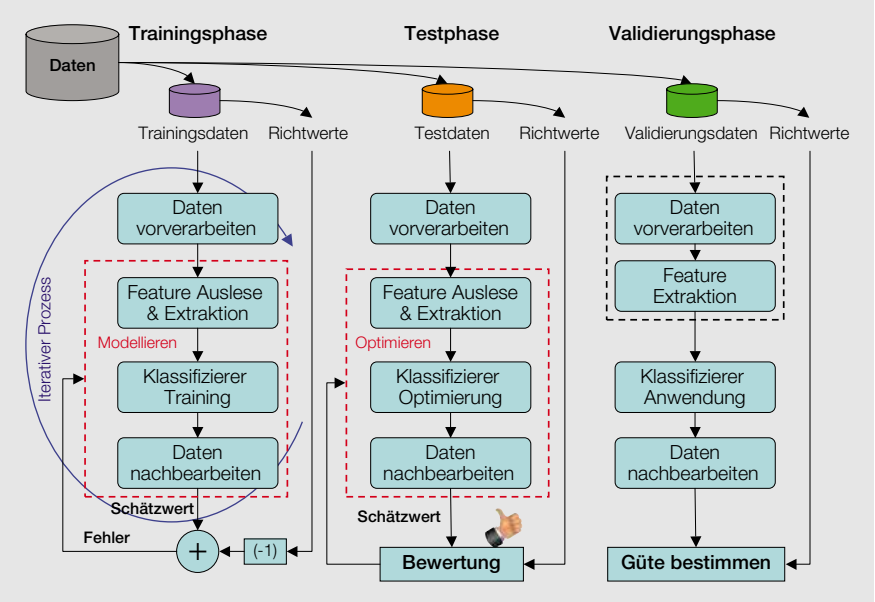

Bild 2 Im Entwicklungsstadium hat ein DM-System eine Trainings-, Test- und Validierungsphase.

nicht proprietäre Standard Crisp-DM, der einem EU-Projekt entstammt.

Natürlich ist DM nicht der «Holy Grail», für den es zurzeit gehalten wird. DM entbindet das Management nicht von der nötigen Kenntnis seines Geschäftes, des Marktes, der Bedeutung und Interpretation seiner Daten oder der Eigenschaften der verwendeten Methoden. Es gilt nach wie vor: «A fool with a tool is still a fool». Aber DM bietet einen klaren Mehrwert, wie zahlreiche Beispiele zeigen.

Daten und Business verstehen

Bevor man sich die Daten anschaut, sollte man wissen, was man erreichen will und welchen Mehrwert sich eine Abteilung, ein Arbeitsprozess oder das Management davon erhofft, d.h. welchem Geschäftszweck das DM dienen soll, ohne den Datenschutz zu verletzen! Zwei Beispiele sollen unterschiedliche DM-Zielvorgaben aufzeigen. Eine Versicherung will Mitarbeitern Richtlinien zur Beurteilung von Entscheidungen zur Verfügung stellen; oder eine Versicherung will frühzeitig erkennen können, welcher Kunde seine Police kündigt, damit sie diesen davon abhalten kann. Wenn der Geschäftszweck klar ist, dann kann man sich Gedanken über die technischen Ziele des DM-Systems machen.

Man unterscheidet zwischen der Entwicklungsphase und dem Einsatz eines DM-Systems. **Bild 1** stellt den Datenfluss eines DM-Systems dar. Basierend auf den aktuellsten Daten, berechnet das DM-System einen Schätzwert. Das kann z.B. eine Empfehlung oder ein Betrugswarnsignal sein. Dieser Schätzwert wird entweder dem Mitarbeiter im Call Center oder einer Serverapplikation zugeführt, um weiter verarbeitet zu werden. So soll z.B. automatisch eine Offerte erstellt, der Kunde beraten, oder firmeninterne Ereignisse ausgelöst werden.

Bild 2 zeigt die Entwicklung eines DM-Systems mit einer Trainings-, Test- und Validierungsphase. Auf die verschiedenen Blöcke wird später noch eingegangen. Grundsätzlich soll basierend auf Daten ein Schätzwert bestimmt werden.

Zum Beispiel soll aus dem Telefonierverhalten eines Kunden festgestellt werden, ob der Kunde sein Abonnement wahrscheinlich kündigen wird. Der Schätzwert beantwortet diese Frage. Damit ein DM-System das Muster in den Daten von solchen Kunden erkennen kann, besitzen die Daten Richtwerte (auch Zielvariable genannt), die besagen, was wirklich passiert ist, z.B. welcher Kunde sein Abo gewechselt hat. Weil die während der Entwicklung verwendeten Daten historisch sind, kennt man natürlich diese Richtwerte, d.h. in diesem Beispiel das echte Kundenverhalten. Der Richtwert wird mit dem Schätzwert verglichen (**Bild 2** unten links) und eine Schlaufe führt den resultierenden Fehlerwert an die relevanten DM-Blöcke zurück.

Wegen der existierenden Richtwerte kann das DM-System aus den Fehlern lernen. Das eigentliche Ziel des Trainings ist, mittels Algorithmen die gesuchten Muster in den Daten so im Klassifizierer abzubilden, dass der Klassifizierer anhand von neuen Daten den korrekten Schätzwert (mit einer akzeptablen Fehlerquote) generieren kann. Um diese Genauigkeit abschätzen zu können, ist neben der Trainingsphase noch eine Testphase, basierend auf unabhängigen (d.h. neuen) Daten, notwendig.

Daten vorbereiten

Die Daten müssen für DM verarbeitbar sein. Lose Blätter mit Zahlenreihen sind unnütz. Das DM ist deshalb wie geschaffen für eCommerce, weil die Daten schon in elektronischer Form vorliegen. Sonst muss man die Daten zuerst suchen,

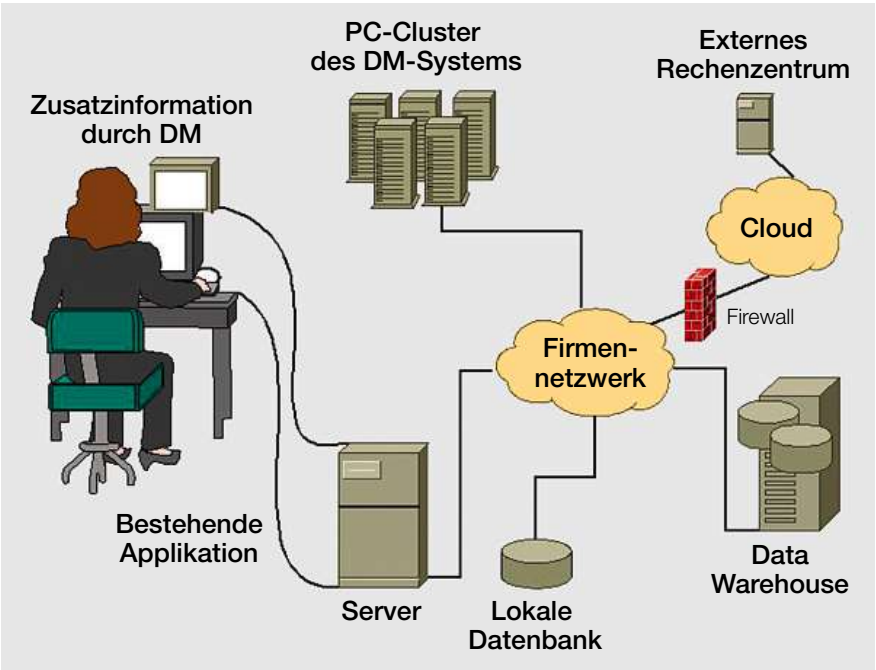

Bild 3 Möglicher Einsatz eines DM-Systems in einer Firma.

sammeln, aufbereiten und speichern. Eine Anpassung der firmeninternen Arbeitsprozesse könnte erforderlich sein und sollte im Einvernehmen mit allen Parteien geschehen, damit Zweck und Nutzen der Umstellung bekannt sind.

Das DM beinhaltet unter anderem den Aufbau einer Datenbank, die Entwicklung eines Modells sowie den Aufbau einer Wissensbank.

Das Modell erfordert:

- Akquisition von Daten zum Modellieren
- Reduktion der Anzahl von Variablen
- Erstellung eines passenden Modells
- Einsatz des Modells in der Firma
- Überwachung der Leistungsfähigkeit
- Versorgung des Modells mit neuen Daten
- Eruierung der Wertschöpfung des Modells

Bei der Reduktion der Anzahl von Variablen muss die Zielvariable bestimmt werden und die Variablen müssen statistisch analysiert werden. Zudem geschieht eine erste Vorselektion von Variablen. Erste Modellierungsversuche zur Identifikation der nützlichen Variablen finden statt.

Jeder dieser vier Schritte beinhaltet wieder Zwischenschritte, wie z.B. Herleitung verschiedener statistischer Werte, Erstellung von Grafiken und weiterer Analysen. Ausserdem müssen die Daten bereinigt werden, d.h. man muss Extremwerte (Ausreisser), fehlende Daten und spezifische Werte behandeln, weil handelsübliche Programme solche Daten u.U. nicht korrekt verarbeiten können.

Es steckt also einiges an Vorarbeit in einem Data Mining System. Die einfach anmutenden Arbeiten wie die Datenbeschaffung benötigen in der Praxis meistens mehr Zeit als angenommen. DM hat nichts mit «Hacking» oder wundersamer Eingebung zu tun, sondern verlangt Engineering.

Ein DM-System besteht im Kern aus einem Modell, welches durch bestehende Daten geformt und anhand neuer Daten regelmässig angepasst wird. Die Qualität der Daten bestimmt die Lernfähigkeit eines Modells. Die Qualität eines Modells hängt aber auch von dessen Architektur und dem gewählten Algorithmus ab. Hier ist Kreativität gefragt.

Modellieren

Das Erstellen eines Modells wird oft als die eigentliche Kunst im DM angesehen. Jedes gewählte Modell, z.B. ein neuronales Netz, hat eine grundlegende Struktur, die mittels Modellparameter konfiguriert (angepasst) werden muss. Der Einfluss jedes Modellparameters muss bekannt sein. Oft wird hier jedoch von Standardwerten ausgegangen, die die DM-Software bereitstellt und die dann mittels einer Optimierungsmethode lokal verändert werden, bis bessere Schätzwerte herauskommen. In der Regel setzt man mehrere komplementierende Algorithmen ein, weil deren Zusammenspiel bessere Resultate liefert. Dann besteht die Kunst in der optimalen Kombination der verschiedenen Lösungen.

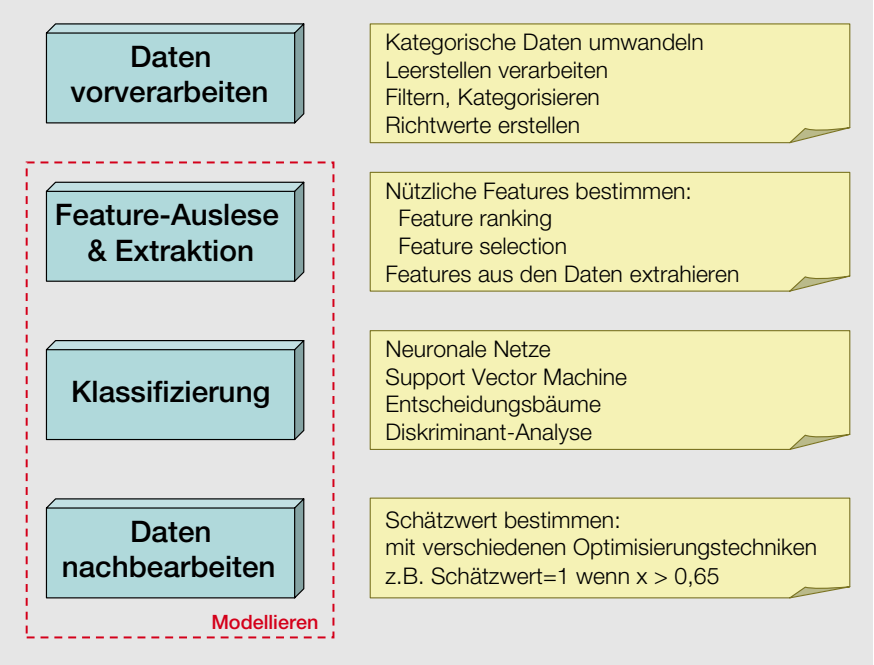

Bild 4 Prinzipieller Aufbau eines DM-Systems.

Beim Modellieren handelt es sich um eine Form des Trainierens, bei welcher der Raum aller möglichen Lösungen durch eine Funktion eingeschränkt wird. Ein Modell wird mittels bestehender Daten trainiert. Zur Modelloptimierung muss man eine Modellvariante beurteilen können. Dazu bedient man sich z.B. grafischer und tabularischer Methoden, wie statistischer Verteilungstabellen oder Kennlinien.

Die Modelloptimierung erfolgt in der Testphase (Bild 2) mit Testdaten (orange), welche nicht während der Trainingsphase verwendet wurden. Damit wird die Genauigkeit des Modells für neue Daten ermittelt, d.h. bewertet. Bei der Modelloptimierung wird auch die Dauer des Trainings bestimmt. Man kann Modelle auch falsch trainieren, z.B. wenn sich das Modell spezifische Muster in den Trainingsdaten merkt und die Informationen in den Daten nicht verallgemeinert.

Beurteilung

Zum Schluss wird in der Validierungsphase mit den Validierungsdaten (grün) die Güte des Modells bestimmt. Ein Durchschnittsvergleich z.B. der korrekt gemachten Schätzwerte zur Zahl aller Schätzwerte ist in der Praxis ungenügend. Man sollte auch die Rate der falsch negativen und falsch positiven Resultate kennen, da diese einen Kompromiss darstellen und unterschiedliche Kosten dafür anfallen. Die Auswirkungen, wenn beispielsweise ein Tumor nicht erkannt wird, können gravierend sein. Dann nimmt man lieber ein paar zusätzliche falsch positive Resultate in Kauf, als dies zu übersehen. Man beachte, dass die Validierung nicht für die Optimierung des Modells verwendet werden darf, sondern einer Endabnahme des DM-Systems entspricht. Die erfasste Güte kann als Abnahmekriterium für den Kunden dienen.

Wenn man mit der Leistung und Güte des Modells zufrieden ist, kann man das Modell in die Produktion (Einsatz) weitergeben. Das bedingt natürlich, dass das DM-System reif für den praktischen Einsatz ist.

Einsatz

Das Entwickeln kann v.a. bei kleineren Datenmengen auf einem PC durchgeführt werden. Obwohl die Modelle bei grösseren Datenmengen besser werden, sollten zu Beginn des Projekts nicht allzu viele Daten eingesetzt werden. Ein DM-System muss schnell, stabil, verlässlich, reproduzierbar, skalierbar und leistungsfähig sein.

Farbe	Gelb	Blau	Grau	Rot
Rot	0	0	0	1
Blau	0	1	0	0
Grau	0	0	1	0
Rot	0	0	0	1
Rot	0	0	0	1
Gelb	1	0	0	0

Tabelle 1 Beispiel einer Datentransformation.

Anwendung	Methoden
Datenerhebung Abhängigkeiten	Statistische Bewertung: Varianz, Mittelwert, Median, Korrelation, Min./Max.-Werte, Histogramm, Verteilung
Klassenzuteilung	Genetische Algorithmen Entscheidungsbäume Diskriminant-Analyse Support Vector Machine Regelbasierte Folgerung
Voraussage	Neuronale Netze Entscheidungsbäume Regressions-Analyse

Tabelle 2 Erfolgsversprechende Methoden für mögliche Einsatzszenarien.

Neben den zahlreichen kommerziellen DM-Paketen existieren auch die kostenlosen Open-Source-Tools wie RapidMiner, KNIME, Weka, SciLab und R. Zur Datenverarbeitung und Bewirtschaftung können verschiedene andere Gratis-Tools genutzt werden (siehe Big Data, Map-Reduce, Hadoop, NoSQL, Hbase oder Cassandra und deren weiterführende Referenzen auf Wikipedia). Die Anwendung und die Leistungsfähigkeit des Tools muss bei der Wahl berücksichtigt werden. Ein Supermarkt wird u.U. eine DM-Analyse an allen Produkten regelmässig durchführen wollen und ein Mobilfunkanbieter will betrügerische (falsche) Telefonate in Echtzeit erkennen können. Beide brauchen deshalb eine effiziente Datenverarbeitung. In einem anderen Szenario begnügt man sich mit einem einmaligen Einsatz und bevorzugt ein intuitives oder kostengünstiges Tool.

Die Infrastrukturkosten halten sich ebenfalls je nach Zweck und Aufwand in Grenzen. Meistens genügt schon ein Laptop, und mit drei bis fünf PCs kann man schon ein leistungsfähiges Rechenzentrum (PC-Cluster) aufbauen, sofern diese Rechenkapazität gebraucht wird. Für eine gute Skalierbarkeit kann man auch Cloud Computing einsetzen und Rechenkapazität extern einkaufen. **Bild 3** zeigt die verschiedenen Hardwarekomponenten eines möglichen DM-Systems. Einer Mitarbeiterin werden die Ausgaben (Resultate) des DM-Systems auf einem zweiten Bildschirm angezeigt, weil z.B. das DM-System noch nicht in die bestehende Applikation integriert ist. Viele Tools erlauben für eine Integration in eine bestehende Applikation den Export eines DM-Modells in C/C++ oder Visual Basic.

Data Mining

Der Aufbau eines DM-Systems ist in **Bild 4** dargestellt. Es erklärt exemplarisch, welche Funktionen die Blöcke enthalten und wozu sie dienen.

In der Datenvorbereitung gibt es viele Aspekte zu beachten und abzuklären. Der «künstlerische» Aspekt in DM behandelt die Wahl und Definition von Variablen, besonders wenn neue Variablen aus bestehenden abgeleitet werden. Zu viele Variablen führen zur Verwirrung im System und zu wenige erschweren die Vorhersagbarkeit. Daher muss der DM-Ingenieur die Variablen, die beim Modellieren verwendet werden sollen, in der Feature-Auslese bestimmen.

Bei der Umwandlung von kategorischen Textdaten wie Farbe, Nationalität oder Religion muss berücksichtigt werden, dass man das Spektrum des Modells nicht einschränkt. **Tabelle 1** zeigt, wie man die Variable «Farbe» mit 4 Ersatzvariablen darstellen kann.

Das Modell wird zwar durch die Ersatzvariablen besser angepasst, aber diese Überanpassung hat ihren Preis, denn das Modell wird weniger flexibel, wenn die Daten z.B. später andere Muster beinhalten. Das Modell hat sich nur die Trainingsdaten gemerkt, nicht aber deren Eigenschaften oder Gesetzmässigkeiten und kann deshalb auf neue Daten nicht korrekt reagieren. Das ist vergleichbar mit einem Schüler, der sich 20 Übungsaufgaben gemerkt hat, nicht aber z.B. die Gesetzmässigkeiten des Bruchrechnens. Neuronale Netze, Entscheidungsbäume und parametrisch statistische Algorithmen reagieren, im negativen Sinn, sensibel auf Überanpassung.

Unvollständige Listen von Variablen (mit Leerstellen), werden im Allgemeinen standardmässig verworfen. Das kann u.U. zu einem voreingenommenen Schätzwert führen. Um dies zu vermeiden, können unvollständige Listen repariert werden. Dabei ist zu beachten, dass die Information in den bestehenden Datenpunkten nicht verzerrt wird.

Gewichtung und Filter

Die Daten müssen eventuell gewichtet oder ausgeglichen werden. Gewisse Algorithmen profitieren davon, wenn z.B. Messdaten gemäss deren Genauigkeit gewichtet werden, oder alle Muster (häufige wie seltene) die Parameter des Modells gleich beeinflussen können. Damit stellt man sicher, dass der Algorithmus nicht nur die gängigen Muster erlernt, sondern andere Muster auch erkennt.

Mit einem Filter lassen sich triviale Werte oder Ausreisser eliminieren, um die Güte des Schätzwertes zu verbessern. Wenn die Daten Ausreisser enthalten, wirken diese wie Störsignale. Wenn das Normalverhalten modelliert werden soll, dann verschlechtern Ausreisser die Vorhersagbarkeit. Andererseits will man Ausreisser bewusst in den Trainingsdaten, wenn man nach Betrug in den Daten sucht.

Wenn die Richtwerte **(Bild 2)** nicht schon explizit vorhanden sind, dann müssen sie auch generiert werden, z.B. aus dem bestehenden Datensatz. Das lässt sich relativ einfach bewerkstelligen, wenn es sich um Informationen wie abgesprungene Kunden oder Maschinendefekte handelt, die in den Daten leicht ersichtlich sind. Kompliziert wird die Sache, wenn sich die Richtwerte nicht direkt aus dem Zweck des DM-Systems, bzw. der verfügbaren Daten, herleiten lassen. Dann ist Kreativität gefragt.

Was ist relevant?

Ein Ziel von Feature-Auslese ist, nur die relevanten Variablen herauszufinden, weil dies die Komplexität und Verarbeitungszeit reduziert. Man bedient sich dazu meist zweier Methoden, «Feature Ranking» oder «Subset Selection». Variablen können alles Mögliche sein, wie z.B. Produkt, Preis, Kaufdatum. Eine erste Auswahl von Variablen beinhaltet z.B. jene, die einen Gini-Index grösser als 0,5 haben, da deren Werte relativ ungleich verteilt sind. **Tabelle 3** zeigt ein weiteres einfaches Beispiel, wie man Variablen mittels der Partial Least Squares Regression (PLSR) Methode eliminiert.

Das Modellieren tangiert alle Blöcke, deren Eigenschaften und Verhalten während der Trainingsphase erheblich

Variable	PLSR
A	0,64056
B	0,22803
C	0,13118
D	0,00021

Tabelle 3 In diesem Beispiel würde Variable D verworfen, weil sie wenig zur PLSR-Summe beiträgt.

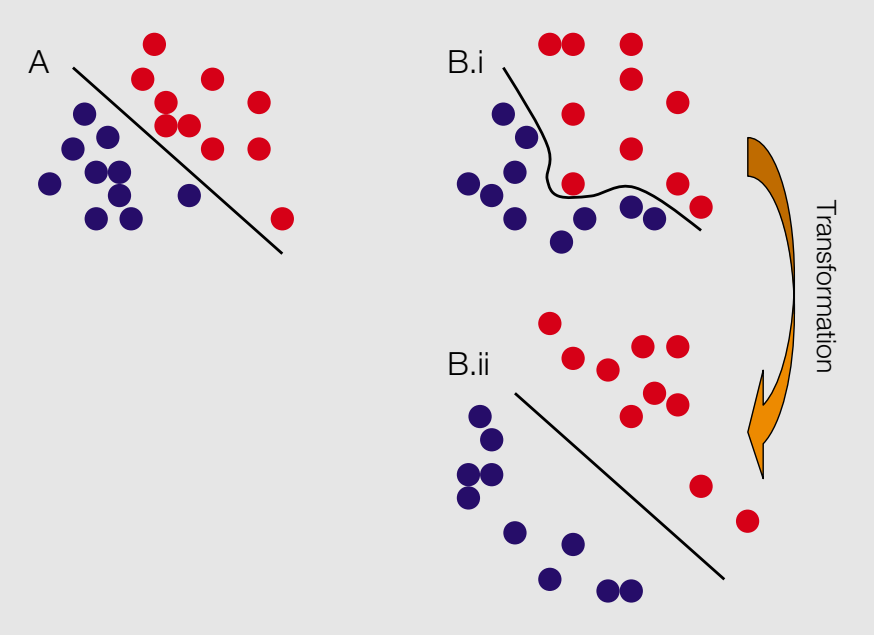

Bild 5 Die SVM wandelt ein nichtlineares Klassifizierungsproblem mittels einer Transformation in ein einfaches lineares Problem um.

manipuliert werden. In der Testphase sollten dann dieselben Blöcke nur noch optimiert werden. In **Bild 4** wurden zur besseren Übersichtlichkeit Feature-Auslese & Extraktion und Klassifizierung getrennt dargestellt. In der Praxis bilden beide Blöcke oft eine verschachtelte Einheit. Der Unterschied zwischen Feature-Auslese und Klassifizierung lässt sich vereinfacht am Beispiel eines Tiefpass-Filters erklären. Zuerst muss man das Filter konstruieren (auswählen), erst danach kann man damit ein elektrisches Signal filtern (klassifizieren).

Der ganze Arbeitsvorgang des Modellierens ist ein iterativer Prozess. Dabei wird auch berücksichtigt, dass das Modell nicht überangepasst wird. Dazu bedient man sich z.B. Algorithmen wie der Vergleichsprüfung oder der «Optimum Stopping Time». Unabhängig von den eingesetzten Methoden sollte man die verschiedenen Resultate vergleichen, um eine abschliessende Selektion der Variablen zu treffen.

Der Klassifizierungsblock wird nun in der Trainingsphase mit den Daten gespiesen und trainiert, bis der Schätzwert dem Richtwert am Nächsten ist. Dazu gibt es eine Reihe von Techniken. Man unterscheidet generell zwischen der klassischen statistischen Analyse – Gruppenanalyse («cluster analysis»), Diskriminantanalyse und Regressionen (Ausgleich) – und sog. «Learning Machines» wie Neuronalen Netzen, Entscheidungsbäumen und genetischen Algorithmen.

Die Wahl der richtigen Methode beruht teilweise auf Intuition und Erfahrung. Aber in den letzten zehn Jahren hat die DM-Gemeinde auch viele Praxiserfahrungen und Empfehlungen publiziert. **Tabelle 2** listet ein paar Anwendungsbeispiele und passende erfolgsversprechende Klassifizierungsmethoden auf.

Stellvertretend für alle Klassifizierungsmethoden soll am Beispiel einer Support Vector Machine (SVM) die Funktionsweise einer Klassifizierung erklärt werden. Wir gehen von einem Klassifizierungsproblem mit 2 Variablen und 2 Klassen (rot, blau) aus. Jeder Datenwert kann dann als farbiger Punkt in 2-D (x,y) dargestellt werden. **Bild 5** (A) zeigt ein lineares Klassifizierungsproblem, bei dem eine Gerade die roten von den blauen Punkten teilt. **Bild 5** (B.i) zeigt ein Szenario, bei dem die beiden Klassen nur mittels einer komplizierten nichtlinearen Linie separiert werden können.

Eine SVM bedient sich sog. Kernels (mathematischer Funktionen) und führt eine Transformation, vom Eingangsraum (B.i) zum Eigenschaftsraum (B.ii), durch. Das Elegante daran ist, dass die Daten im Eigenschaftsraum nun linear separierbar sind. Dies erlaubt den Einsatz von einfa-

Résumé

L'exploration de données: un eldorado quelque peu différent

Fonctionnement et possibilités d'utilisation

L'exploration de données se consacre à l'identification, dans les données commerciales existantes, de rapports et de modèles qui pourraient se révéler précieux. Il existe deux approches différentes qui reposent soit sur des données, soit sur des hypothèses. Dans le premier cas, l'objectif consiste à identifier dans les données des échantillons d'informations jusqu'ici inconnues ou inhabituelles (à savoir des événements), telles que celles recueillies lors de l'utilisation frauduleuse de cartes de crédit. Dans la seconde approche, un modèle décrivant les données est créé afin de pouvoir effectuer des prédictions, notamment quant à l'évolution du cours des actions ou du comportement d'un client. Ce modèle est ensuite optimisé par le biais d'entraînements sur des données qui font office d'exemples. Ces dernières valent de l'or à l'heure actuelle et la taille des mémoires, également appelées « entrepôts de données », ne cesse d'augmenter.

Les petites sociétés se contentent la plupart du temps des données stockées sur leur serveur. En principe, l'exploration de données peut s'appliquer à l'ensemble des secteurs d'activité et des sociétés, même aux PME, et toutes les branches sont en mesure d'en tirer bénéfice.

Un projet d'exploration de données réunit différents acteurs: un client, un chef de projet, un analyste et un ingénieur spécialisés dans l'exploration de données, sans oublier le département Informatique du client. Les données doivent être tout d'abord retravaillées afin de pouvoir être exploitées dans ce contexte. Ce travail prend généralement un temps considérable. C'est la raison pour laquelle l'exploration de données est particulièrement adaptée au commerce électronique.

Outre les nombreuses offres commerciales en la matière, une multitude d'outils open source est également disponible.

Un modèle est alors créé et évalué en trois étapes: l'entraînement, l'essai et la phase de validation. Une fois la mise en service d'un système d'exploration de données effectuée au sein de la société se pose encore la question de sa maintenance.

L'avenir révélera sur le long terme si l'avantage commercial généré par l'exploration de données peut être conservé lorsque tout le monde fera la même chose et utilisera des algorithmes mathématiques identiques. En effet, l'exploration de données ne peut pas remplacer l'innovation. La créativité restera un atout essentiel. No

chen und effizienten linearen Techniken zur Trennung und Unterscheidung der zwei Datenklassen.

Die Zukunft wird noch einiges an neuen Methoden zum Modellieren hervorbringen, denn das Data Mining von Audio, Bild und Videodaten stellt uns vor neue Herausforderungen.

Am Schluss des Modellierens werden die vom Klassifizierer ausgegebenen Werte bei Bedarf nachbearbeitet. Das sind üblicherweise relativ einfache Verfahren, bei denen z.B. ein Schwellwert gesucht wird, der bestimmt, ob der Schätzwert eine 0 oder 1 ist.

Nach der Inbetriebnahme eines DM-Systems in der Firma drängt sich die Frage des Unterhalts auf. Die Lebensdauer eines DM-Systems hängt von der Anwendung und vor allem vom Geschäftsumfeld ab. Finanzorganisationen mussten z.B. ihr DM-System seit 2007 regelmässig überarbeiten, weil sich die ökonomischen Gegebenheiten änderten. Wer Daten der Qualitätssicherung verarbeitet, der kann wahrscheinlich sein System jahrelang nutzen. Wer Daten verarbeitet, die Modetrends unterliegen oder aus sozialen Netzwerken stammen, wird periodisch sein DM-System anpassen müssen (d.h. das Modell nachtrainieren), um den Veränderungen gerecht zu werden.

Schlusswort

Zurzeit wird viel Geld in Data Mining investiert, sei es in Firmen oder bei Anbietern, denn es besteht noch Optimierungsbedarf und -potenzial. Zum Beispiel macht es Sinn, den CO_2-Ausstoss oder die Abfallproduktion durch optimierte Herstellungsprozesse zu reduzieren. Es wird sich aber auch noch zeigen, ob ein Marktvorteil durch DM aufrechterhalten werden kann, wenn alle dasselbe tun und die gleichen mathematischen Algorithmen einsetzen. Schliesslich ist DM kein Ersatz für Innovation. Dann ist wieder Kreativität gefragt.

Literatur

- P. Janert, Data Analysis, O'Reilly, 2010.
- R. Nisbert et al., Handbook of Statistical Analysis, AP, 2009.
- H. Marmanis, D. Babenko, Algorithms of the Intelligent Web, Manning, 2009.
- R. Duda, P. Hart, D. Stork, Pattern Classification, Wiley, 2001.

Angaben zum Autor

Dr. **Rudolf Tanner** ist Gründer der iCloudius GmbH und befasst sich seit 20 Jahren mit Algorithmen. Er arbeitete vorher als F&E-Abteilungsleiter.
iCloudius GmbH, 9478 Wartau
rt@icloudius.ch