



In the past few years, there has been much activity in the market for data mining software. Companies interested in data mining can choose between a large number of tools. What strengths and weakness do the current tools display when really put to the test?

Data mining has now become a strategic competitive factor in many industries. This has led to a steady expansion of the application areas: be it the optimization of manufacturing in the automotive sector, risk management and fraud detection at banks and insurance companies, or – thanks to the increasing pervasiveness of RFID technology – supply chain analyses with logistics service providers and in the retail sector.

In the classic data mining domains of marketing and sales, analytical issues are still right at the top of the list of priorities. The current Capgemini CRM Barometer 2008 reveals that 40% of the companies questioned regard customer value modeling and segmentation as the most important CRM topics at the moment. This does not just apply to large companies – data mining applications are also becoming more popular in the midmarket. The tool providers are profiting from this, too: According to the Business Application Research Center (BARC), the German market for business intelligence software is forecast to grow by between 10% and 12% a year until 2012.

Growing requirements

However, today's data mining tools have to meet tough requirements. Access to additional data sources means that the quantities of data to be analyzed are multiplying in ever shorter intervals. The number of data records to be analyzed and the attributes that describe them are likewise on the increase. What's more, fueled by plummeting memory prices, data is gathered and evaluated in even greater levels of detail. And – to add to this – the constant emergence of new data formats makes processing more complex.

Independently of one another, the Aberdeen Group and Rexer Analytics investigated in May and September 2008 respectively the selection criteria companies use when they look for data mining software. The results show that stability, the simple handling of large quantities of data, the automation of standard tasks, and the quality and interpretability of the results are high on the agenda. Furthermore, not only the employees in the user departments demand usability without the need for extensive training.

Design of study: 12 data mining products scrutinized

The data mining software market comprises many different species of solution, which complicates matters and makes selecting the right one more difficult. The current mayato study "Data Mining Software 2009" makes it easier to arrive at a decision. In this study, 12 data mining tools were selected and their functions compared. In addition, four of the products were thoroughly examined with regard to their practicality using several test data records. Factors that were assessed include usability, stability, system response with large data quantities, documentation, and the overall efficiency of the analysis process, which encompasses criteria such as speed, level of automation, and quality of results. The latter was determined using a manageable test file (30,000 data records, 15 variables). The system response with large data quantities was testing by importing a substantial volume of data (100,000 data records, 450 variables).

Basic types of data mining products

Figure 1 shows the basic types of data mining products with the corresponding selection of tools discussed in the study. Classic data mining suites (for example, from SAS and SPSS) with their comprehensive selection of data preparation functions and data mining methods are now also available from open source providers such as Rapid-I (Rapidminer), the University of Konstanz, Germany (KNIME), and the University of Waikato, New Zealand (Weka) – and with similarly powerful features.

Then there is the group of slimmer data mining tools with reduced functionality, such as those from Viscovery (SOMine), prudsys (Discoverer/Basket Analyzer), and Bissantz (Delta Master). These tools are specially designed for certain application areas (for example, controlling) or analysis cases (for example, forecasting and classifying tasks) or a combination of the two. They require largely cleansed data, because their data preparation options are usually very restricted. For instance, statistical methods such as main component analysis or factor analyses are in many cases not implemented.

The implementation of self-acting data mining in a software context occupies a special position in this category. This highly automated approach manages without manual data preparation and parameterization to a great extent.

A further category of software in which functions for explorative data analysis are increasingly found is business intelligence (BI) environments. Numerous database and BI providers like SAP (SAP NetWeaver BI), Oracle (Data Mining) and Microsoft (SQL Server Analysis Services) have in some cases quite comprehensive data mining functions, and some even have their own self-developed methods (such as the Oracle partitioning algorithm called Orthogonal Partitioning Clustering) integrated into their products. In the BI area, a convergence of data management and data analysis systems can generally be observed. This combination makes sense, because much analysis-relevant data already exists in a consolidated form in data warehouse systems, so the implemented data mining methods can access the data directly without it having to be extracted from other systems.



Figure 1]: Classification of current data mining solutions

Flexible price models mean better value for the midmarket

The product decision is not one to be made lightly: Depending on the features and number of users, a client/server license may well cost several €100,000 to purchase, plus €100,000 in annual maintenance fees.

There are, however, much cheaper options: Specialized data mining tools can be bought for under €10,000; and for open source solutions, the only costs incurred are any annual support fees, which carry a four-digit euro price tag. Making use of database licenses that you already have is another way of implementing data mining projects on a low budget: Data mining functions are included, for example, in the Enterprise licenses for the databases from Oracle and Microsoft (SQL Server). Both products cost around €27,000 for a single processor license, and it is also possible to upgrade from existing standard licenses.

Furthermore, the providers' price models are becoming ever more flexible, which is good news for the midmarket. For instance, some data mining tools can be rented on a monthly basis for rates of a few thousand euros. Another factor is that the full product does not necessarily have to be ordered. Instead, it is often now possible to purchase application-specific packages of functions or even individual data mining methods separately. Such offerings are especially appealing to companies that initially plan smaller projects, like determining potential for crossselling or up-selling, but that want to keep open the option of topping up their license if they are successful.

Usability and efficiency

The application test revealed that, as a rule, usability has improved greatly thanks to graphical interfaces. Nevertheless, data mining suites in particular – compared with specialized tools – require not only more initial training but also sound background knowledge. Another factor is that people in the user departments usually have different usability requirements than their colleagues in the IT department. Rapidminer und KXEN attempt to cater to this by, for instance, providing assistants that predefine the order of the analysis steps and systematically retrieve the entries required in each step.

For the quality test (see excerpt from Table 1), the test file mentioned above (30,000 data records) was read and a classification model was created with it using all the tools. After this, the evaluation took place with a uniform cost matrix, followed by conversion to amounts per data record, to enable a standardized comparison. To test the robustness of the generated models, these were additionally applied to another data record (15,000 lines) (penultimate column in the table). The total gain in each case with 50,000 data records is listed as an absolute value in the final column.

In the calculation, there are first of all substantial differences in the runtimes, and they become all the more pronounced with large data quantities (see following section). SAS Enterprise Miner emerged as having the best model quality overall. The quality of results for Rapidminer is less impressive in comparison. In particular, model transferability is poor. KXEN's model quality is good, and its high level of automation puts it ahead in terms of analysis efficiency. What's more, KXEN is unbeatably fast in calculating models. Performance and runtime for the models were a fair average in SAP NetWeaver.

Bewertungskriterien	Beschreibung
Funktionsumfang	Umfasst Datenvorverarbeitungsfunktionen, Data-Mining- Verfahren, Funktionen zur Darstellung der Ergebnisse.
Bedienung	Umfasst die Ergonomie und das beim Anwender notwendige Fachwissen zur Erzielung adäquater Analyseergebnisse.
Systemverhalten bei großen Datenmengen	Testergebnisse „Test große Datenmenge“ (KDD 98-Datensatz).
Stabilität	Systemstabilität im Test.
Dokumentation	Übersichtlichkeit, Verständlichkeit und Vollständigkeit der Dokumentation einschließlich Onlinehilfe und eventl. vorhandener Tutorials.
Gesamteffizienz des Analyseprozesses	Gesamtbewertung der Effizienz des Analyseprozesses. Setzt sich zu gleichen Teilen aus dem Automatisierungsgrad, der Ausführungsgeschwindigkeit und der Modellqualität zusammen.
Automatisierungsgrad	Automatisierungsgrad, gemessen über den Analyseprozess: Datenvorverarbeitung, Datenanalyse, Darstellung der Ergebnisse.
Ausführungsgeschwindigkeit	Umfasst die Rechenzeit der Datenvorverarbeitungsfunktionen, der Data-Mining-Verfahren sowie die Start- und Reaktionszeit der Software im praktischen Einsatz.
Modellqualität	Bewertung der Messwerte des Modellqualitätstests (adult 1994-Datensatz) in Gewinn pro Datensatz des besten Modells.

[Tab. 1]: Model quality for the tested products (excerpt)

Large data volumes – the Achilles heel

The test with large data volumes also uncovered considerable differences between the products. KXEN was least fazed by the increased data quantity. With other tools, even the reading of the data proved problematic and in some cases necessitated time-consuming manual interventions. In SAP NetWeaver, for example, a separate InfoObject has to be created for each attribute, which is hardly practical with 450 attributes. With attributes running into four-digit figures – which is quite feasible, for example, in the case of telecommunications service providers, large insurance companies, and mail order companies – the underlying databases also have their technical limits: With Microsoft and ORACLE, for instance, no more than 1,024 columns per table are permitted.

Analysts sometimes have to be patient if they want to subsequently calculate descriptive statistics or graphically display frequency distributions. And the same applies to the data mining analysis itself: Because each further attribute included in the calculation adds another dimension, the runtimes usually increase disproportionately, depending on the method used. In this respect, Rapidminer stood out with very long runtimes, sometimes resulting in terminations due to main memory overflow. But it doesn't have to be that way: KXEN analyzed the entire data in less than 10 minutes.

Data mining tools maturing steadily

“Contrary to current thinking, our industry is still in its infancy,” said Usama Fayyad, one of the pioneers of data mining, in his keynote at the KDD 2007 in San José, California. And the speedy development of data mining solutions does indeed indicate that the tool providers are also contemplating many more – and major – improvements. These include, for example, greater automation of recurring routine tasks, the efficient handling of large data quantities, and – in particular – comprehensive support for users in need of



a business-based introduction to analysis. Here, for example, it would be conceivable to have preconfigured analysis modules based on typical business questions to be answered using data analyses.

The increasing prevalence of data mining and its more intensive use in the business world mean that companies wish to integrate their analysis results directly and immediately into transactional processes, and then use them there. This begs the general question of how large data mining suites, in particular, will be able to meet these requirements in the future.

Yet there is no mistaking the fact that data mining solutions have become much more mature in the past few years. Approaches like self-acting data mining, which deliberately steer clear of the plethora of algorithms found in traditional data mining suites and instead significantly improve analysis efficiency overall using as much automation as possible, are important steps toward meeting the current demands of data mining solutions. It is particularly pleasing that, thanks to flexible price models and a wide range of product offerings, companies are finding a way into explorative data analysis that is more straightforward and affordable than ever before.