

Data Mining wird anwenderfreundlicher

Beim Data Mining schreitet die Automatisierung voran und die Bedienung vereinfacht sich, sodass mehr und mehr Anwender differenzierte Analysemöglichkeiten nutzen können.

von peter neckel * | werner.fritsch@informationweek.de

Bei den Anbietern im Bereich Data Mining (DM) hat sich in den letzten Jahren einiges getan. Neben den klassischen DM-Suiten mit sehr vielen Funktionen und einer wachsenden Anzahl von Open-Source-Paketen ähnlichen Umfangs gibt es inzwischen zahlreiche spezialisierte DM-Werkzeuge für bestimmte Analyseaufgaben. Außerdem integrieren Hersteller von herkömmlicher Business-Intelligence-Software zunehmend DM-Funktionalität in ihre Produkte. Insgesamt sind hierzulande derzeit rund 150 DM-Angebote verfügbar.

autonom Regelmäßigkeiten und bis dato unbekannt Zusammenhänge in den Daten zutage fördern.

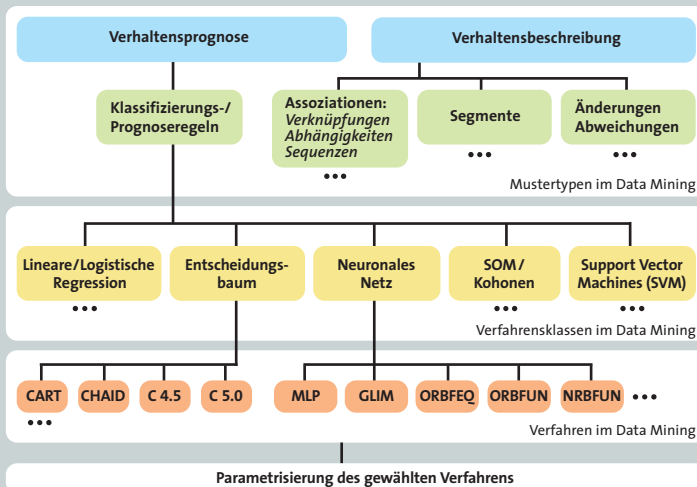
Zunehmende Automatisierung

Für die Studie »Data-Mining-Software 2009« hat das Beratungshaus mayato zwölf Angebote einem Funktionsvergleich unterzogen. Vier der Pakete durchliefen außerdem einen detaillierten Praxistest: die kommerzielle Suite SAS Enterprise Miner, die Open-Source-Software Rapid Miner, das Werkzeug für Self-Acting Data Mining KXEN Analytic Framework sowie die Software SAP NetWeaver 7.0 Data Mining Workbench.

Der Praxistest und die Funktionsbewertung orientierten sich am Ablauf des klassischen DM-Prozesses: Am Anfang steht die Selektion der Daten, die von der Software durch unterschiedliche Eingabeformate oder Funktionen zur Auswahl von Datensätzen unterstützt wird. Die Exploration der Daten ermöglicht dann die Berechnung von statistischen Kennzahlen oder die grafische Aufbereitung und Präsentation. Noch vor der eigentlichen Analyse liegt die umfangreichste Aufgabe in einem DM-Projekt: Die Modifikation der Daten. Dafür bieten viele Tools eine Reihe von Funktionen zur Zusammenführung, Anreicherung und Kodierung der Daten bis hin zur komplexen Berechnung zusätzlicher Kennzahlen. Spezialisierte DM-Werkzeuge decken meist nicht alle Mustertypen (siehe Kasten auf Seite 23) ab, sodass damit nur bestimmte Fragestellungen angegangen werden können.

Zu den Auswahlkriterien für DM-Software gehören hohe Stabilität, der unkomplizierte Umgang mit großen Datenmengen, die Automatisierung von Standardaufgaben, die Qualität und Interpretierbarkeit der Ergebnisse sowie nicht zuletzt einfache Bedienbarkeit ohne lange Einarbeitungszeiten. Auch auf die Effizienz des Analyseprozesses, die Anwendbarkeit der Programme sowie die daraus resultierenden Gesamtkosten kommt es an. Klar erkennbar: Der Stellenwert der Automatisierung nimmt zu. Denn mit anschwellenden Datenmengen und wachsendem Analysebedarf steigt der Anteil an Standardaufgaben wie Datenvorverarbeitung und Parametrisierung. Und durch deren Automatisierung können mehr Analyseergebnisse in kürzerer Zeit erzielt

Systematik des Data Mining



Im klassischen Data Mining gibt es Hunderte unterschiedlicher Verfahren, die einzeln parametrisiert werden müssen. Bei Self-Acting Data Mining entfällt dieser Aufwand.

Quelle: mayato

Die Datensammlungen der Anwenderunternehmen erreichen, nicht zuletzt wegen sinkender Speicherpreise, immer öfter Terabyte-Dimensionen. Um möglichst viel nützliche Information aus dem unüberschaubaren Datenvolumen abzuleiten, werden explorative Analyseansätze wichtiger. Sie sind im Gegensatz zu konfirmativen Analysen, bei denen von konkreten Annahmen ausgegangen wird, durch offene Fragestellungen gekennzeichnet. Die Tools sollen dabei möglichst

Muster des Data Mining

Mit DM-Methoden lassen sich vier unterschiedliche Typen von Mustern aufspüren.

1. Klassifizierungs- und Prognoseregeln dienen zum Beispiel der Vorhersage des Abwanderungsverhaltens oder der Kampagnenplanung, der Zielgruppenselektion oder der Kundenwertberechnung.
2. Assoziationen sind die Grundlage für Warenkorbanalysen sowie die Ermittlung von Cross- und Up-Selling-Potenzialen.
3. Segmente helfen bei der Markt- und Kundensegmentierung sowie der Analyse der Kundenentwicklung.
4. Mit Verfahren zur Bestimmung von Änderungen und Abweichungen lassen sich Datensätze ermitteln, die im Vergleich zu Referenzwerten stark abweichen. Ausreißer können zum Beispiel auf Betrugsversuche hinweisen.

werden. Dadurch verbessert sich die Effizienz des gesamten Analyseprozesses erheblich, da mehr Zeit für anspruchsvollere Aufgaben wie die Ergebnisinterpretation verbleibt – Tätigkeiten, in denen der menschliche Analyst den automatisierten Verfahren auf absehbare Zeit noch überlegen sein wird.

Die genannten Produkte wurden anhand von Testdatensätzen detailliert auf ihre Praxistauglichkeit geprüft, zunächst mittels einer überschaubaren Testdatei mit 30 000 Datensätzen und 15 Variablen. Insbesondere das Systemverhalten bei großen Datenmengen wurde anschließend durch Einlesen eines umfangreichen Volumens mit 100 000 Datensätzen und 450 Variablen gemessen. Die Palette der Benchmarkdaten enthielt eine Reihe typischer Datenqualitätsprobleme, mit denen DM-Tools umgehen müssen: beispielsweise korrelierende Variablen, fehlende Werte oder Ausreißer.

Große Performance-Unterschiede

Schon im ersten Durchgang ergaben sich erhebliche Unterschiede in den Laufzeiten, die sich bei der Verarbeitung größerer Datenmengen noch verstärkten. Dem SAS Enterprise Miner gelang insgesamt die beste Modellqualität. Die Bedienung der Suite geht trotz des großen Funktionsumfangs nach einer kurzen Eingewöhnungsphase relativ rasch von der Hand. Allerdings erfordert das Erstellen qualitativ hochwertiger Modelle Fingerspitzengefühl bei der Parametrisierung und ein gewisses Maß an Erfahrung.

Die Ergebnisqualität bei Rapid Miner fiel im Vergleich ab, insbesondere die Übertragbarkeit der Modelle erwies sich als gering. Ein erheblicher Nachteil besteht darin, dass Rapid Miner keine Möglichkeit bietet, die mit unterschiedlichen Verfahren erstellten Modelle automatisch zu vergleichen. KXEN lieferte eine gute Modellqualität und zeigte sich zudem bei der Berechnung als sehr schnell. Mit dem Ansatz des Self-Acting Data Mining weist KXEN das modernste Gesamtkonzept auf, bei dem die Datenvorbereitung nahezu automatisch abläuft. Beim Test mit großen Datenvolumina veränderte sich die Performance von KXEN Analytic Framework kaum. Bei anderen Produkten gestaltete sich hingegen bereits das Einlesen der Daten problematisch und erforderte

teilweise langwierige manuelle Eingriffe. In SAP Net-Weaver etwa musste für jedes Attribut ein separates Infoobjekt angelegt werden – bei 450 Variablen ein erheblicher Aufwand. Die Resultate und die Laufzeit waren hier guter Durchschnitt. Allerdings bietet SAP nur sehr rudimentäre Unterstützung beim Vergleich verschiedener Modelle. Vorteilhaft allerdings ist die Integration der DM-Funktionalität in die BI-Umgebung von SAP: Der Analyseprozessdesigner APD hat vollen Zugriff auf die mächtigen Funktionen der Datentransformation.

Die Berechnung deskriptiver Statistiken und der Aufbau grafischer Darstellungen nahm bei einigen Produkten sehr viel Zeit in Anspruch. Die Laufzeiten der eigentlichen DM-Analyse schließlich stiegen je nach Verfahren häufig überproportional an. Insbesondere Rapid Miner fiel auf: mit sehr langen Laufzeiten von mehr als drei Stunden bis zu Abbrüchen wegen Hauptspeicherüberlaufs. Mit KXEN nahm die Analyse des kompletten Datensatzes hingegen weniger als zehn Minuten in Anspruch.

Grafische Oberflächen

Im Anwendungstest zeigte sich, dass die Bedienung dank grafischer Benutzeroberflächen leichter von der Hand geht als noch vor wenigen Jahren. Data Mining kann dadurch von einem breiteren Anwenderkreis genutzt werden als früher. Damit weicht auch die Sonderstellung als Spezialistendisziplin allmählich auf. Diese Entwicklung wird zusätzlich dadurch getrieben, dass die explorative Datenanalyse für viele Unternehmen immer wichtiger wird. Dennoch erfordern insbesondere die DM-Suiten weiterhin einen erhöhten Einarbeitungsaufwand und fundiertes Hintergrundwissen, um vollen Nutzen aus dem großen Funktionsangebot ziehen zu können. Billig ist DM nach wie vor im Allgemeinen nicht, aber eine gewisse Flexibilisierung zeichnet sich ab, sodass auch Mittelständler nicht mehr darauf verzichten müssen. ■

* **Peter Neckel** ist Analyst bei dem auf Business Intelligence und Business Analytics spezialisierten Beratungshaus mayato in Berlin.