

Open Source Data Mining mit RapidMiner



Ingo Mierswa

Einführung

Data Mining, also die Entdeckung verborgener Zusammenhänge mittels Methoden des statistischen und des maschinellen Lernens, wird gemeinhin als ein Feld für Spezialisten betrachtet. Diese erstellen mit häufig sündhaft teuren Softwarelösungen mehr oder weniger komplexe Analyseprozesse, um beispielsweise drohende Kündigungen oder die Verkaufszahlen eines Produkts zu prognostizieren. Der wirtschaftliche Nutzen liegt auf der Hand, und so galt lange Zeit, dass die Anwendung von Data Mining Tools auch mit hohen Kosten für Softwarelizenzen und den auf Grund der Komplexität der Materie oft notwendigen Support verbunden war. Dass Softwarelösungen für Data Mining jedoch nicht zwingend teuer oder schwer zu bedienen sein müssen, daran dürfte spätestens seit der Entwicklung der Open Source Software RapidMiner wohl niemand mehr ernsthaft zweifeln.

Begonnen wurde die Entwicklung von RapidMiner unter dem Namen „Yet Another Learning Environment“ (YALE) am Lehrstuhl für künstliche Intelligenz der Universität Dortmund unter der Leitung von Prof. Dr. Katharina Morik. Mit der Zeit wurde die Software immer ausgereifter, beinahe eine halbe Million Downloads wurden seit dem Entwicklungsstart im Jahre 2001 verzeichnet. Unter den vielen Tausend Anwendern waren auch viele Unternehmen, welche nach einem Partner mit entsprechender Data Mining Kompetenz für Dienstleistungen und Projekte suchten. Diesem Bedarf folgend, wurde von den RapidMiner-Entwicklern das Unternehmen Rapid-I gegründet, welches heute auch für die Weiterentwicklung und Wartung der Software verantwortlich ist. Im Zuge der Unternehmensgründung wurde die Software YALE ihrer neuen Bedeutung entsprechend in RapidMiner umbenannt. Damit befinden sich RapidMiner und das dahinter stehende Unternehmen Rapid-I auf einem guten Wege: Rapid-I erreichte den vierten Platz beim nationalen Start-Up Wettbewerb „start2grow“ und gewann bei Europas höchstdotiertem IT-Wettbewerb „Open Source Business Award“ den ersten Preis. RapidMiner selbst wurde auf dem bekannten Data Mining Portal „KDnuggets“ zwei Mal in Folge zur meistverwendeten Open Source Data Mining Lösung gewählt – und auch insgesamt machte RapidMiner mit einem knappen zweiten Platz unter den mehr als 30 auch proprietären Lösungen eine mehr als gute Figur.

Flexibilität und Funktionsvielfalt

Was genau macht RapidMiner aber zur weltweit führenden Open Source Data Mining Software? Gemäß einer unabhängigen Vergleichsstudie der TU Chemnitz, die beim internationalen Data Mining Cup 2007 (DMC-2007) vorgestellt wurde, schneidet RapidMiner unter den wichtigsten Open Source Data Mining Tools sowohl hinsichtlich der Technologie als auch der Anwendbarkeit am besten ab. Dies spiegelt auch den Fokus der Entwick-

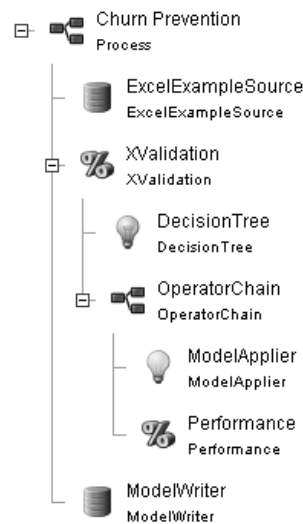


Abb. 1: Operator Tree für Churn Prevention

lungsarbeit wieder, der stets auf eine benutzerfreundliche Kombinierbarkeit der aktuellsten sowie der bewährten Data Mining Techniken abzielte.

Diese Kombinationsfreudigkeit verschafft RapidMiner eine hohe Flexibilität bei der Definition von Analyseprozessen. Prozesse können aus einer großen Zahl von nahezu beliebig schachtelbaren Operatoren erzeugt und schließlich durch sogenannte Operator Trees repräsentiert werden (siehe Abbildung 1). Der Prozessaufbau wird intern durch XML beschrieben und mittels einer graphischen Benutzeroberfläche entwickelt. Auch wenn das Konzept des Operator Trees zunächst verwundern mag, so gewöhnt man sich bereits nach kurzer Zeit an diese Arbeitsweise und wird den Geschwindigkeitsvorteil beim Prozessdesign und die klar strukturierte Arbeitsweise kaum noch missen wollen. Ein weiterer Vorteil liegt an den besser unterstützten Prüfungen, ob der Prozess auch syntaxkonform ist sowie den Möglichkeiten zur Definition von Breakpoints und Building Blocks. Damit kombinieren die Operator Trees von RapidMiner die Mächtigkeit von IDEs mit der Einfachheit von visueller Programmierung. Das modulare Vorgehen hat zudem den Vorteil, dass auch die internen Analyseabläufe genauestens geprüft und ausgenutzt werden können. Analysten können so beispielsweise auch in die einzelnen Folds einer Kreuzvalidierung hineinsehen oder den Effekt der Vorverarbeitung ebenfalls evaluieren – was mit anderen Lösungen typischerweise nicht möglich ist und oftmals in zu optimistischen Fehlerabschätzungen resultiert.

Insgesamt beinhaltet RapidMiner mehr als 500 Operatoren für alle Aufgaben der Wissensentdeckung in Datenbanken, d.h. Operatoren für Ein- und Ausgabe sowie der Datenverarbeitung (ETL), maschinelles Lernen und Data Mining. Aber auch Methoden des Text Mining, Web Mining, der automa-

tischen Stimmungsanalyse aus Internet-Diskussionsforen (Sentiment Analysis, Opinion Mining) sowie der Zeitreihenanalyse und Prognose stehen dem Analysten zur Verfügung. Zusätzlich beinhaltet RapidMiner mehr als 20 Verfahren, auch hochdimensionale Daten und Modelle zu visualisieren (siehe Abbildung 2). Darüber hinaus wurden auch alle Lernverfahren und Gewichtungsfaktoren der Weka Toolbox vollständig und nahtlos in RapidMiner integriert, so dass zu dem bereits enormen Funktionsumfang von RapidMiner auch noch einmal der vollständige Funktionsumfang des gerade in der Forschung ebenfalls weit verbreiteten Weka kommt.

Skalierbarkeit

Im November 2008 erschien die Version 4.3 von RapidMiner, im März 2009 die Version 4.4. Die Stoßrichtung wird in diesen beiden Versionen mehr als deutlich: zusätzlich zur großen Funktionsvielfalt liegt der Hauptfokus auf eine Optimierung hinsichtlich der Skalierbarkeit auch auf große Datenmengen. Schon immer war eine der Haupteigenschaften von RapidMiner ein Konzept ähnlich zu dem von relationalen Datenbanken, welches verschiedene Sichten auf Datenquellen ermöglicht. Dieses Konzept hat RapidMiner weiter verfeinert und bietet nun die Möglichkeit, eine Vielzahl solcher Sichten so zu kombinieren, dass die Daten on-the-fly transformiert und Datenkopien weitestgehend unnötig werden. Hierdurch erreicht RapidMiner einen im Vergleich oftmals deutlich niedrigeren Speicherverbrauch und kann bei entsprechender Konfiguration auch mit mehreren 100 Millionen Datensätzen spielend leicht umgehen.

Weitere Neuerungen wie die verbesserten Lift Charts von RapidMiner unterstützen die Optimierung von Direct-Mailing- und Marketing-Kampagnen, die Kündigungsprävention (Churn Reduction), die Erhöhung der Kundenbindung und die Kosten-Nutzen-optimierte Neukundengewinnung. Erweiterte Pivotisierungen, neue Aggregationsfunktionen, eine umfangreiche Datums- und Zeitbehandlung, die vereinfachte funktionsbasierte Konstruktion neuer Attribute, optimierte Wizards unter anderem für die automatische Optimierung von Data Mining Prozessparametern sowie neue Visualisierungen mit Zooming und Panning ermöglichen ebenfalls verbesserte Analysen und Datentransformationen und erleichtern die Bedienung zudem enorm.

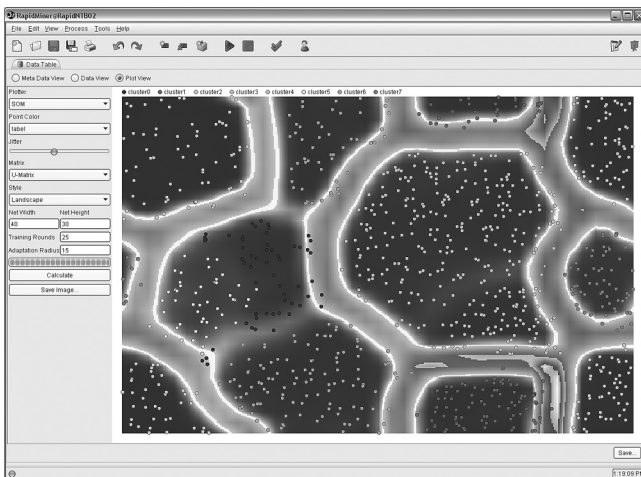


Abb 2: Visualisierung mittels SOM

Eine Frage des Formats

Ein weiterer Schwerpunkt von RapidMiner ist die hohe Konnektivität zu den verschiedensten Datenquellen wie z.B. Oracle, IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL und Ingres, dem Zugriff auf Excel-, Access- und SPSS-Dateien sowie zahlreichen anderen Datenformaten. Zusammen mit den hunderten Operatoren zur Datenvorverarbeitung lässt sich RapidMiner neben der Datenanalyse damit auch hervorragend zur Datenintegration und -transformation (ETL) einsetzen.

Und auch bei der Software selbst hat der Anwender die Wahl aus verschiedenen Formaten. RapidMiner gibt es einmal in der freien RapidMiner Community Edition, welche jederzeit und kostenlos von der Website heruntergeladen werden kann und in der Enterprise Edition, welche die Vorteile der freien Community Edition mit unternehmensspezifischen Lösungen für professionelle Anwender wie parallele Verarbeitung und Reporting kombiniert. Die RapidMiner Enterprise Edition schließt neben einem vollständigen professionellen Support mit garantierten Antwortzeiten auch alle notwendigen Garantien mit ein.

Damit hat es RapidMiner in kürzester Zeit geschafft, sich einen breiten Anwenderkreis in über 40 Ländern weltweit zu erarbeiten. Zu diesem gehören neben großen namhaften Unternehmen auch viele Unternehmen aus dem Mittelstand sowie zahlreiche Forschungseinrichtungen, bei denen die Analyse von Ergebnisdaten häufig zur Standardauswertung zählt.

Firmenprofil

Rapid-I ist Anbieter von Software, Lösungen und Dienstleistungen für Data Mining und Text Mining. Rapid-I, gegründet 2006 mit Hauptsitz in Dortmund, hat einen breiten Kundenkreis in über 40 Ländern weltweit zu dem u.a. Unternehmen wie E.ON, Sanofi Aventis, GfK, Libri, Lufthansa Systems, mobilkom austria, Schott, Pervasive, Allianz, comdirekt bank, ThyssenKrupp und Siemens gehören.

Kontakt

Dipl.-Inform. Ingo Mierswa
 Rapid-I GmbH
 Stockumer Str. 475
 44227 Dortmund
 Tel.: +49 (0)231-425-786-90
<http://rapid-i.com/>
contact@rapid-i.com



Ingo Mierswa ist Gründer und geschäftsführender Gesellschafter von Rapid-I. Er ist verantwortlich für die Bereiche Technologie & Entwicklung, Personalmanagement und Marketing und arbeitet darüber hinaus als Softwareentwickler und Berater für Rapid-I. Seit sieben Jahren koordiniert und leitet er die Entwicklung der Software RapidMiner mit weltweit 30 Entwicklern.