



Self-Acting Data Mining
verspricht deutlich
„schlankere“ Datenanalysen.

Datenanalyse mit Autofokus

Data-Mining-Analysen gelten als komplex, nur von Spezialisten beherrschbar und daher als teuer. Der neue Analyse-Ansatz Self-Acting Data Mining verbindet die Schnelligkeit von OLAP-Analysen mit der Suchraumgröße des traditionellen Data Mining.

KOMPAKT

- ▶ Herkömmliche Datenanalysen sind aufwändig und komplex
- ▶ Neues Paradigma verkürzt Prozess der Knowledge Discovery
- ▶ Parametrisierung der Analysemodelle wird automatisiert

HOCHWERTIGE Informationen haben sich in vielen Branchen zu einem strategischen Wettbewerbsfaktor entwickelt. Insbesondere im Customer Relationship Management besteht ein immenser und stetig steigender Bedarf an entscheidungsrelevantem Wissen. Der Rohstoff für dieses wertvolle Wissen liegt vielfach in Form von unausgewerteten Daten

vor – er muss lediglich zutage gefördert und zu Wissen veredelt werden.

Unbegrenzt Datenwachstum

Dies ist jedoch bei den heute üblichen Datenmengen eine zunehmend schwierige Aufgabe: Das Data Warehouse von Wal-Mart ist mittlerweile auf über 700 Terabyte angewachsen – und das, obwohl die Transaktionsdaten bereits nach zwei Jahren gelöscht werden. Auch Banken, Versicherungen und Telekommunikationsanbieter verfügen über sehr große Datenbestände, die rasant steigen: Die Winter Corp. hat ermittelt, dass die größten Data-Warehouse-Systeme ihren Umfang alle zwei Jahre verdreifachen.

In vielen Unternehmen liegt das

Potenzial dieser Daten brach: Sie werden kaum ausgewertet, sondern verursachen als „Datenfriedhöfe“ Kosten in erheblicher Höhe. „Wir ertrinken in Informationen, aber uns dürstet nach Wissen“, sagte Zukunftsforscher John Naisbitt treffend. Dieses Missverhältnis kann nur durch hochautomatisierte Datenanalysen ausgeglichen werden.

Konfirmative Datenanalyse: Mit der Spitzhacke auf Goldsuche

Gibt es eine konkrete Vorstellung davon, welche Frage durch eine Datenanalyse beantwortet werden soll, werden hypothesengeleitete Datenanalysen durchgeführt. Beispiele sind:

- „Wie viele und welche Kunden ha-

ben neben einem Tagesgeldkonto auch einen Aktiensparplan abgeschlossen?“

- „Welche zehn Kunden haben sich im letzten Jahr am häufigsten beschwert?“
- „Welche Artikel haben den negativen Deckungsbeitrag des Kunden 21493 verursacht?“

Typische Methoden der konfirmativen Datenanalyse sind statistische Analysen, Standardberichte, Ad-hoc-Datenbankabfragen oder Online Analytical Processing (OLAP) mit multidimensionalen Datenbanken. Konfirmative Analysen sind die Spitzhacken im Datenbergbau: Sie liefern nur dann gute Ergebnisse, wenn der Anwender genau weiß, wo er nach interessanten Wissensschätzen suchen muss. Dieser Ansatz liefert nur Antworten auf bereits bekannte Hypothesen; der Suchraum ist stark eingeschränkt. Zudem müssen solche Analysen weitgehend

manuell durchgeführt werden – die heute üblichen Datenmengen lassen sich so nicht umfassend auswerten.

Explorative Datenanalyse: Mit Data Mining das Terrain sondieren

Der Data-Mining-Ansatz verspricht Abhilfe: Er ist auf große Datenmen-

genzung durch Annahmen vorgegeben ist. Aufgrund seiner zur konfirmativen Analyse komplementären Charakteristik eignet sich Data Mining besonders gut als vorgeschaltetes Verfahren: Man spürt damit zunächst explorativ Zusammenhänge auf, kann daraufhin Hy-

Konfirmative Analysen sind die Spitzhacken im Datenbergbau.

gen spezialisiert und kann deutlich offenere Fragestellungen untersuchen:

- „Welche Charakteristika kennzeichnen unsere Stammkunden?“
- „Welche Produkte verkaufen sich besonders gut zusammen?“
- „Welchen Artikel sollte ich im nächsten Monat welchen Kunden bevorzugt anbieten?“

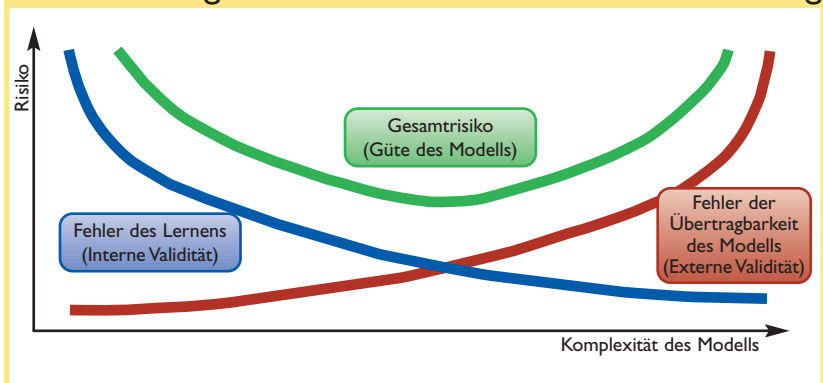
In diesen Fällen ist der Suchraum wesentlich komplexer, da keine Be-

pothesen bilden und diese dann zum Beispiel mittels OLAP gezielt näher untersuchen.

Die Ineffizienz des KDD-Prozesses

Leider besitzen traditionelle Data-Mining-Analysen in der Praxis vor allem wegen des komplexen Ablaufs des Prozesses der Knowledge Discovery in Databases (KDD), in den Data-Mining-Analysen einge-

Abbildung 1: Automatisierung der Data-Mining-Modellbildung durch strukturierte Risikominimierung



bettet sind, gravierende Nachteile. Denn der KDD-Prozess, der unter anderem die wichtigen Aufgaben der Datenvorverarbeitung enthält, ist hochgradig ineffizient, da er

und müssen daher häufig mehrfach ausgeführt werden. Dies vervielfacht die Durchlaufzeit und damit die Kosten.

Für die Praxis wäre ein Analyse-

Im Self-Acting Data Mining kann die aufwändige Parametrisierung automatisiert werden.

nicht geradlinig abläuft, sondern nur über zahlreiche Umwege zum Ziel kommt: Die Aufgaben sind durch gegenseitige Abhängigkeiten extrem ineinander verschachtelt

ansatz wünschenswert, der die Schnelligkeit und einfache Erstellung von OLAP-Analysen mit dem großen Suchraum des traditionellen Data Mining kombiniert.

Self-Acting Data Mining: Das neue Paradigma der Datenanalyse

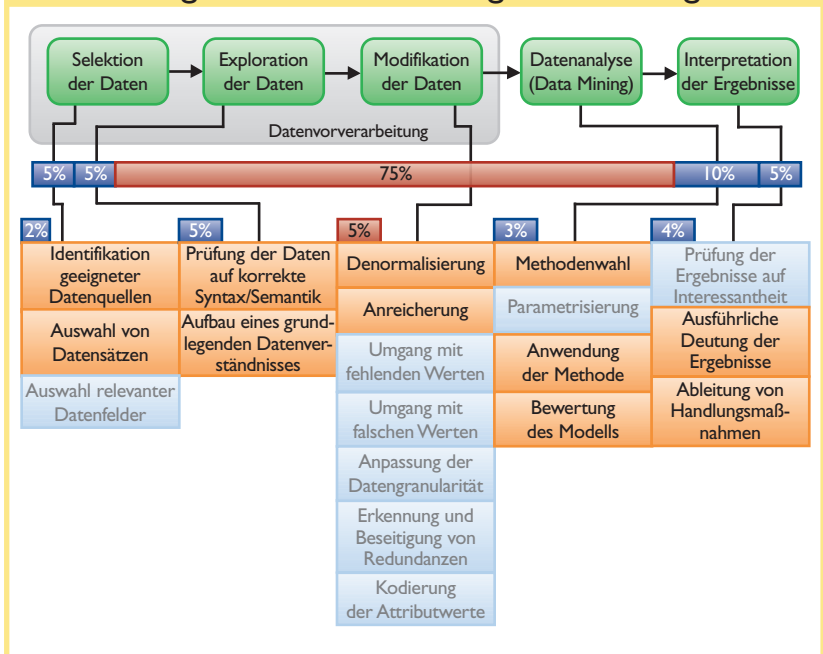
Self-Acting Data Mining bietet genau diese Kombination. Ein Beispiel für die Software-technische Umsetzung dieses neuen Analyse-Ansatzes ist das Analytic Framework von KXEN. Die darin verwendeten Verfahren machen sich die Erkenntnisse der strukturierten Risikominimierung zunutze. Das Grundprinzip: Ein Data-Mining-Modell ist stets ein Kompromiss zwischen der Gültigkeit der Ergebnisse im aktuellen Analysefall (interne Validität) und der Übertragbarkeit der Ergebnisse auf neue, unbekannte Daten (externe Validität).

Im Self-Acting Data Mining kann die aufwändige Parametrisierung der Verfahren automatisiert werden, indem sie in ein mathematisches Optimierungsproblem umgewandelt wird (Abbildung 1): Die Fehler der beiden Größen „Externe Validität“ (rote Kurve) und „Interne Validität“ (blaue Kurve) werden beide gleichzeitig minimiert. Die Güte des Modells insgesamt ist in der grünen Kurve dargestellt; deren Minimum stellt das minimale „Risiko“ der beiden Fehler dar und ist somit das optimale Modell. Over- oder Underfitting wird so vermieden.

Nicht nur die Data-Mining-Analyse allein, sondern auch zahlreiche weitere Teilaufgaben des KDD-Prozesses können durch Self-Acting Data Mining automatisiert werden, wie Abbildung 2 zeigt. Die Schritte des KDD-Prozesses sind grün dargestellt, die zugehörigen Teilaufgaben orange. Die lange Zeitleiste zeigt den sehr unterschiedlich verteilten Zeitaufwand der einzelnen Aufgaben im traditionellen Data Mining: Dort fallen beispielsweise etwa 75 Prozent der Gesamtzeit allein für die Modifikation der Daten an.

Der Zeitbedarf im Self-Acting Data Mining hingegen ist – maßstabsgetreu – jeweils separat direkt über den jeweiligen gelb unterlegten Teilaufgaben angegeben. Auffällig ist die deutliche Zeitersparnis, die

Abbildung 2: Der klassische KDD-Prozess im Vergleich mit Self-Acting Data Mining



hauptsächlich durch die Reduktion der Datenmodifikation von 75 Prozent auf fünf Prozent zustande kommt. Wie lässt sich dies erreichen? Einige der in der orange hinterlegten Tabelle aufgeführten Teilaufgaben sind ausgegraut, was bedeutet, dass diese im Self-Acting Data Mining automatisiert ablaufen oder sogar entfallen. Das Ergebnis ist ein sehr viel schlanker Prozess, dessen Zeitbedarf bei unter 20 Prozent des klassischen KDD-Prozesses liegt.

Selektion der Daten

Im klassischen Data Mining muss eine Auswahl der Datenfelder getroffen werden, die in die Analyse einbezogen werden sollen. Das führt dazu, dass der Grundansatz der explorativen, datengetriebenen Analyse insgesamt verwässert wird. Denn welche Attribute etwa in einem Prognosemodell die Kündigungswahrscheinlichkeit eines Kunden beeinflussen, lässt sich ex ante nur schwer bestimmen. Genau dies soll schließlich durch die Data-Mining-Analyse herausgefunden werden. Zum anderen ist die Auswahl relevanter Datenfelder auch eine für die Prozesseffizienz kritische Aufgabe, weil sie häufiger Auslöser für den Start einer neuen Iteration des Gesamtprozesses ist.

Dies alles entfällt im Self-Acting Data Mining vollständig, denn ne-

ben der beschriebenen „Autojustierung“ können die Verfahren der strukturierten Risikominimierung mehrere tausend Variablen verkraften, ohne dass die Rechenzeit inakzeptable Höhen erreicht. Auch werden durch das Einbeziehen von irrelevanten Attributen die Ergebnisse nicht beeinträchtigt.

Modifikation der Daten

Im traditionellen Data Mining muss viel Zeit auf den Umgang mit fehlenden und falschen Werten verwendet werden. Im Self-Acting Data Mining werden fehlende Werte und Ausreißer automatisch erkannt, in eigene Wertklassen gruppiert und gehen explizit in die Analyse mit ein.

Auch die Beseitigung von Redundanzen ist nicht mehr notwendig; Verfahren des Self-Acting Data Mining können auch mit stark untereinander korrelierenden Variablen umgehen. Weil herkömmliche Verfahren in diesen Fällen unbrauchba-

Der Autor



Peter Neckel arbeitet seit 2007 für das Analytischen- und Beraterhaus mayato. Bereits seit 2001 ist er als Berater in zahlreichen Praxisprojekten tätig, vorwiegend in den Bereichen CRM und Data Mining. Seine Erfahrungen bündelte er 2005 im Buch „Customer Relationship Analytics – Praktische Anwendungen des Data Mining im CRM“ (dpunkt.verlag).

re Ergebnisse liefern, müssen solche Korrelationen unter den Variablen zuvor zum Beispiel mithilfe einer Faktoranalyse ermittelt und beseitigt werden. Bei mehreren Hundert Variablen eine sehr aufwändige Arbeit, zumal auch dabei wiederum ein kompletter Analyseprozess zu durchlaufen ist.

Die Anpassung der Datengranularität, insbesondere für kontinuierlich verteilte Attributwerte wie Umsatz oder Deckungsbeiträge notwendig, ist eine erfolgskritische und im traditionellen Data Mining sehr mühselige Aufgabe, die im Self-Acting Data Mining entfällt: Die KXEN-Komponente

„Consistent Coder“ gruppiert beispielsweise die Attributausprägungen automatisch in angemessene Werte-Intervalle. Solche Diskretisierungen werden für jedes Attribut unter Berücksichtigung der individuellen Verteilung der jeweiligen Attributwerte automatisch berechnet. Die Zeitersparnis im Vergleich zum

traditionellen Vorgehen ist erheblich und steigt überproportional mit zunehmender Anzahl von Datenfeldern.

Datenanalyse (Data Mining)

Traditionelle Data-Mining-Verfahren müssen an zahlreichen Stellenschrauben mit viel Erfahrung und Fingerspitzengefühl parametrisiert werden. Durch die Automatisierung dieser Aufgabe können – neben der einmaligen Zeitersparnis – zahlreiche weitere Prozessdurchläufe gespart werden. Zudem entfällt die

Notwendigkeit eines Experten, so dass Self-Acting Data Mining die Potenziale explorativer Datenanalysen auch für Anwender aus den Fachabteilungen öffnet.

Bei weniger Kosten fast garantierter Informationsgewinn

Bisher ging man offenen analytischen Fragestellungen lediglich in ausgewählten Fällen nach – meist aus der Annahme heraus, den Kosten eines Data-Mining-Projekts stünde kein adäquater Informationsgewinn gegenüber. Durch Self-Acting

Data Mining hat sich die Kosten-Nutzen-Relation nun derart grundlegend verschoben, dass viele dieser Entscheidungen überdacht werden müssen.

Self-Acting Data Mining führt in der Praxis zu deutlich schlankeren Datenanalyseprojekten, die lediglich einen Bruchteil der Kosten traditioneller Data-Mining-Projekte verursachen. Durch den Zeitgewinn kann zudem sehr viel schneller auf aktuelle Ereignisse oder ein verändertes Kundenverhalten reagiert werden – ein zunehmend wichtiger Faktor. ◀

Data Mining bei den DEVK Versicherungen: Verbesserte Responsequoten und sechsstellige Budgeteinsparungen

Die DEVK Versicherungen versenden im Jahr mehrere Millionen Kundenbriefe zur Information über neue Produkte und mit Angeboten zur Ergänzung des bestehenden Versicherungsschutzes.

Zur Effizienzsteigerung hat neben der inhaltlichen Optimierung die Identifikation verschiedener Kundengruppen für eine zielgerichtete Ansprache einen sehr hohen Stellenwert. Für diesen Zweck wurde vor mehreren Jahren nach einem erfolgreichen Test die Beschaffung eines Data-Mining-Tools beschlossen. Hauptkriterien für die Toolauswahl waren sowohl das Preis-Leistungs-Verhältnis und die leichte Erlernbar- und Bedienbarkeit als auch die einfache Installation und Wartung des Programms im Fachbereich. Die Software *STATISTICA Data Miner* hat diese Kriterien am besten erfüllt, wurde beschafft und auf einer Workstation installiert.

Ein bemerkenswertes Ergebnis stand am Anfang der Analysen: Die Methode Classification and Regression Tree (C&RT) deckte auf, dass das Merkmal „Dima“, welches die bisherige Direktmarketing-Affinität abbildet, keine Differenzierung ermöglicht. Der Wert „0 = bisher keine Information vorliegend“ trat zu häufig auf. Dies war auch nach der Datenexploration mittels Histogrammen schon vermutet worden.

Die Übersetzung weiterer Analyse-Ergebnisse in Aussagen über die Wirkung der einzelnen Merkmalsausprägungen brachte zum Teil die Bestätigung von vermuteten Sachverhalten, aber auch neue Erkenntnisse. Dass zum Beispiel die Zahlungsweise (ZW) in diesem Fall die dominante Größe für das Responsever-

halten ist, war eine interessante, aber für Zusatzprodukte schon bekannte Tatsache. Es ist offensichtlich leichter, eine Ergänzung des Versicherungsschutzes „für 3 Euro im Monat“ als „für 36 Euro im Jahr“ zu kommunizieren.

Ebenso bekannt war bereits, dass Kunden mit mehreren bestehenden Verträgen viel häufiger zum Abschluss neigen als jene mit wenigen Verträgen. Unbekannt dagegen war die Tatsache, dass im letzteren Fall ältere Kunden eher zum Abschluss neigen als jüngere. Offensichtlich haben jüngere Kunden mit einem vorhandenen breiten Basisschutz schon entschieden, welche Produkte sie nicht oder noch nicht benötigen.

Die Differenzierung der Merkmalsausprägungen wurde in einen gemeinsamen Scorewert für die Responsewahrscheinlichkeit umgerechnet. Von insgesamt rund 330 000 Kunden wurden die 95 000 mit der höchsten Responsewahrscheinlichkeit ausgewählt.

Anhand der Rückläufer wurden die Vorhersagen überprüft: Im Vergleich mit dem im Vorjahr durchgeführten Mailing (gleicher Kundenkreis, kein erneutes Anschreiben) ergibt sich eine um 41 Prozent höhere Responsequote. Innerhalb der angeschriebenen Kunden liegt die Quote bei den Kunden mit dem geringsten Scorewert etwa 20 Prozent unter der Quote der Kunden mit dem höchsten Wert.

Die Anzahl der versendeten Briefe konnte bei nahezu gleicher Responsemenge gegenüber dem Vorjahr erheblich gesenkt werden. Die Einsparungen bei der DEVK betragen schon bei einer einzigen Aktion mehr als die Lizenzkosten für *STATISTICA Data Miner* für mehrere Jahre. dk

Der Anwender



„Die schnelle Einsatzbereitschaft und der Betrieb von *STATISTICA Data Miner* durch die Fachabteilung ermöglichte uns ein schlankes Projektvorgehen.“

Bruno Küpper, Referent für Kommunikationstechnik im Bereich Kundenbindung und Dialog-Marketing