mayato

act on facts

Data Mining Software 2009

Funktionsvergleich und Benchmarkstudie

Zusammenfassung

mayato®-Studie „Data Mining Software 2009“

Analyseerfolge zu überschaubaren Kosten

Der Markt für Data-Mining-Software ist in den letzten Jahren spürbar in Bewegung geraten. Interessierten Firmen steht eine Vielzahl von Werkzeugen zur Auswahl. Welche Stärken und Schwächen zeigen die aktuellen Tools im Praxistest?
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 & Vertrieb“ stehen analytische Fragestellungen nach wie vor ganz oben auf der Prioritätenliste: Im aktuellen CRM-Barometer 2008 sehen 40% 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 BARC wächst der deutsche Markt für BI-Software insgesamt voraussichtlich um 10 bis 12 % jährlich bis zum Jahr 2012.

Steigende Anforderungen

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-Software 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: 12 Data-Mining-Produkte im Test

Der Data-Mining-Softwaremarkt ist durch eine große „Artenvielfalt“ an Lösungen gekennzeichnet, was zu einer gewissen Unübersichtlichkeit führt und die Auswahl erschwert. Die aktuelle mayato-Studie „Data Mining Software 2009“ erleichtert die Entscheidungsfindung: Darin wurden zwölf ausgewählte Data-Mining-Tools einem Funktionsvergleich unterzogen; vier Produkte sind zudem anhand mehrerer Testdatensätze ausführlich auf ihre Praxistauglichkeit geprüft.
Bewertet wurden u.a. 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.

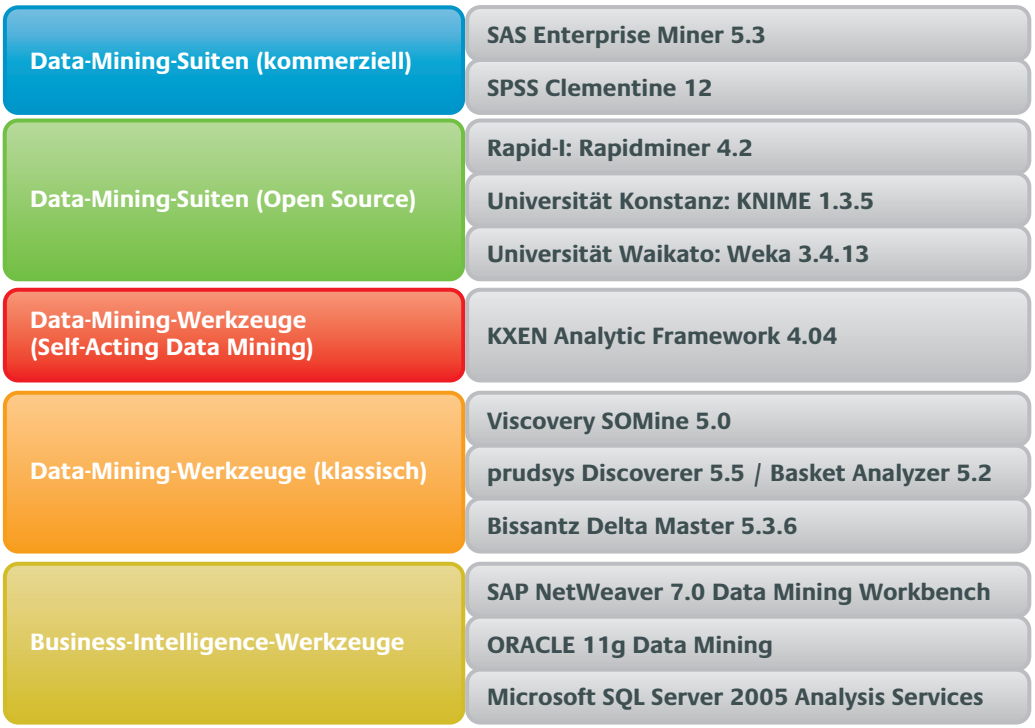
Grundtypen von Data-Mining-Produkten

Die Grundtypen mit der zugehörigen Auswahl der in der Studie besprochenen Tools zeigt Abb. 1: Die klassischen Data-Mining-Suiten (z.B. 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 Gruppe der schlankeren Data-Mining-Werkzeuge mit reduzierter Funktionalität, wie sie etwa Viscovery (SOMine), prudsys (Discoverer/Basket Analyzer) oder Bissantz (Delta Master) anbieten.
Diese Werkzeuge sind auf bestimmte Anwendungsgebiete (z.B. Controlling) oder Analysefälle (z.B. Prognose- und Klassifizierungsaufgaben) oder einer Kombination aus beidem spezialisiert. Sie setzen weitgehend bereinigte Daten voraus, da ihre Datenvorverarbeitungsmöglichkeiten meist stark eingeschränkt sind: So sind z.B. statistische Verfahren wie Hauptkomponenten- oder Faktorenanalysen vielfach nicht implementiert.

Eine Sonderstellung in dieser Kategorie nimmt die softwaretechnische Umsetzung des Self-Acting Data Mining ein – dieser hochautomatisierte Ansatz kommt weitgehend ohne manuelle Datenvorverarbeitung und Parametrisierung aus.
Eine weitere Kategorie von Software, in der zunehmend Funktionen zur explorativen Datenanalyse zur Verfügung stehen, sind BI-Umgebungen: Zahlreiche Datenbank- und BI-Anbieter wie SAP (NetWeaver BI), ORACLE (Data Mining) oder Microsoft (SQL Server Analysis Services) haben in manchen Fällen recht umfangreiche Data-Mining-Funktionen, z.T. gar selbst entwickelte Verfahren (wie etwa den ORACLE-Segmentierungsalgorithmus „Orthogonal Partitioning Clustering“) in ihre Produkte integriert.

Im BI-Bereich ist generell ein Zusammenwachsen von Datenverwaltungs- und Datenanalysesystemen festzustellen. Diese Kombination ist nahe liegend, da viele analyserelevante Daten bereits in konsolidierter und qualitätsgesicherter Form in Data-Warehouse-Systemen vorliegen, auf die implementierte Data-Mining-Verfahren direkt zugreifen können, ohne dass die Daten aus anderen Systemen extrahiert werden müssen.



[Abb.1]: Taxonomie aktueller Data-Mining-Lösungen

Auch für Mittelständler attraktiv: Flexible Preismodelle

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 jährliche Wartungskosten von 100.000 Euro 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 lediglich ggf. 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 z.B. 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. Die Preismodelle der Anbieter werden zudem immer flexibler, was besonders dem Mittelstand entgegenkommt: Einige Data-Mining-Werkzeuge lassen sich z.B. 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 interessant, 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 & Analyseeffizienz

Im Anwendungstest zeigte sich, dass die Bedienung dank grafischer Benutzeroberflächen grundsätzlich 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 dem Rechnung zu tragen, indem sie z.B. Assistenten anbieten, die eine feste Abfolge an Analyseschritten vorgeben und die erforderlichen Eingaben dazu systematisch abfragen.

Für den Qualitätstest (siehe auszugsweise Tabelle 1) wurde die oben genannte Testdatei (30.000 Datensätze) 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 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 liegt durch den hohen Automatisierungsgrad bei der Analyseeffizienz vorn. KXEN ist zudem bei der Modellberechnung unerreicht schnell. Güte und Laufzeit der Modelle in SAP NetWeaver BI waren guter Durchschnitt.

	Rechenzeit in sec	Verfahren	Gewinn pro Datensatz (Modellerstellung)	Übertragbarkeit des Modells	Gewinn pro Datensatz (Modellanwendung)	Gesamtgewinn bei 50.000 Datensätzen
SAS Enterprise Miner 5.3	93	Entscheidungsbaum	1,10 €	90%	0,99 €	49.781,95 €
Rapidminer 4.2	327	Entscheidungsbaum	0,13 €	40%	0,05 €	2.600,00 €
KXEN Analytic Framework 4.04	14	Robust Regression	0,98 €	82%	0,81 €	40.261,65 €
SAP NetWeaver 7.0 Data Mining Workbench	66	Entscheidungsbaum	0,78 €	66%	0,51 €	25.704,81 €
			bei Zufallsauswahl (zum Vergleich): -9,23 €			

[Tab.1]: Modellqualität der getesteten Produkte (Auszug)

Achillesferse große Datenvolumina

Auch der Test mit großen Datenvolumina deckte erhebliche Unterschiede zwischen den Produkten auf: KXEN ließ sich von der gestiegenen Datenmenge am wenigsten beeindrucken. Bei anderen Tools war bereits das Einlesen der Daten problematisch und erforderte zuweilen langwierige manuelle Eingriffe: In SAP NetWeaver BI muss z.B. für jedes Attribut ein separates Infoobjekt angelegt werden, was sich bei 450 Stück als wenig praxistauglich erweist.
Bei einer vierstelligen Anzahl an Attributen – durchaus gängige Größenordnungen z.B. bei Telekommunikationsdienstleistern, großen Versicherungen und im Versandhandel – stößt man zusätzlich an technische Grenzen der zugrunde liegenden Datenbanken, die z.B. im Falle von Microsoft und ORACLE nicht mehr als 1.024 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 meist – je nach verwendetem Verfahren – deutlich überproportional. Insbesondere Rapidminer fiel diesbezüglich mit sehr langen Laufzeiten bis zu Abbrüchen wegen Hauptspeicherüberlaufs auf.
Aber es geht auch anders: KXEN analysierte den kompletten Datensatz als einziges Tool im Test 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 z.B. 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.
Denkbar wären hier z.B. vorgefertigte Analysebausteine, die sich an typischen fachlichen Fragestellungen orientieren, die mittels Datenanalysen beantwortet werden sollen.

Die zunehmende Verbreitung des Data Mining und die immer intensivere Nutzung in der Unternehmenspraxis führt zu dem Wunsch, die erzeugten Analyseergebnisse direkt und ohne Zeitverlust in die operativen Prozesse zu integrieren und dort verwenden zu können.
Es stellt sich generell die Frage, in welcher Weise insbesondere große Data-Mining-Suiten diese Anforderungen in Zukunft erfüllen können.

Dennoch ist unübersehbar, dass sich der Reifegrad von Data-Mining-Lösungen in den letzten Jahren deutlich erhöht hat. Ansätze wie Self-Acting Data Mining, die sich bewusst von der Algorithmenvielfalt traditioneller Data-Mining-Suiten absetzen und stattdessen durch einen möglichst hohen Automatisierungsgrad die Analyseeffizienz insgesamt deutlich verbessern, sind wichtige Schritte dazu, den aktuellen Anforderungen an Data-Mining-Lösungen gerecht zu werden.
Erfreulich ist insbesondere, dass für die Unternehmen der Einstieg in die explorative Datenanalyse dank flexibler Preismodelle und eines vielfältigen Produktangebotes unkomplizierter und attraktiver ist als je zuvor.

⇢ Hinweis

Die vollständige Studie können Sie zum Preis von 199 € (99.- € für Studenten und Hochschulen (mit Nachweis) käuflich erwerben.

Im Internet unter **www.mayato.com** bzw. per Email an **sales@mayato.com** mayato.

Notizen

mayato® GmbH | Am Borsigturm 9 | 13507 Berlin | T +49 30 4174.8657 | info@mayato.com | **www.mayato.com**

Gestaltung: sign-grafikdesign.de