



Einstieg ins Data Mining wird leichter

Data Mining hat sich inzwischen in vielen Branchen zu einem strategischen Wettbewerbsfaktor entwickelt. Das hat in den letzten Jahren zu einer stetigen Ausweitung der Anwendungsgebiete geführt: Von der 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.

Das BI-Beratungshaus mayato hat in seiner aktuellen Studie „Data Mining Software 2009“ zwölf Data-Mining-Suiten und -Werkzeuge einem Funktionsvergleich unterzogen. Vier der Lösungen durchliefen zudem einen detaillierten Praxistest. Das Ziel der Studie besteht darin, Unternehmen bei der Tool-Auswahl im stark diversifizierten Softwaremarkt für Data Mining (DM) zu unterstützen. Die Studie berücksichtigt zahlreiche verschiedene Typen von Analysesoftware – für die Frage, welche Art von Werkzeug das richtige ist, sollte jedes Unternehmen im Vorfeld klar seine Zielsetzung definieren.

Ergebnis der Untersuchung: Die qualitativ hochwertigsten Modelle ließen sich mit SAS erstellen – Open-Source-Suiten wie RapidMiner blieben hingegen hinter den Erwartungen zurück. KXEN Analytic Framework konnte sich als Werkzeug für automatisiertes Data Mining hinsichtlich Effizienz, Bedienbarkeit, Performance bei großen Datenmengen und Geschwindigkeit der Modellerstellung überraschend an erster Stelle positionieren. Durch flexiblere Preismodelle der Anbieter rückt zudem das bisher als kostenintensiv und risikobehaftet geltende Data Mining auch für Mittelständler in den Fokus.

Das Resümee von mayato lautet: Der Markt für Data-Mining-Software ist in den letzten Jahren spürbar in Bewegung geraten und der Einstieg in DM-Projekte heute deutlich leichter.

Untersuchte Anwendungen

Neben den kommerziellen Data-Mining-Suiten SAS Enterprise Miner und Clementine von SPSS untersuchte die Studie die drei Open-Source-Suiten RapidMiner, KNIME und Weka. Eine weitere Kategorie bestand in schlankeren Data-Mining-Werkzeugen mit reduzierter Funktionalität, die in der Regel auf bestimmte Anwendungsgebiete wie Controlling oder auf Analyseszenarien wie Prognose- und Klassifizierungsaufgaben spezialisiert sind. Dazu zählen Viscosity SOMine 5.0, prudsys Discoverer 5.5 und

Data-Mining-Suiten (kommerziell)	SAS Enterprise Miner 5.3
	SPSS Clementine 12
Data-Mining-Suiten (Open Source)	Rapid-I: Rapidminer 4.2
	Universität Konstanz: KNIME 1.3.5
	Universität Waikato: Weka 3.4.13
Data-Mining-Werkzeuge (Self-Acting Data Mining)	KXEN Analytic Framework 4.04
Data-Mining-Werkzeuge (klassisch)	Viscosity SOMine 5.0
	prudsys Discoverer 5.5 / Basket Analyzer 5.2
	Bissantz Delta Master 5.3.6
Business-Intelligence-Werkzeuge	SAP NetWeaver 7.0 Data Mining Workbench
	ORACLE 11g Data Mining
	Microsoft SQL Server 2005 Analysis Services

Taxonomie aktueller Data-Mining-Lösungen

Bissantz Delta Master 5.3.6. Eine Sonderstellung in dieser Kategorie nimmt das Self-Acting Data Mining am Beispiel von KXEN Analytic Framework ein. Der hochautomatisierte Ansatz kommt weitgehend ohne manuelle Datenvorverarbeitung und Parametrisierung aus. Weiterhin kamen die integrierten BI-Lösungen Oracle 11g Data Mining, SAP NetWeaver 7.0 Data Mining Workbench und Microsoft SQL-Server 2005 Analysis Services auf den Prüfstand.

Detaillierter Praxistest

Im Blickfeld der Untersuchung standen die häufigsten Auswahlkriterien von Anwendern. Das sind einer Studie der Aberdeen Group von 2008 zufolge: Hohe Stabilität, der unkomplizierte Umgang mit großen Datenmengen, die Automatisierung von Standardaufgaben sowie die Qualität und Interpretierbarkeit der Ergebnisse und eine einfache Bedienbarkeit ohne lange Einarbeitungszeiten. Besonderes Augenmerk galt auch der Gesamteffizienz des Analyseprozesses und der Anwendung der einzelnen Werkzeuge und Suiten sowie den daraus resultierenden Gesamtkosten. KXEN, RapidMiner, SAS und SAP NetWeaver wurden zusätzlich anhand von Testdatensätzen detailliert auf ihre Praxistauglichkeit geprüft.

Das Testen von Kriterien wie Geschwindigkeit, Automatisierungsgrad und Ergebnisqualität erfolgte mittels einer überschaubaren Testdatei mit 30.000 Datensätzen und 15 Variablen. Das Systemverhalten bei großen Datenmengen wurde anschließend durch Einlesen eines umfangreichen Datenvolumens mit 100.000 Datensätzen und 450

Variablen gemessen. Die Palette der Benchmarkdaten enthielt eine Reihe typischer Probleme für DM-Tools.

Ergebnisse im Praxistest

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 Ergebnisqualität bei RapidMiner fiel im Vergleich ab, insbesondere die Übertragbarkeit der Modelle erwies sich als gering. KXEN konnte eine gute Modellqualität vorweisen und zeigte sich zudem bei der Berechnung als unerreichbar schnell. Resultate und Laufzeit der Modelle in SAP NetWeaver waren guter Durchschnitt.

Beim Test mit großen Datenvolumina veränderte sich die Performance von KXEN Analytic Framework auch bei gestiegenen Datenmengen kaum. Bei anderen Produkten gestaltete sich bereits das Einlesen der Daten problematisch und erforderte teilweise langwierige manuelle Eingriffe. In SAP NetWeaver beispielsweise muss für jedes Attribut ein separates Info-Objekt angelegt werden – für 450 Variablen ein erheblicher Aufwand. Auch die anschließende Berechnung deskriptiver Statistiken und der Aufbau grafischer Darstellungen – beispielsweise für Häufigkeitsverteilungen – nahm bei einigen Produkten sehr viel Zeit in Anspruch. Die Laufzeiten der eigentlichen Data-Mining-Analyse schließlich stiegen je nach verwendetem Verfahren häufig überproportional an: Jedes weitere in die Berechnung eingehende Attribut fügt dem Algorithmus eine

Klassifizierungs-/Prognoseregeln

- › **Was?** Kundenverhalten verstehen und vorhersagen
- › **Wozu?** Kampagnenplanung im Direktmarketing, Zielgruppenselektion, Kundenwertberechnung (Ermittlung relevanter Einflussgrößen), Vermeidung von Kundenabwanderung (Ermittlung von Abwanderungswahrscheinlichkeiten), ...

Assoziationen (Verknüpfungen, Abhängigkeiten, Sequenzen)

- › **Was?** Ermittlung von Beziehungen und (zeitlichen) Abhängigkeiten zwischen Produkten
- › **Wozu?** Umfangreiche Analyse der Warenkörbe der Kunden, Cross-/Up-Selling, Hinweise für absatzfördernde Bundle-Angebote, Erkennung typischer Kaufrhythmen, Bestimmung des "Share of Wallet", ...

Segmente

- › **Was?** Gruppierung von Objekten (z.B. Kunden) aufgrund ähnlicher Marktmalsausprägungen
- › **Wozu?** Markt- und Kundensegmentierung, Analyse der Kundenentwicklung (Ermittlung von Segmentwechseln der Kunden), ...

Änderungen/Abweichungen

- › **Was?** Ermittlung von Datensätzen, deren Ausprägungen im Vergleich zu Referenzwerten stark abweichen
- › **Wozu?** Erkennung und Vorhersage von Trends oder Zyklen in Zeitreihen (z.B. Entwicklung von Preisen, Abverkaufsmengen, Deckungsbeiträgen,...), Identifikation von Ausreißern und Frühwarnindikatoren, z.B. in Produktions- oder Controllingdatenbeständen

Data-Mining-Mustertypen

zusätzliche Dimension hinzu. Insbesondere RapidMiner fiel diesbezüglich mit sehr langen Laufzeiten von über drei Stunden bis zu Abbrüchen wegen Hauptspeicherüberlaufs auf. Mit KXEN nahm die Analyse des kompletten Datensatzes hingegen weniger als zehn Minuten in Anspruch.

Data Mining für Mittelstand greifbarer

Der Vergleich zeigt: Je nach Funktionsumfang und Nutzerzahl kann eine Client/Server-Lizenz mehrere 100.000 Euro in der Anschaffung sowie jährliche Wartungskosten von 100.000 Euro verursachen. Spezialisierte Data-Mining-Werkzeuge sind jedoch auch bereits für unter 10.000 Euro zu haben. Für Open-Source-Lösungen fallen maximal jährliche Supportgebühren im vierstelligen Eurobereich an. Für Unternehmen, die bereits über Enterprise-Lizenzen für Datenbanken wie Oracle oder Microsoft SQL-Server verfügen, bietet es sich an, die darin enthaltenen Data-Mining-Funktionen zu nutzen. Die Preise liegen für beide Produkte bei rund 27.000 Euro für die Einprozessor-Lizenz, auch ein Upgrade von vorliegenden Standardlizenzen ist möglich. Ein kostengünstiger Einstieg in das Data Mining wird zudem durch zunehmend flexible Preismodelle erleichtert. Einige Data-Mining-Werkzeuge lassen sich auf Monatsbasis zu Preisen im unteren vierstelligen Eurobereich mieten. Hinzu kommt, dass nicht mehr zwingend das Komplettprodukt bestellt werden muss. Oft können auch anwendungsspezifisch zu-

sammengestellte Pakete an Funktionen oder einzelne Data-Mining-Methoden separat erworben werden. Diese Angebote eignen sich vor allem für Unternehmen, die zunächst eher überschaubare Projekte etwa zur Ermittlung von Cross- oder Up-Selling-Potenzialen planen, und nur im Erfolgsfall ihre Lizenz aufstocken wollen.

Bedienung wird einfacher

Im Anwendungstest zeigte sich, dass die Bedienung dank grafischer Benutzeroberflächen grundsätzlich leichter von der Hand geht als noch vor wenigen Jahren. 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 diesem Umstand Rechnung zu tragen, indem sie beispielsweise Assistenten anbieten, die eine feste Abfolge an Analyseschritten vorgeben und die erforderlichen Eingaben dazu systematisch abfragen. Als Trend zeichnet sich ab, dass Data Mining künftig seine Sonderposition als Spezialistendisziplin verlieren wird.

Die prädiktive Datenanalyse steht für viele Unternehmen immer stärker im Vordergrund. Telekommunikationsanbieter, Banken und Versicherungen kommen beispielsweise schon seit einiger Zeit nicht mehr ohne Data Mining aus.

Im Anbieterlager ist derzeit eine Orientierung hin zu einfacherer Bedienbarkeit und einer verstärkten Automatisierung von Routineaufgaben der Datenvorbereitung zu verzeichnen.

Die komplette Studie ist zu einem Preis von 3.799,- Euro zzgl. MwSt. unter <http://www.mayato.com> erhältlich. Im März dieses Jahres veröffentlicht RAAD Research gemeinsam mit mayato eine neue Studie zum Thema „BI Potenziale 2009“.

Peter Neckel

Analyst und Leiter der Studie „Data Mining Software 2009“ bei der mayato GmbH