ARTICLE

BEZIEHUNG ZWISCHEN DATA MINING UND STATISTIK

Dr. Diego Kuonen, Statoo Consulting, Schweiz
kuonen@statoo.com

Im Data Mining geht es – wie auch in der Statistik – darum, «aus Daten zu lernen» oder «Daten in Wissen zu verwandeln». Der vorliegende Artikel behandelt die Beziehung zwischen Data Mining und Statistik.

Was ist Statistik und warum ist sie notwendig?
Statistik ist die Wissenschaft vom Lernen aus Daten. Sie umfasst den kompletten Vorgang vom Planen der Datensammlung und dem Datenmanagement bis hin zu abschliessenden Aktivitäten, bei denen aus numerischen Fakten Schlüsse gezogen und Ergebnisse präsentiert werden. Statistik berührt eines der grundlegendsten menschlichen Bedürfnisse: mehr herauszufinden über unsere Welt und wie sie angesichts von Veränderung und Unsicherheit funktioniert. Die wachsenden Einsatzmöglichkeiten für Statistik machen es unumgänglich, statistisches Denken zu verstehen und anzuwenden. Oder, um es mit den Worten von Herbert G. Wells zu sagen: «Statistisches Denken wird für den mündigen Bürger eines Tages dieselbe Bedeutung haben wie die Fähigkeit, lesen und schreiben zu können».

Aber warum ist Statistik notwendig? Wissen ist das, was wir kennen. Daten an sich stellen noch kein Wissen dar. Der Weg von Daten zu Wissen verläuft zunächst von den Daten zur Information: Daten werden zu Information, indem sie relevant für Entscheidungsfragen werden (Information unterscheidet sich von Daten klar in ihrem Kontextbezug und in der Anwendung von Qualität); und weiter von der Information zum Wissen: Information wird zu Wissen, indem sie durch die Daten gestützt wird und dazu dient, einen Entscheidungsprozess erfolgreich abzuschliessen. Statistik ist daher unerlässlich. Sie resultiert aus der Anforderung, dass Wissen auf systematischen Aussagen basieren muss.

Was ist Data Mining?
Data Mining liegt an der an der Schnittstelle von Statistik, Computerwissenschaft, Künstlicher Intelligenz, Mustererkennung, Maschinellem Lernen, Datenbankmanagement und Datenvisualisierung (um einige Bereiche zu nennen). Die Definition von Data Mining variiert daher je nach Blickwinkel:

«Data Mining ist der Prozess der Exploration und Analyse großer Datenmengen, mit dem Ziel, bedeutsame Muster und Regeln zu entdecken.» (Michael J. A. Berry und Gordon S. Linoff)

«Data Mining ist das Erkennen interessanter Strukturen (Muster, statistische Modelle, Zusammenhänge) in Datenbanken.» (Usama M. Fayyad, Surajit Chaudhuri und Paul S. Bradley)

«Data Mining ist ein analytischer Prozess der Datenexploration, bei dem nach konsistenten Mustern und/oder systematischen Beziehungen zwischen Variablen gesucht wird, um dann die Erkenntnisse zu validieren, indem die entdeckten Muster auf neue Daten angewendet werden.» (StatSofts «Electronic Statistics Textbook»)

Data Mining soll hier verstanden werden als ein nicht-trivialer Prozess, mit dem valide, neuartige, potenziell nützliche und zudem verständliche Muster oder Modelle in Daten identifiziert werden, um wichtige Businessentscheidungen zu treffen. «Nicht-trivial» bedeutet, dass es sich um die Suche nach Erkenntnissen handelt und nicht um einfache Berechnungen wie die eines Mittelwerts. «Valide» heisst, dass die Muster allgemein Bestand haben, also mit einer gewissen Sicherheit auch für neue Daten gelten müssen. Mit «neuartig» ist gemeint, dass die Muster nicht bereits

bekannt sein dürfen. «Potenziell nützlich» heisst, dass sie für den Anwender zu einem Mehrwert führen müssen. «Verständlich» bedeutet, dass die Muster nachvollziehbar und interpretierbar sein sollen. Damit beschränkt sich Data Mining – wie auch die Statistik – weder auf Modellbildung und Prognose, noch handelt es sich um ein fertiges Produkt, das erworben werden könnte. Data Mining ist vielmehr ein iterativer Prozess oder Zyklus der Problemlösung, der nur durch Teamarbeit bewältigt werden kann. Um es mit den Worten von Henry Ford zu sagen: «Zusammenkommen ist ein Beginn, Zusammenbleiben ist ein Fortschritt, Zusammenarbeiten ist Erfolg».

Beim erfolgreichen Data Mining ist das Schwierigste die Definition der Businessaufgabe, da es sich hierbei ausschließlich um eine Frage der Kommunikation handelt. Ein Analytiker muss jedoch verstehen, worum es geht. Auch die hochentwickelsten Algorithmen können nicht herausfinden, welche Anforderungen im Business wirklich bestehen. Vergessen wir nicht, dass «garbage in» zu «garbage everywhere» und «garbage out» führt. Ein anderes Schlüsselelement im Data Mining ist die Datenvorverarbeitung. Qualitativ gute Entscheidungen und Ergebnisse basieren auf qualitativ hochwertigen Daten. In der Realität gibt es jedoch keine perfekten Daten. Sie müssen beispielsweise aus mehreren Quellen zusammengeführt werden; weisen Missing Data, d.h. fehlende Werte, auf; sind «unsauber», enthalten also fehlerhafte und inkonsistente Werte; weisen Ausreisser auf; oder liegen nicht in dem benötigten Aggregationsniveau vor.

Hauptbestandteil des Data Minings sind die Datenanalyse und der Einsatz von Data-Mining-Techniken zur Suche nach Mustern in Datenbeständen. Hierbei ist es die Aufgabe des Computers, über verborgene Regeln und Eigenschaften der Daten Muster zu identifizieren. Welche Kombination an Data-Mining-Techniken zum Einsatz kommen kann, hängt von der Aufgabenstellung und den verfügbaren Daten ab. Die Grundidee ist es, dass es möglich ist, auch unerwartet auf Gold zu stossen; denn Data-Mining-Techniken extrahieren bisher unbekannte Muster. Da sie nicht offensichtlich sind, hat sie niemand zuvor bemerkt. Ausgehend von einem Datenbestand setzt man im Analyseprozess Methoden ein, die dazu dienen, die Datenstruktur optimal abzubilden und dabei Wissen aufzubauen. Einmal erlangtes Wissen lässt sich auf grössere Datenbestände übertragen. Dabei wird davon ausgegangen, dass diese Daten eine ähnliche Struktur besitzen wie die Stichprobendaten. Die Vorgehensweise ist analog zum Bergbau, wo groe Materialmengen von geringem Wert gesiebt werden, um etwas Kostbares zu finden.

Das klingt vertraut. Erinnern wir uns, dass wir Statistik als die Wissenschaft des Lernens aus Daten definiert haben. Die wesentliche Abfolge von den Daten zum Wissen war dabei: von Daten zu Information, und von Information zu Wissen. Diese Sequenz sei kurz veranschaulicht: Daten sind das, was gesammelt und gespeichert werden kann (z.B. Kundendaten, Lagerdaten, demographische oder geographische Daten). Sie werden zu Information, sobald sie für Entscheidungsfragen relevant sind. Information setzt einzelne Elemente der Daten zueinander in Beziehung (z.B. X lebt in Z; S ist Y Jahre alt; X und S sind umgezogen; W besitzt Geld in Z). Information wird zu Wissen, wenn sie für den erfolgreichen Abschluss des Entscheidungsprozesses eingesetzt wird. Wissen ordnet nun Elemente der Information einander zu (z.B. die Menge Q des Produkts A wird in Region Z konsumiert; Kundengruppe L nutzt N% von C während der Periode D). Tatsächlich handelt es sich um einen Ausschnitt der Business-Intelligence-Chain: von den Daten zur Information, von der Information zum Wissen, vom Wissen zur Entscheidung, von der Entscheidung zum Handeln (z.B. eine Promotionaktion für Produkt A in der Region Z; ein Werbemailing an Haushalte mit dem Profil P; ein Cross-Selling der Dienstleistung B zu Kundengruppe E). Die Hauptaufgabe liegt darin, von den Daten zum Wissen zu gelangen. Mit den Worten John Naisbitts: «Wir versinken in einer Flut an Informationen und dürsten zugleich nach Wissen.» Die Lösung für dieses Problem liegt im statistischen Data Mining, das den Wissenserwerb systematisch und wissenschaftlich stützt. Mit Data-Mining-Techniken können Unternehmen das bisherige Verhalten ihrer Kunden analysieren und so strategische Entscheidungen für die Zukunft treffen. Die Anwendungs-

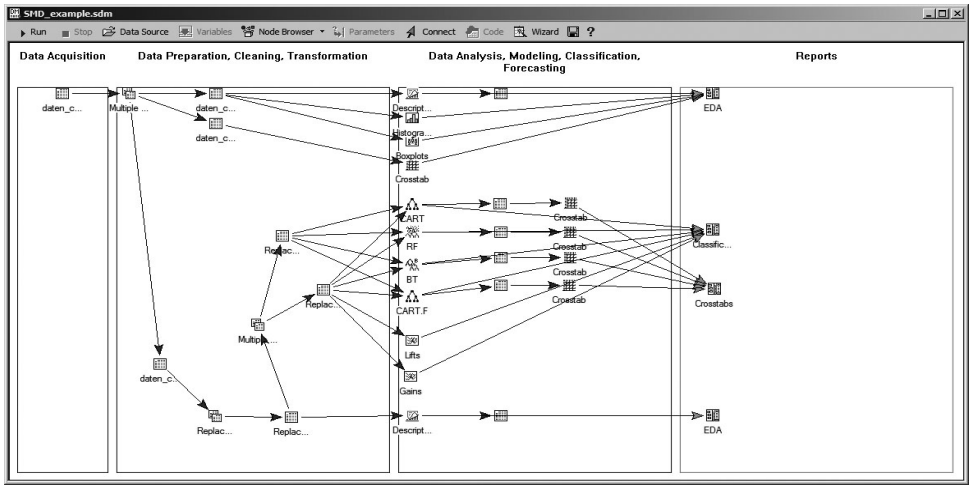


Abbildung 1.
Screenshot einer kommerziellen Data-Mining-Software: der «STATISTICA Data Miner» von StatSoft, der eine der umfassendsten Sammlungen an Data-Mining-Algorithmen enthält, die auf dem Markt verfügbar ist.

möglichkeiten für Data-Mining-Techniken und -Tools erstrecken sich dabei über so unterschiedliche Felder wie Strafverfolgung, Radioastronomie oder Prozesslenkung in Medizin und Industrie (um nur einige Bereiche zu nennen).

Warum gerade jetzt?

Die meisten Data-Mining-Techniken existieren, zumindest in akademischen Kreisen, seit vielen Jahren. Kommerzielles Data Mining hat sich jedoch erst in den letzten Jahren in grösserem Massstab durchgesetzt. Der Grund liegt im Aufeinandertreffen einer Vielzahl von Faktoren in den 90er Jahren: Daten wurden in bis dahin nicht gekanntem Masse erzeugt, gesammelt und (um den Zugriff zu verbessern) in Data Warehouses abgelegt. In der Hard- und Softwaretechnologie wurden Fortschritte erzielt (z.B. durch kontinuierlich sinkende Preise für Datenträger, bei Speicherkapazität und Rechengeschwindigkeit, und bei der Verfügbarkeit kommerzieller Data-Mining-Software; s. exemplarisch Abbildung 1). Der Wettbewerbsdruck dabei ist enorm: Unternehmen sehen sich hohen Businessanforderungen ausgesetzt, die den Einsatz von Data Mining erfordern (z.B. um das Marketing zu verbessern, Geschäftsbetrug aufzudecken, Herstellungsprozesse zu optimieren oder die Qualität der Kundenbeziehungen zu halten). Der wichtigste Aspekt ist jedoch, dass Data Mining Werkzeuge, mit denen man aus Daten lernen kann, in einen systematischen und iterativen Prozess einbindet.

Wieso Data Mining?

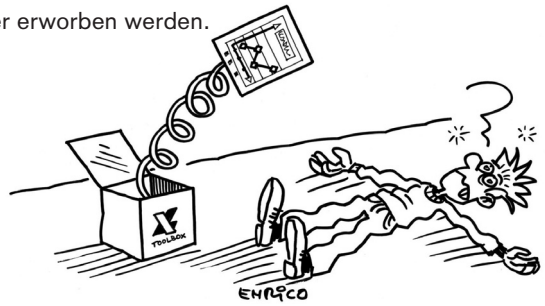
Ein kundenorientiertes Unternehmen sieht in jedem aufgezeichneten Kundenkontakt die Chance, etwas zu lernen. Lernen erfordert jedoch mehr als blosses Sammeln von Daten. Viele Unternehmen horten hunderte von Giga- oder Terabyte an Kundendaten, ohne wirklich etwas zu erfahren. Die Daten werden lediglich für Verwaltungsaufgaben genutzt, etwa für die Inventur oder Fakturierung. Sobald die Daten diese Aufgabe erfüllt haben, lagern sie nur noch auf Band oder werden archiviert. Der eigentliche Wert der Daten bleibt weitgehend unerkannt. Lernen erfordert, dass Daten zunächst aus vielen Quellen zusammengeführt und in einer konsistenten und verwertbaren Weise organisiert werden, damit Informationen zu Analysezwecken gewonnen werden können. Man spricht hier von Data Warehousing. Data Warehousing versetzt Unternehmen in die Lage, sich ihrer Kunden zu erinnern. Das Data Warehouse ist das Gedächtnis eines Unternehmens. Ohne intelligenten Gebrauch ist ein Gedächtnis jedoch nur von geringem Nutzen. Genau hier kommt Data Mining ins Spiel. Intelligenz erlaubt, das Gedächtnis zu durchkämmen und dabei Muster zu erkennen, Regeln zu aufzuspüren, Ideen zu entwickeln und Aussagen über die Zukunft zu treffen. Daten müssen analysiert, verstanden und in verlässliches Wissen überführt werden. Data Mining bietet Werkzeuge und Techniken, die dem Data Warehouse Intelligenz verleihen. Data Mining versieht ein Unternehmen mit Intelligenz. Anhand von Data Mining gewonnene Erkenntnisse können für profitablere Entscheidungen, Initiativen und den eigenen Wettbewerbsvorteil eingesetzt werden. Es erlaubt einem Unternehmen, riesige Datenmengen, die durch Kontakte mit Kunden und Interessenten entstehen, nutzbar zu machen, um mehr über diese zu erfahren.

- Welche Kunden werden am ehesten auf ein Mailing reagieren?
- Gibt es Kunden(segmente) mit ähnlichen Eigenschaften und typischem Verhalten?
- Bestehen interessante Zusammenhänge zwischen Kundenmerkmalen?
- Welches sind die loyalen Kunden und wer steht kurz davor, verloren zu gehen?
- Wo sollte die nächste Niederlassung angesiedelt werden?
- Welches Produkt oder welche Dienstleistung wird dieser Kunde als nächstes nachfragen?
- Woran erkennt man missbräuchliche Transaktionen?
- Wie hoch ist die Life-Time-Profitabilität dieses Kunden?

Die Antworten auf solche Fragen liegen in Unternehmensdaten verborgen und es braucht mächtige Data-Mining-Werkzeuge, um an sie heranzukommen. Data Mining darf hier nicht ignoriert werden – Daten und Data-Mining-Techniken sind zahlreich vorhanden und die Vorteile, die Data Mining für ein Unternehmen bietet, enorm. Unternehmen, deren Bemühen im Data Mining auf einer «Mythodologie» basiert, werden sich in einem ernsthaften Wettbewerbsnachteil gegenüber solchen Unternehmen sehen, die einen rationalen, informationsbasierten Ansatz verfolgen. Oder, um es mit Joseph P. Bigus zu sagen: «Wenn Sie trotz aller Vorzüge kein Data Mining betreiben, müssen sie sich vorwerfen lassen, eine der wichtigsten Ressourcen Ihres Unternehmens ungenutzt zu lassen».

Schlussbemerkung

Data Mining basiert auf drei aufeinander abgestimmten Standbeinen: Computerwissenschaft, Statistik und Businesswissen. Nur eine oder zwei Säulen genügen nicht, selbst drei Standbeine nicht, solange sie nicht ausbalanciert sind. Erfolgreiches Data Mining erfordert erhebliche Zusammenarbeit der drei Bereiche. Alle Beteiligten müssen ihren Horizont weiten, damit echte Zusammenarbeit zustande kommt und die Suche nach dem «Gold» Realität wird. Die entscheidende Herausforderung liegt darin, die Chance auf einen gemeinsamen Erfolg zu erkennen.
Viele Data-Mining-Analysen lassen sich leicht durchführen, sobald man das benötigte «Black-Box»-Software-Paket bedienen kann. Dieser unbedarfte Umgang birgt jedoch die Gefahr des Missbrauchs und weist deutliche Fallstricke auf. Er kann zu praktisch wertlosen Ergebnissen und irreführenden Schlussfolgerungen führen – und tut dies wahrscheinlich auch häufig. Data Mining kann leicht misslingen. Es ist daher wichtig, dass sowohl Anwender als auch der Konsument genügend über die Eigenschaften des Data Minings wissen, d.h. über Vorteile und Fehlerquellen. Erst dann kann eine sachkundige Entscheidung getroffen werden, welche Methoden zum Einsatz kommen sollen. Und erst dann lassen sich eigene Ergebnisse und die anderer beurteilen. Idealerweise sollte dieses Verständnis unabhängig von einem Softwareanbieter erworben werden.



Über den Autor

Der promovierte Statistiker Dr. Diego Kuonen ist Gründer und CEO von Statoo Consulting, Schweiz. Statoo Consulting ist ein softwareunabhängiges Schweizer Beratungsunternehmen mit Schwerpunkt auf statistische Beratung und Schulung, Datenanalyse, Data Mining und Dienstleistungen für analytisches CRM. Dr. Diego Kuonen gibt seit einigen Jahren in ganz Europa softwareunabhängige Methodenkurse in statistischem Data Mining, für namhafte internationale Unternehmen, aber auch öffentlich. Gegenwärtig ist er ebenfalls Vizepräsident der «Schweizerischen Gesellschaft für Statistik» und dort Präsident der Sektion «Statistik in Business und Industrie».

Sind Sie schon «Statoowiert» worden? Falls nicht, finden Sie weiterführende Informationen auf www.statoo.info/.