Data Mining in der Praxis

Praxis und Forschung: Wo liegen die Unterschiede?

Data Mining entwickelt sich an zwei Fronten: Zum einen in der privaten und universitären Forschung auf der Suche nach effizienten neuen Algorithmen, die beliebige Daten durchforsten, um interessante Informationen zu gewinnen. Zum anderen gibt es eine Entwicklung in der Praxis, in der Data Mining als Mittel eingesetzt werden soll, um Wettbewerbsvorteile zu erzielen.

Bereits die unterschiedliche Zielsetzung macht deutlich, daß sich beide Bereiche auch unterschiedlicher Methoden bedienen, um am Ziel anzukommen. Für die Praxis benötigt man aufgrund der pragmatischeren Zielsetzung eine erweiterte Definition des Begriffs Data Mining:

Data Mining ist der Prozeß, automatisch, vorher unbekannte, statistisch korrekte, interessante und interpretierbare Zusammenhänge in großen Datenmengen zu finden und diese für wichtige Unternehmensentscheidungen zu verwenden.

Besonders in Branchen wie Banken, Bausparkassen, Versicherungen, Versender, Handel und Telekommunikationsunternehmen, die große Datenmengen über ihre Kunden besitzen, können fortschrittliche Datenanalysemethoden für Marketingzwecke genutzt werden. Aufgabenstellungen sind dabei das Erkennen von homogenen Kundengruppen, die Auswahl von Kundengruppen für zielgerichtete Mailings, die Kreditwürdigkeitsanalyse, Kündigungsanalyse und Warenkorbanalyse (Assoziationsanalyse). Um diese Aufgaben zu lösen, kommen verschiedene Techniken des Data Mining zum Einsatz. Hauptvertreter bei den Techniken sind mit Sicherheit die Entscheidungsbäume: Über 80% der angebotenen Programme bieten diese an. Der Grund liegt vor allem in der Endanwenderfreundlichkeit - der verständlichen Ergebnisdarstellung, der Manipulationsmöglichkeiten (Marketing kann auch einmal das Alter als oberstes Trennkriterium einführen und sehen wie sich die Altersgruppen unterscheiden) und der Geschwindigkeit. Darüber hinaus kommen Clustering-Techniken (Finden von Kundengruppen), "black-box" Techniken wie neuronale Netze (vor allem bei Mailing-Auswahl und Kreditwürdigkeitsanalyse) sowie Assoziationsregelverfahren für die Warenkorbanalyse zum Einsatz.

Fallstricke in der Praxis
Beim praktischen Einsatz der Data Mining Techniken liegen die wahren Fallstricke nicht in einer mangelnden Güte der Verfahren, sondern in der Datengrundlage, den verfügbaren Tools sowie der Einbettung in die Geschäftsprozesse.

Datengrundlage. Hier erwachsen die Probleme in kleinen aus schlecht kodierten Datenfeldern (z.B. veraltete Sortimentsgliederungen, die nicht mehr die aktuelle Artikelstruktur widerspiegeln oder Felder, in die mehrere Informationen komprimiert gespeichert werden oder Felder, die erst zusammen mit anderen Informationen sinnhaft sind, im Extremfall durch über die Zeit unterschiedliche Semantik). Im Großen entstehen ungünstige Analysesituationen dadurch, daß die Daten für die operativen Abläufe in verschiedenen Geschäftsbereichen mit mehr oder weniger verschiedenen Datenmodellen gesammelt wurden. Es gilt, Daten zu einem Kunden zu bündeln, die in verschiedenen Bereichen gespeichert wurden (z.B. speichern Versicherungen die Daten getrennt nach den einzelnen Branchen). Nur aktuelle Daten sind sofort verfügbar, ältere Daten dagegen nur durch aufwendiges Prozedere zu erhalten. Kann man mit Data Mining auf einem speziell für die Analyse aufgebauten Data Warehouse aufsetzen, so ist schon viel gewonnen. Mit heutigen Data Mining Tools direkt auf diesen relationalen oder multidimensionalen Strukturen arbeiten zu können, ist jedoch eine Illusion: aus diesen Daten müssen trotzdem zunächst für den jeweiligen Zweck der Analyse relevante Informationen gebildet werden. Insbesondere ist die relationale Struktur dabei aufzulösen. Das Bilden von Sekundärmerkmalen kann wichtiger sein als der Einsatz eines besseren Verfahrens!

Tools. Das flexible Bilden solcher Merkmale (z.B. die Interessenlage aus granularen Einzelkäufen) übersteigt die Möglichkeiten heutiger Data Warehouse/Data Mining Tools und erfordert intensives Expertenwissen. Um solche komplexen Ableitungen aus Originaldaten effizient halten zu können, sind Überlegungen aus der Datenbankforschung von materialisierten Sichten notwendig, Ideen temporaler Datenbanken müssen Eingang in Datenbank-Produkte finden. Data Mining Tools setzen in der Regel auf flachen Tabellen auf, in denen das Untersuchungsziel (also etwa der Kunde) Schlüssel ist. Weder unterstützen sie besonders bei der Verwaltung von Analysen (wird z.B. ein Mailscoring durchgeführt, sind die Daten auf den Stand der

Aussendung an die Testgruppe zu speichern, die Analyse kann erst Wochen später - nach erfolgtem Rücklauf - durchgeführt werden), noch sind sie auf die Aufgabenstellungen optimiert. In der Regel handelt es sich lediglich um eine Ansammlung analytischer Verfahren. Das zweckbezogene Sammeln relevanter Daten, die Identifikation der Problemart und oft sogar das Ermitteln der Parameter eines Verfahrens werden den Endbenutzer aufgebürdet. Lediglich spezielle Database Marketing Systeme gehen die Aufgabenstellungen direkt an, ihnen fehlt aber noch die Fähigkeit, auf wirklich großen Datenmengen zu arbeiten.

Einbettung. Ein wesentlicher Punkt des Data Mining in der Praxis besteht auch in der Umsetzung der gewonnenen Erkenntnisse. Jedes Unternehmen hat verschiedene Kontaktpunkte zum Kunden; die einen sind vom Kunden initiiert (Call-Center, WWW, Filialbesuch), die anderen vom Unternehmen gesteuert (Brief-, Telefonaktionen, Vertreterbesuche). Besonders interessant aber auch schwierig ist die Nutzung der vom Kunden ausgehenden Kontakte. Dies bedeutet, daß Wissen an verschiedenste Stellen des Unternehmens transportiert werden muß und dort eventuell online ein Verfahren der künstlichen Intelligenz zum Einsatz kommen muß, um eine aktuelle Einschätzung vorzunehmen. Wird an dieser Stelle keine Automatisierung angestrebt, muß jedesmal von neuem um die Einbettung neuer Erkenntnisse in die Systeme der DV gekämpft werden. Was das heutige Data Mining auch übersieht, ist der zeitliche Aspekt von Analysen. Verfahren setzen zu einem Zeitpunkt auf Daten auf, Wissen wird gewonnen. Wann soll die nächste Analyse gemacht werden? Wird bereits bekanntes Wissen genutzt und nur noch auf die Veränderungen aufmerksam gemacht? Welche Veränderungen in der kontaktierten Kundengruppe gibt es? Wann ist eine aus dem Data Mining entstandene Regel zu korrigieren oder zu streichen?

Folgerungen für die zukünftige Entwicklung des Data Mining
- Aus den Ausführungen lassen sich Aussagen darüber treffen, wie sich Data Mining weiterentwickeln muß, um auch in der Praxis verbreiteten Einsatz zu finden:

- In der Analyse steht Data Mining nicht isoliert. Es ist umgeben von Mechanismen, die automatisiert die optimale Datengrundlage herstellen und die Informationen aus allen Geschäftsbereichen zu einem Kunden bündeln.

- Auf die Analyse folgt die Maßnahmendefinition. Idealerweise ist die Wahl des richtigen Ansprachemediums ebenfalls gesteuert durch Ergebnisse von früheren Data Mining Analysen.

- Die Maßnahmen sind zu verwalten, deren Wirkung auf die Kunden ist in zukünftigen Data Mining Schritten zu überprüfen.

- Data Mining darf nicht jedesmal von Null anfangen. Gewonnenes Wissen muß in zukünftigen Data Mining Schritten berücksichtigbar sein.

- Auf diese Weise ist ein echter Kreislauf anzustreben, indem als wertvoll erkannte Regeln sich mit der Zeit selbst anpassen und bei den verschiedensten Kundenkontakten genutzt werden.

- Zur Akzeptanz neuer Verfahren sind Darstellungen des gefundenen Wissens und das Einbringen von Domänenwissen (Eingriff seitens des Experten) notwendig.

- Verfahren müssen verstärkt zeitliche Entwicklungen aufspüren.

Der erste Punkt fordert nicht nur eine Fortentwicklung heutiger Data Warehouse Ansätze, sondern auch Automatismen zur Generierung von informativen Sekundärmerkmalen aus granulieren Daten je nach Untersuchungsziel. Ansätze wären einfache Attributwichtigkeitsanalysen mittels statistischer Tests oder durch Bayes Netze gepaart mit einem Mechanismus zur Durchforstung des Universums möglicher Attribute.

Die Maßnahmendefinition wird noch lange Domäne des menschlichen Experten sein, trotzdem lassen sich für abgegrenzte Problemstellungen aufgrund von Wissen aus früheren Aktionen, Maßnahmen und Wege der Kundenansprache optimieren.

Das auffälligste Manko bestehender Data Mining Produkte ist die verfehlte Annahme, daß außer den aktuellen Daten kein weiteres Wissen zur Verfügung steht. Diesem Manko ist leicht beizukommen, wenn man sich Methoden des Wissensvergleichs mit früheren Analyseläufen überlegt. Dies erfordert eine Speicherung des gewonnenen Wissens. Idealerweise kann auch Wissen aus verschiedenen Verfahren gemeinsam betrachtet werden.

Neue Analysemethoden müssen in der Lage sein, die Spreu der Erkenntnisse vom Weizen zu trennen. Solange dies nicht vollkommen automatisiert möglich ist, muß es eine einleuchtende Wissensrepräsentation geben, aus der der Experte schnell erkennen kann, ob es sich um einen Goldklumpen oder um Katzengold handelt. Weiterhin muß es dem Experten möglich sein, seine Vorstellungen in den Analyseprozeß einfliegen zu lassen, um in der Interaktion mit der Datenanalyse valide Zusammenhänge zu erarbeiten.

Es ist notwendig, ein Analysesystem zu schaffen, das eng mit den operativen Abläufen verzahnt ist. Es bedient sich sämtlicher relevanter Daten, das Wissen wird in möglichst allen Kontakten zum Kunden zum Einsatz gebracht, die ergriffenen Maßnahmen werden bewertet. Das gewonnene Wissen wird über das veränderte Kundenverhalten stetig angepaßt. Dieser Kreislauf stellt sicher, daß stets mit dem aktuellsten Wissen an den Kunden herangetreten wird. Dabei wird das üblicherweise als Data Mining Prozeß bezeichnete Modell auf das Gesamtgeschäft erweitert und somit in einen größeren Kreislauf eingebettet. Tools decken die Analyseschritte und Maßnahmenkontrolle ab, in Projektarbeit geschieht die Anbindung an die operativen Systeme.

Literatur

Faydad, Usama M.; Piatetsky-Shapiro, Gregory; Padhraic Smyth. The KDD Process for Extracting Useful Knowledge from Volumes of Data, in Communications of the ACM, Vol.39, No. 11,5. 27-34, Oktober 1996.

Lee, Hing-Yan; Ong, Hwee-Leng. Visualization Support for Data Mining, in IEEE Expert, S. 69-75, Oktober 1996.

Myrach, Thomas. TSQL2: Der Konsens über eine temporale Datenbanksprache, in Informatik-Spektrum 20, 5. 1 43-1 50, Springer-Verlag, 1997.

Napirokowski, Gregory; Borghard William. Modeling of Customer Response to Marketing of Local Telephone Services, in Dynamic Competitive Analysis in Marketing, Lecture Notes in Economics and Mathematical Systems 444, Steffen Jorgensen, Georges Zaccour (Hrsg.), 5. 252-283, Springer verlag, 1996.

Autor: Joachim Feist

Zurück