

Die Spreu vom Weizen trennt sich bei Analyseeffizienz und großen Datenmengen

Der Markt für Data-Mining-Software ist in den letzten Jahren spürbar in Bewegung geraten. Welche Stärken und Schwächen die aktuellen Tools im Praxistest zeigen, verrät die im Folgenden exklusiv vorgestellte Mayato-Studie.

KOMPAKT

- ▶ Test der Analyseeffizienz von Data-Mining-Tools
- ▶ Betrachtung von Data-Mining-Suite mit Open-Source-Lizenz
- ▶ Flexible Preismodelle erleichtern Unternehmen den Einstieg

DATA MINING hat sich inzwischen in vielen Branchen zu einem strategischen Wettbewerbsfaktor entwickelt, was in den letzten Jahren zu einer stetigen Ausweitung der Anwendungsgebiete geführt hat: Ob Fertigungsoptimierung im Automobilsektor, Risikomanagement und Betrugserkennung bei Banken und Versicherungen, oder – dank zunehmender Verbreitung der RFID-Technik – Supply-Chain-Analysen bei Logistikdienstleistern und im Handel.

Auch in der klassischen Data-Mining-Domäne Marketing und Vertrieb stehen analytische Fragestellungen nach wie vor ganz oben auf der Prioritätenliste: Im aktuellen CRM-Barometer 2008 sehen 40 Prozent der befragten Unternehmen die Kundenwertmodellierung und -segmentierung als die derzeit wichtigsten CRM-Themen an. Dies gilt nicht nur für Großunternehmen – im Mittelstand steigt die Verbreitung von Data-Mining-Anwendungen ebenso. Von dieser Entwicklung profitieren auch die Toolanbieter: Laut dem Business Application Research Center (BARC) wächst der deutsche Markt für Business-Intelligence-Software insgesamt voraussichtlich um 10 bis 12 Prozent jährlich bis zum Jahr 2012.

Steigende Anforderungen wegen vielfältiger Entwicklungen

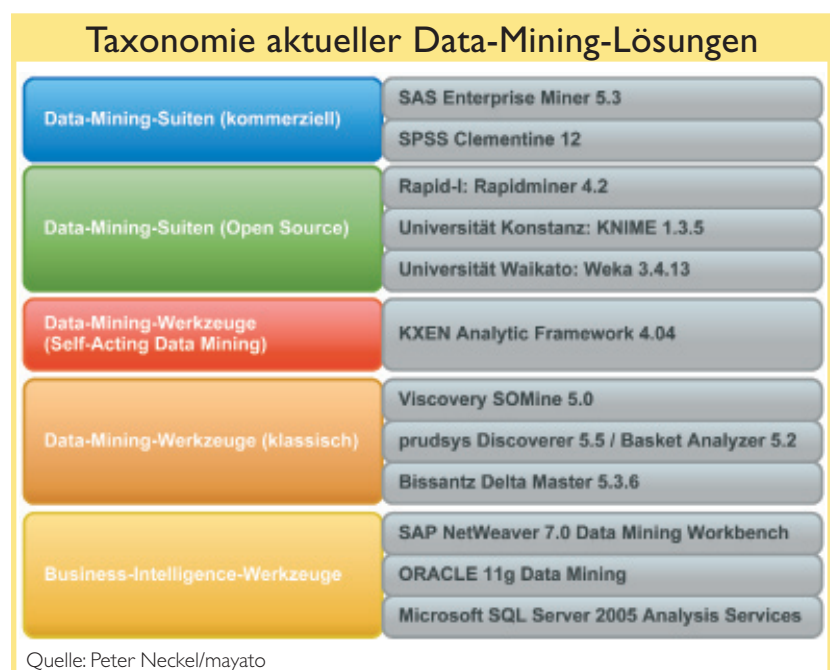
An heutige Data-Mining-Tools werden jedoch hohe Anforderungen gestellt. Durch den Zugriff auf zusätzliche Datenquellen vervielfachen sich die zu analysierenden Datenmengen in immer kürzeren Abständen. Die Anzahl der zu analysierenden Datensätze und die der sie beschreibenden Attribute steigen gleichermaßen. Gefördert durch inflationär sinkende Speicherpreise werden die Daten zudem in immer feineren Granularitätsstufen erhoben und ausgewertet. Immer neue Datenformate verkomplizieren zusätzlich die Verarbeitung.

Die Aberdeen Group und Rexer Analytics haben im Mai bzw. September 2008 unabhängig voneinander erhoben, welche Auswahlkriterien Unternehmen für Data-Mining-Soft-

ware zugrunde legen. Das Ergebnis: Hohe Stabilität, der unkomplizierte Umgang mit großen Datenmengen, die Automatisierung von Standardaufgaben sowie die Qualität und Interpretierbarkeit der Ergebnisse stehen ganz oben auf der Liste. Dazu fordern nicht nur die Anwender aus den Fachabteilungen eine einfache Bedienbarkeit ohne lange Einarbeitungszeiten.

Studiendesign für zwölf Data-Mining-Produkte im Test

Der Markt für Data-Mining-Software ist durch eine große „Artenvielfalt“ an Lösungen gekennzeichnet, was zu einer gewissen Unübersichtlichkeit führt und die Auswahl erschwert. mayato stellt auf der Nürnberger CRM-Expo 2008 seine aktuelle Data-Mining-Studie 2009 vor, die die Entscheidungs-



EXKLUSIV

findung erleichtert. Darin wurden zwölf ausgewählte Data-Mining-Tools einem Funktionsvergleich unterzogen; vier Produkte sind zudem anhand mehrerer Testdatensätze ausführlich auf ihre Praxistauglichkeit geprüft. Bewertet wurden unter anderem Bedienung, Stabilität, Systemverhalten bei großen Datenmengen, Dokumentation und die Gesamteffizienz des Analyseprozesses, in die Kriterien wie Geschwindigkeit, Automatisierungsgrad und Ergebnisqualität eingehen. Letzteres wurde anhand einer überschaubaren Testdatei (30 000 Datensätze, 15 Variablen) festgestellt; das Systemverhalten bei großen Datenmengen wurde durch Einlesen eines umfangreichen Datenvolumens (100 000 Datensätze, 450 Variablen) getestet.

Grundtypen von Data-Mining-Produkten

Die Grundtypen mit der zugehörigen Auswahl der in der Studie besprochenen Tools zeigt die Abbildung auf Seite 22: Die klassischen Data-Mining-Suiten (zum Beispiel von SAS oder SPSS) mit ihrem umfassenden Angebot an Datenvorverarbeitungsfunktionen und Data-Mining-Verfahren werden – mit durchaus vergleichbar mächtigem Funktionsumfang – inzwischen auch von Open-Source-Anbietern wie Rapid-I (*Rapidminer*), der Uni Konstanz (*KNIME*) oder der Uni Waikato (*Weka*) offeriert. Daneben gibt es die schlankeren Data-Mining-Werkzeuge mit reduzierter Funktionalität: Sie sind in der Regel auf bestimmte Anwendungsgebiete (beispielsweise Controlling) oder Analysefälle (etwa Prognose- und Klassifizierungsaufgaben) spezialisiert. Eine Sonderstellung in dieser Kategorie nimmt die softwaretechnische Umsetzung des Self-Acting Data Mining ein – dieser hoch automatisierte Ansatz kommt weitgehend ohne manuelle Datenvorverarbeitung und Parametrisierung aus.

Weiterhin haben zahlreiche Datenbank- und Business-Intelligence-Anbieter wie SAP, Oracle oder Microsoft in manchen Fällen recht umfangreiche Data-Mining-Funktionen, zum Teil gar selbst entwickelte Verfahren in ihre Produkte integriert – wie etwa den Oracle-Segmentierungsalgorithmus „Orthogonal Partitioning Clustering“.

Preismodelle auch für Mittelständler

Die Produktentscheidung will gut überlegt sein: Je nach Funktionsumfang und Nutzerzahl kann eine Client-Server-Lizenz durchaus mehrere 100 000 Euro in der Anschaffung sowie sechsstellige jährliche Wartungskosten verursachen.

Es geht aber auch bedeutend günstiger: Spezialisierte Data-Mining-Werkzeuge sind bereits für unter 10 000 Euro zu haben; für Open-Source-Lösungen fallen gegebenenfalls lediglich jährliche Supportgebühren im vierstelligen Eurobereich an.

Die Ausnutzung bereits vorhandener Datenbanklizenzen ist eine weitere Möglichkeit, Data-Mining-Projekte zu geringen Kosten durchzuführen: In den Enterprise-Lizenzen der Datenbanken von Oracle oder Microsoft (*SQL Server*) sind etwa die Data-Mining-Funktionen bereits enthalten. Die Preise liegen für beide Produkte bei rund 27 000 Euro für die Einprozessor-Lizenz; auch ein Upgrade von vorliegenden Standardlizenzen ist möglich.

Der Autor



Peter Neckel arbeitet seit 2007 für das Analytischen und Beraterhaus mayato mit Sitz in Berlin. Bereits seit 2001 ist er als Berater in zahlreichen Praxisprojekten tätig, vorwiegend in den Bereichen CRM und Data Mining.

Die Preismodelle der Anbieter werden zudem immer flexibler, was besonders dem Mittelstand entgegenkommt: Einige Data-Mining-Werkzeuge lassen sich etwa auf Monatsbasis zu Preisen im unteren vierstelligen Eurobereich mieten. Weiterhin muss nicht zwingend das Komplettprodukt bestellt werden; vielmehr können oft auch anwendungsspezifisch zusammengestellte Pakete an Funktionen oder gar einzelne Data-Mining-Methoden separat erworben werden. Derartige Angebote sind vor allem für Unternehmen interes-

sant, die zunächst eher überschaubare Projekte etwa zur Ermittlung von Cross- oder Upsellingpotenzialen planen, sich aber die Möglichkeit offen lassen wollen, im Erfolgsfall ihre Lizenz aufzustocken.

Bedienbarkeit und Analyseeffizienz

Im Anwendungstest zeigte sich, dass die Bedienung dank grafischer Benutzeroberflächen grundsätzlich

Spezialisierte Data-Mining-Werkzeuge sind bereits für unter 10 000 Euro zu haben.

leichter von der Hand geht. Dennoch erfordern insbesondere die Data-Mining-Suiten im Vergleich zu spezialisierten Werkzeugen nicht nur einen erhöhten Einarbeitungsaufwand, sondern auch fundiertes Hintergrundwissen. Dazu kommt, dass Anwender aus den Fachabteilungen meist andere Anforderungen an die Bedienung stellen als IT-Nutzer. *Rapidminer* und *KXEN* versuchen diesen Aspekt zu berücksichtigen, indem sie beispielsweise

Assistenten anbieten, die eine feste Abfolge an Analyseschritten vorgeben und die erforderlichen Eingaben dazu systematisch abfragen.

Für den Qualitätstest (siehe auszugsweise die Tabelle unten auf dieser Seite) wurde die oben genannte Testdatei mit 30 000 Datensätzen eingelesen und darauf mit allen Tools ein Klassifizierungsmodell erstellt. Danach erfolgte die Bewertung mit einer einheitlichen Kostenmatrix und die Umrechnung auf Beträge pro Datensatz, um einen normierten Vergleich zu ermöglichen. Um die

Achillesferse große Datenvolumina

Auch der Test mit großen Datenvolumina deckte erhebliche Unterschiede zwischen den Produkten auf: *KXEN* ließ sich etwa von der gestiegenen Datenmenge wenig beeindrucken. Bei anderen war bereits das Einlesen der Daten problematisch und erforderte zuweilen langwierige manuelle Eingriffe. In *SAP NetWeaver* muss zum Beispiel für jedes Attribut ein separates Infoobjekt angelegt werden, was sich bei 450 Stück als wenig praxistaug-

meist – je nach verwendetem Verfahren – deutlich überproportional. Insbesondere *Rapidminer* fiel diesbezüglich mit sehr langen Laufzeiten bis hin zu Abbrüchen wegen Hauptspeicherüberlaufs auf. Aber es geht auch anders: *KXEN* analysierte den kompletten Datensatz in weniger als zehn Minuten.

Stetig verbesserter Reifegrad von Data-Mining-Tools

“...contrarily to current thinking, our industry is still in its infancy” stellte Usama Fayyad, einer der Pioniere des Data Mining, in seiner Keynote auf der KDD’07 in San José fest. Die zu beobachtende zügige Weiterentwicklung der Data-Mining-Lösungen spricht in der Tat dafür, dass auch die Toolanbieter noch zahlreiche weitere – auch größere – Verbesserungsschritte vor sich haben. Dazu gehören beispielsweise die höhere Automatisierung immer wiederkehrender Routineaufgaben, der effiziente Umgang mit großen Datenmengen und vor allem die umfangreiche Unterstützung der Anwender beim fachlichen Einstieg in die Analyse.

Dennoch ist unübersehbar, dass sich der Reifegrad von Data-Mining-Lösungen in den letzten Jahren deutlich erhöht hat. Erfreulich ist insbesondere, dass für die Unternehmen der Einstieg in die explorative Datenanalyse dank flexibler Preismodelle, eines vielfältigen Produktangebotes und neuen Ansätzen wie Self-Acting Data Mining unkomplizierter und attraktiver ist als je zuvor. ◀

Für den Qualitätstest wurden mit allen Tools ein Klassifizierungsmodell erstellt.

Robustheit der generierten Modelle zu prüfen, sind diese zusätzlich auf einen weiteren Datensatz (15 000 Zeilen) angewendet worden (vorletzte Tabellenspalte). Als absolute Größe ist in der letzten Tabellenspalte der jeweilige Gesamtgewinn bei 50 000 Datensätzen aufgeführt.

Bei der Berechnung zeigen sich zunächst erhebliche Unterschiede in den Laufzeiten; diese verstärken sich bei größeren Datenmengen (siehe folgender Abschnitt). Dem *SAS Enterprise Miner* gelang insgesamt die beste Modellqualität. Die Ergebnisqualität bei *Rapidminer* fällt im Vergleich ab; insbesondere die Übertragbarkeit der Modelle ist gering. *KXEN* kann eine gute Modellqualität vorweisen und ist zudem bei der Berechnung unerreicht schnell. Güte und Laufzeit der Modelle in *SAP NetWeaver* waren guter Durchschnitt.

lich erweist. Bei einer vierstelligen Anzahl an Attributen – durchaus gängige Größenordnungen beispielsweise bei Telekommunikationsdienstleistern, großen Versicherungen und im Versandhandel – stoßen Anwender zusätzlich an technische Grenzen der zugrunde liegenden Datenbanken, die etwa im Falle von Microsoft und Oracle nicht mehr als 1024 Spalten pro Tabelle zulassen.

Für die anschließende Berechnung deskriptiver Statistiken oder zur grafischen Darstellung von Häufigkeitsverteilungen muss der Analytiker zuweilen Geduld aufbringen. Dies gilt auch für die eigentliche Data-Mining-Analyse: Da jedes weitere in die Berechnung eingehende Attribut dem Algorithmus eine zusätzliche Dimension hinzufügt, steigen die Laufzeiten

Modellqualität der getesteten Produkte (Auszug)

Produkt	Rechenzeit in sec	Verfahren	Gewinn pro Datensatz (Modellerstellung)	Gewinn pro Datensatz (Modellanwendung)	Gesamtgewinn bei 50.000 Datensätzen
SAS Enterprise Miner 5.3	93	Entscheidungsbaum	1,10 €	0,99 €	49.781,95
Rapidminer 4.2	327	Entscheidungsbaum	0,41 €	0,05 €	2.600,00
KXEN Analytic Framework 4.04	14	Robust Regression	0,98 €	0,81 €	40.261,65
SAP NetWeaver 7.0 Data Mining Workbench	66	Entscheidungsbaum	0,78 €	0,51 €	25.704,81

bei Zufallsauswahl (zum Vergleich): -9,23

Quelle: Peter Neckel/mayato