mayato Data Mining Studie 2010: Entscheidungshilfe bei der Data-Mining-Toolauswahl

Die Einsatzfelder von Data-Mining-Analysen haben sich in den letzten Jahren sprunghaft vervielfältigt. Die Toolhersteller reagieren auf diese Entwicklung mit einer Vielzahl unterschiedlicher Produkte, deren Funktionalitäten, Bedienkonzepte und Algorithmen sich erheblich unterscheiden. Welches Werkzeug sich für welchen Analysebedarf am besten eignet, klärt die aktuelle Data Mining Studie 2010 des BI-Analysten- und Beratungshauses mayato.

Immer mehr Unternehmen erkennen, dass sich durch den überlegten Einsatz von Data Mining vielfach bereits kurz- und mittelfristig Wettbewerbsvorteile erzielen lassen – eine willkommene Chance, wenn man sich die angespannte Wettbewerbssituation in zahlreichen Branchen vor Augen führt.

Der gewinnbringende Nutzen von Data Mining lässt sich daher am ehesten im Marketing & Vertrieb begründen: Denn die Unternehmen müssen größte Anstrengungen darauf verwenden, neue Kunden zu finden, bestehende Kundenbeziehungen zu festigen sowie abgewanderte, aber profitable Kunden zurückzugewinnen. Für dieses Management der Kundenbeziehungen ist eine kontinuierliche Versorgung mit Informationen elementare Voraussetzung, die sich durch die Ergebnisse von Data Mining Analysen erzeugen lassen.

Besonders die Bedeutung des Cross- und Upselling hat in letzter Zeit enorm zugenommen: Insbesondere im (Web)Versandhandel gehören Empfehlungen der Art „Kunden, die dieses Produkt gekauft haben, kauften auch:…“ zum Tagesgeschäft. Die Erfolgsquote dieser zusätzlichen Produktangebote kann durch Data-Mining-Analysen des Verbund-Kaufverhaltens (Assoziations- und Sequenzanalyse) stark verbessert werden, was nicht nur (kurzfristig) den Umsatz erhöht, sondern meist auch (langfristig) den Kundenwert steigert.

Diese Erkenntnisse sind aber auch in anderen Branchen von Bedeutung. Insbesondere Finanzdienstleister und Versicherungen profitieren von den gewonnenen Erkenntnissen, um z. B. Direktmarketingaktionen zielgenauer umzusetzen.

Die Data-Mining-Produktvielfalt

Bereits die Grundtypen an Analysewerkzeugen sind vielfältig und deren jeweilige Vertreter nicht für jedes Einsatzgebiet passend. Um die Auswahl zu erleichtern, zeigt Abb. 1 einen Überblick über die wichtigsten Kategorien und Tools: Die klassischen Data-Mining-Suiten (z. B. von SAS, SPSS oder StatSoft) mit ihrem umfassenden Angebot an Datenvorverarbeitungsfunktionen und Data-Mining-Verfahren werden seit einiger Zeit auch in der Open-Source-Variante angeboten und erfreuen sich wachsender Beliebtheit.

Data-Mining-Suiten (kommerziell)	SAS Enterprise Miner 6.1	Im Test
	SPSS PASW Modeler 13	
	StatSoft: STATISTICA Data Miner 9	Im Test
Data-Mining-Suiten (Open Source)	Rapid-I: Rapidminer 4.6	
	Universität Konstanz: KNIME 2.0.3	Im Test
	Universität Waikato: Weka 3.6	
Data-Mining-Werkzeuge (Self-Acting Data Mining)	KXEN Analytic Framework 5.1.1	Im Test
Data-Mining-Werkzeuge (klassisch)	Viscovery SOMine 5.1	
	prudsys Realtime Decisioning Engine (RDE)	
	Bissantz Delta Master 5.4.1	
Business-Intelligence-Werkzeuge	SAP NetWeaver 7.0 Data Mining Workbench	Im Test
	ORACLE 11g Data Mining	
	Microsoft SQL Server 2008 Analysis Services	

Abb. 1: Taxonomie aktueller Data-Mining-Lösungen

Daneben gibt es die schlankeren Data-Mining-Werkzeuge mit reduzierter Funktionalität: Sie sind in der Regel auf bestimmte Anwendungsgebiete (z. B. Controlling) oder Analysefälle (z. B. Prognose- und Klassifizierungsaufgaben) spezialisiert. 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 und eignet sich daher insbesondere für Einsteiger.

Weiterhin haben zahlreiche Datenbank- und BI-Anbieter wie SAP, ORACLE oder Microsoft in manchen Fällen recht umfangreiche Data-Mining-Funktionen integriert, die häufig nicht zusätzlich lizenziert werden müssen.

Mittelstandstaugliche Preismodelle

Da Verbundkaufanalysen andere Schwerpunkte als Prognosemodelle im Marketing oder Zeitreihenanalysen bei einer Versicherung erfordern, ist es sinnvoll, sich vor der Auswahlentscheidung die konkreten, unternehmensspezifischen Einsatzmöglichkeiten von Data Mining vor Augen zu führen. Es lohnt sich, die Produktentscheidung gut zu überdenken: Je nach Funktionsumfang und Nutzerzahl liegt eine Client/Server-Lizenz typischerweise im mittleren sechsstelligen Eurobereich, zum Teil auch deutlich darüber. Die jährlichen Wartungskosten können zusätzlich noch ebenfalls Kosten im sechsstelligen Eurobereich verursachen.

Es geht aber auch günstiger: Selbst mächtige Data-Mining-Suiten können z. B. im Falle des STATISTICA Data Miner für moderate 20.000 Euro für die lokale Einzelplatzlizenz erworben werden – bei vollem Funktionsumfang. Spezialisierte Data-Mining-Werkzeuge sind vielfach noch günstiger zu erwerben. Weiterhin besteht bei vielen Anbietern die Möglichkeit, nur einzelne, wirklich benötigte Komponenten separat zu lizenzieren, was den Preis deutlich reduziert.
Für Open-Source-Lösungen entfällt der Anschaffungspreis; hier sind maximal jährliche Supportgebühren im vierstelligen Eurobereich zu zahlen.

Typische Auswahlkriterien für Data-Mining-Software

Der typische Anwender stellt dafür mittlerweile hohe Ansprüche an moderne Data-Mining-Tools: Zum einen wird der Umgang mit großen und sehr großen Datenmengen immer wichtiger, zum Anderen soll die Bedienung auch für Fachabteilungsnutzer ohne lange Einarbeitungszeiten möglich sein. Weiterhin stehen eine hohe Stabilität, die Automatisierung von Standardaufgaben sowie die Qualität und Interpretierbarkeit der Ergebnisse ganz oben auf der Wunschliste.

Studienumfang
Der Schwerpunkt der aktuellen Data-Mining-Studie 2010 liegt in der Analyse von Cross- und Upselling-Potenzialen mittels Assoziations- und Sequenzanalysen: Dazu mussten fünf Data-Mining-Tools und -suiten ein umfangreiches Testszenario absolvieren:
- o SAS Enterprise Miner 6.1
- o StatSoft STATISTICA Data Miner 9
- o KNIME 2.0.3
- o KXEN Analytic Framework 5.1.1
- o SAP NetWeaver 7.0 Data Mining Workbench.

Anhand einer Fallstudie und einem großen Datensatz mit 1,8 Mio. Zeilen wurde der gesamte Data-Mining-Prozess durchlaufen – von der Datenvorverarbeitung über die Datenexploration bis hin zur (grafischen) Darstellung und Interpretation der Ergebnisse. 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. Die Ausführungsgeschwindigkeit wurde mit einer Vielzahl unterschiedlicher Parametereinstellungen gemessen und dokumentiert. Zusätzlich zum umfangreichen Praxistest wurde für jedes getestete Werkzeug eine detaillierte Funktionsübersicht rund um die Assoziations- und Sequenzanalyse erstellt.

Testergebnisse
Der Funktionsumfang und die Laufzeit der Verfahren boten im Praxistest die größten Überraschungen.

Zahlreiche, mitunter gravierende Einschränkungen in der Funktionalität sind aus den Produktbeschreibungen mancher Hersteller gar nicht oder nur sehr mühsam herauszulesen. SAP BW und KNIME stellen beispielsweise keine Sequenzanalyse zur Verfügung, sodass etwa der zeitliche Abstand zwischen Kauftransaktionen nicht ausgewertet werden kann. Auch die zum Teil gravierenden Unterschiede in der Laufzeit können in der Praxis ein entscheidender Faktor sein – gerade bei Assoziationsanalysen, die typischerweise mehrere Millionen Transaktionen in kurzer Zeit analysieren müssen.

Die Bedienung geht hingegen dank grafischer Benutzeroberflächen grundsätzlich bei allen Testkandidaten leicht von der Hand. Dennoch erfordern insbesondere die mächtigen Data-Mining-Suiten im Vergleich zu spezialisierten Werkzeugen nicht nur einen erhöhten Einarbeitungsaufwand, sondern auch fundiertes Hintergrundwissen. StatSoft und KXEN kommen dem Gelegenheitsanwender entgegen, indem sie z. B. Assistenten anbieten, die eine feste Abfolge an Analyseschritten vorgeben und die erforderlichen Eingaben dazu systematisch abfragen.

Die gerade bei der Assoziationsanalyse wichtige Funktion, die Fülle der Ergebnisse in aussagekräftiger Form grafisch darzustellen, ist in den letzten Jahren spürbar ausgebaut und stark verbessert worden. Hier ist ein deutlicher Vorsprung der kommerziellen Data-Mining-Suiten vor spezialisierten Tools und Open-Source-Suiten wahrzunehmen. Das SAP BW und KNIME bieten diesbezüglich z. B. nur eine rudimentäre Unterstützung – der Anwender ist bei großen Datenmengen mit der Interpretation nicht sortierbarer Standardlisten, die mehrere Hundert Assoziationsregeln enthalten, deutlich überfordert. Positiv bleiben bei SAP der souveräne Umgang mit großen Datenmengen und die insgesamt gute Systemstabilität.

Dass es auch komfortabler geht, zeigen der SAS Enterprise Miner und der STATISTICA Data Miner. Insbesondere die von beiden Tools gebotenen Optionen zur grafischen Aufbereitung und Exploration der Assoziationsregeln sind im Testfeld eine Klasse für sich. Hier kann KXEN nicht ganz mithalten, überzeugt aber insgesamt vor allem mit einem durchgängig einsteigerfreundlichem Bedienkonzept und der Geschwindigkeit des selbstentwickelten Assoziations-Algorithmus.

Stetig verbesserter Reifegrad von Data-Mining-Tools
„The fruits of knowledge growing on the tree of data are not easy to pick". Diese Einschätzung des Data-Mining-Experten William Frawley aus dem Jahr 1991 hat im Grundsatz auch heute noch seine Berechtigung. Dennoch hat sich der Reifegrad von Data-Mining-Lösungen deutlich erhöht. Neue Ansätze wie Self-Acting Data Mining ermöglichen die einfache Nutzung von Analyse-Ergebnissen und den Einsatz für nahezu jede (unternehmens-)spezifische Aufgabe. Gerade Cross-Selling-Analysen eignen sich aufgrund der geringen Datenanforderungen und der vielfältigen Anwendungsmöglichkeiten besonders gut als Einstieg in die explorative Datenanalyse.
Die vollständige Studie kann ab sofort in der Druckversion oder als PDF über www.mayato.com erworben werden.

Peter Neckel
Analyst und Leiter der Studie beim BI-Analysten- und Beratungshaus mayato
peter.neckel@mayato.com
www.mayato.com