mayato

act on facts

Data Mining Studie 2010

Praxistest & Benchmarking

Zusammenfassung

Data Mining Studie 2010 Praxistest & Benchmarking

Vielfältiger Data-Mining-Softwaremarkt

Die immer stärkere Verbreitung des Data Mining in zahlreiche Branchen und in immer mehr Anwendungsgebiete hinein führt zu einer stetigen Diversifizierung der Data-Mining-Softwareprodukte. Zusammen mit der zunehmend flexiblen Preisgestaltung birgt diese Entwicklung für potenzielle Anwender und Unternehmen zahlreiche Vorteile: Sie können sich etwa für jeden auch etwas spezielleren Analysewunsch und passend zur jeweils vorliegenden IT-Infrastruktur individuell das geeignete Werkzeug auswählen.

Der Hauptnachteil liegt allerdings in der zunehmenden Unübersichtlichkeit des Angebots, was die konkrete Auswahl erschwert. Die mayato Data Mining Studie hat sich zum Ziel gesetzt, diese Auswahl durch eine umfangreiche Evaluation aktueller Tools in einem typischen Praxisszenario zu erleichtern.

Denn die unterschiedlichen „Verpackungen“, in denen die Anbieter ihre Softwareprodukte offerieren, sind zahlreich: Neben den klassischen „All-in-One-Produkten“, den Data-Mining-Suiten, die es seit einiger Zeit auch in der Open-Source-Variante gibt, werden zahlreiche spezialisierte Data-Mining-Werkzeuge für bestimmte Analyseaufgaben angeboten. Auch Hersteller von BI-Software haben vielfach Data-Mining-Funktionen in ihre Produkte integriert.

Cross- und Upselling-Potenziale realisieren mit Data Mining

Bei den Anwendungsgebieten steht nach wie vor der Marketing- und Vertriebsbereich im Vordergrund – im Speziellen gilt das Cross- und Upselling aktuell als die wichtigste Data-Mining-Anwendung: Kaum noch ein (Web-)Versandhändler kommt ohne Empfehlungen der Art *„Kunden, die dieses Produkt gekauft haben, kauften auch:...“* oder *„Was kaufen Kunden, nachdem Sie diesen Artikel angesehen haben?“* aus.
Die Erfolgsquote dieser zusätzlichen Produktangebote kann durch Data-Mining-Analysen des Verbundkaufverhaltens (Assoziations- und Sequenzanalyse) stark erhöht werden. Die Erkenntnisse werden auch in anderen Branchen wie z.B. bei Finanzdienstleistern und Versicherungen verstärkt dazu genutzt, Direktmarketingaktionen zielgenauer planen und umsetzen zu können.

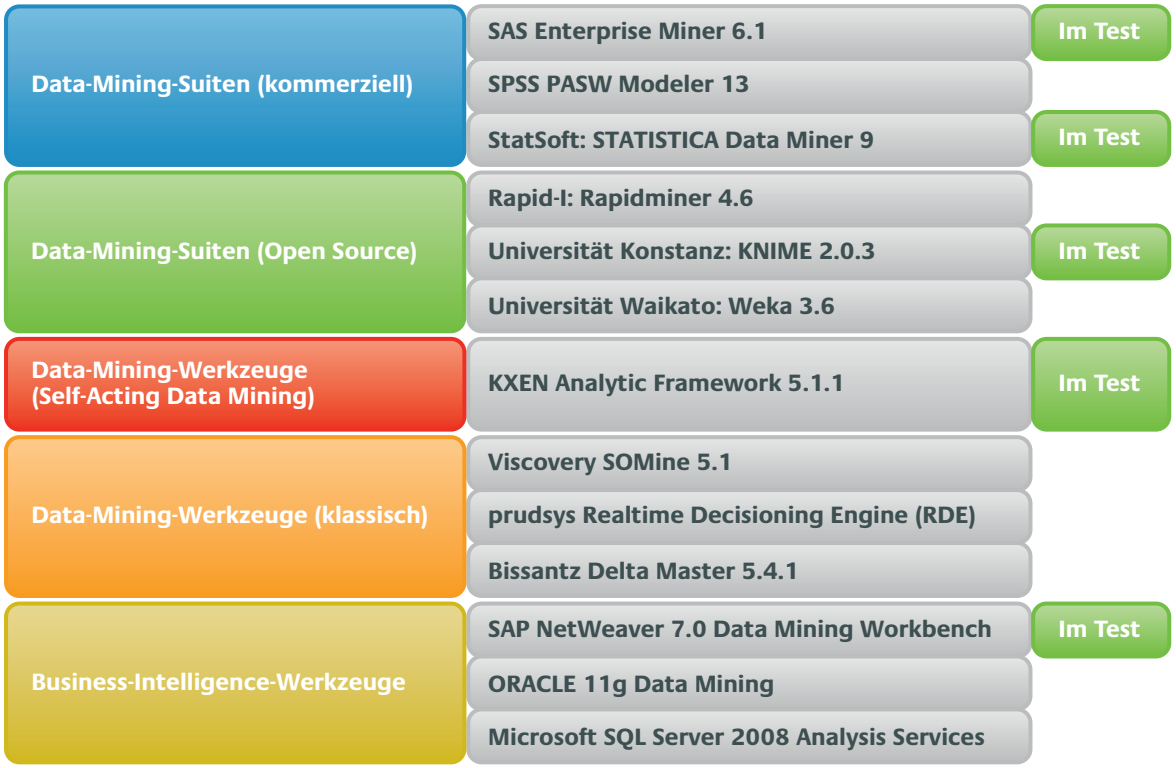
Grundtypen von Data-Mining-Software

Die Data Mining Studie 2010 vergleicht anhand eines großen Testdatensatzes (1,8 Mio. Zeilen) fünf Data-Mining-Tools und -Suiten:

› SAS Enterprise Miner 6.1
› StatSoft STATISTICA Data Miner 9
› KNIME 2.0.3
› KXEN Analytic Framework 5.1.1
› SAP NetWeaver 7.0 Data Mining Workbench.

Das Ziel bestand darin, die verschiedenen Analyseansätze und -konzepte anhand einer typischen, praxisnahen Fragestellung im Live-Einsatz zu prüfen. Daher wurden bewusst Data-Mining-Produkte aus gänzlich unterschiedlichen Tool- und Preiskategorien getestet (siehe Abb. 1): Die klassische, funktionsmächtige Data-Mining-Suite findet sich ebenso im Testfeld wie das spezialisierte, schlanke Data-Mining-Werkzeug. 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.

Weiterhin ist ein Business-Intelligence-Werkzeug getestet worden, das Data-Mining-Verfahren eher als Zusatz zu mächtigen Datenverwaltungsfunktionen offeriert. Um die Frage zu klären, ob sich die Analyseaufgabe auch mit einer kostenlosen Data-Mining-Suite zufriedenstellend und mit ähnlichem Bedienkomfort lösen lässt, wurde zusätzlich eine Open-Source-Variante in den Test aufgenommen.



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

Mittelstandstaugliche Preismodelle

Die Produktentscheidung will gut überlegt sein: Je nach Funktionsumfang und Nutzerzahl kann eine Client/Server-Lizenz mehrere Hunderttausend Euro in der Anschaffung sowie jährliche Wartungskosten ebenfalls im sechsstelligen Eurobereich verursachen.

Es geht aber auch günstiger: Spezialisierte Data-Mining-Werkzeuge sind bereits für unter 10.000 Euro zu haben. Aber auch funktionsmächtige 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. Für Open-Source-Lösungen entfällt der Anschaffungspreis; hier fallen lediglich jährliche Supportgebühren an, die sich im mittleren vierstelligen Eurobereich bewegen.

Studiendesign

Die diesjährige Ausgabe hat die Untersuchung von Cross-/Upselling-Potenzialen zum Schwerpunkt. Anhand einer Fallstudie wird der gesamte Data-Mining-Prozess durchlaufen – vom Einlesen der Daten über die Datenvorverarbeitung und die Datenexploration bis hin zur Durchführung der Assoziations- und Sequenzanalyse sowie der (grafischen) Darstellung und Interpretation der Ergebnisse.
Ein derart aufwändiges und umfangreiches Testverfahren liefert wertvolle praxisrelevante Fakten und Erkenntnisse im direkten Vergleich, die nicht aus den Produktbeschreibungen der Hersteller hervorgehen.

Die Bewertung stützt sich u.a. auf Bedienung & Dokumentation, Funktionsumfang, Systemverhalten bei großen Datenmengen und Stabilität. Weiterhin wurde die Ausführungsgeschwindigkeit 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.

Große Unterschiede im Praxistest bei Funktionsumfang, Laufzeit und Ergonomie

Im Praxistest zeigte sich, dass die Bedienung dank grafischer Benutzeroberflächen mittlerweile grundsätzlich leicht von der Hand geht. Ein dominierendes Bedienkonzept lässt sich jedoch weiterhin nicht erkennen; manche Hersteller gehen sogar dazu über, mehrere Modi für unterschiedliche Anwendertypen anzubieten.

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 versuchen dem Gelegenheitsanwender entgegen zu kommen, indem sie z.B. Assistenten anbieten, die eine feste Abfolge an Analyseschritten vorgeben und die erforderlichen Eingaben dazu systematisch abfragen.

Deutliche Unterschiede lassen sich auch beim Funktionsumfang feststellen. Bei SAP und KNIME zeigen sich hier z.B. praxisrelevante Lücken: Beide Tools bieten keine Sequenzanalyse an, so dass die zeitlichen Abstände zwischen den Kauftransaktionen nicht berücksichtigt werden können. Das bedeutet in der Praxis eine deutliche Einschränkung, da so nicht nur zahlreiche analytische Anwendungsszenarien im Umfeld des Cross- und Upselling, sondern auch im Risikomanagement oder in der Betrugserkennung bei Banken und Versicherungen ausgeklammert werden.

Beide Werkzeuge bieten zudem nur eine rudimentäre Unterstützung bei der Ergebnisauswertung – bei großen Datenmengen ist der Anwender mit der Interpretation nicht sortierbarer Standardlisten, die mehrere Hundert Assoziationsregeln enthalten, deutlich überfordert. KNIME fällt zusätzlich bei der Laufzeit ab – wesentlich verursacht durch die zeitraubende Datenvorverarbeitung, die für die Erstellung der zur internen Berechnung verwendeten Datenstruktur notwendig ist.

Dass es auch komfortabler geht, zeigen der SAS Enterprise Miner und der STATISTICA Data Miner. Sie punkten beide mit umfangreicher Funktionalität, einer hohen Ausführungsgeschwindigkeit und sicherem Umgang mit großen Datenmengen.
Insbesondere die von beiden Tools gebotenen, vielfältigen Optionen zur grafischen Aufbereitung und Exploration der Assoziationsregeln sind im Testfeld eine Klasse für sich.

Bei der Ergebnisauswertung verfügt das Analytic Framework von KXEN lediglich über eine – allerdings komfortabel konfigurierbare – textuelle Ausgabe.
KXEN überzeugt besonders mit seinem einsteigerfreundlichen Bedienkonzept und mit dem unerreicht schnellen, selbstentwickelten Assoziationsverfahren. Gerade das routinemäßige Analysieren großer Datenmengen geht mit diesem Tool leicht von der Hand – auch aufgrund der sehr guten Systemstabilität.

Assoziations- und Sequenzanalysen ideal zum Einstieg ins Data Mining

„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 ist unübersehbar, dass der Einstieg in die explorative Datenanalyse im Jahr 2010 bedeutend leichter fällt als noch vor wenigen Jahren:
Dank eines immer vielfältigeren Produktangebotes, verbesserter Bedienbarkeit und neuen Ansätzen wie Self-Acting Data Mining gibt es mittlerweile für nahezu jede (unternehmens-)spezifische Analyseaufgabe das passende Data-Mining-Werkzeug.

Auch am Ende des Datenanalyseprozesses – der Ergebnisinterpretation – ist eine positive Entwicklung zu verzeichnen: 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.

Dieser Fortschritt ist im Umfeld der Assoziations- und Sequenzanalyse besonders wichtig, denn derartige Analysen eignen sich aufgrund ihrer eingängigen Warenkorbmetapher, der geringen Datenanforderungen und der vielfältigen Anwendungsmöglichkeiten besonders gut als Einstieg in die explorative Datenanalyse.

Peter Neckel, Analyst und Leiter der Studie beim Analysten- und Beratungshaus mayato.

Kontakt:

mayato GmbH
Am Borsigturm 9
D-13507 Berlin
www.mayato.com

Georg Heeren
Tel. +49 171 481.8877
georg.heeren@mayato.com

⇢ Hinweis

Die vollständige Studie können Sie zum Preis von 299 € käuflich erwerben. (99.- € für Studenten und Hochschulen)

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