Data Mining

<<TableOfContents>>

= Data Mining =

== Data Mining als Teilbereich der KI ==

WikiPedia:Data-Mining ist ein Teilbereich der Künstlichen Intelligenz (KI) in der Informartik. WikiPedia:Data_Mining ist ein vereinfachter Begriff für Knowledge Discovery in Databases (KDD). D.h. KDD ist der Versuch aus vorhandenen Datenmassen in Datenbanken Erkenntnisse zu bekommen. Bei nicht als Datenbanken vorliegenden Texten wird dabei zuerst [[http://wissensexploration.de/textmining-vs-datamining.php|Text Mining]] betrieben.

Bei Zahlendaten ist dieses nicht anderes als ein numerisches Verfahren, d.h. mathematische Funktionen werden durch Polynome angenähert, also eine kompliziertere Version der linearen Regression.

Bei nicht numerischen Daten versucht Data-Mining ebenfalls die Daten durch eine Funktion zu approproximieren. Diese ist dann allerdings keine analytische Funktion, sondern eine abstrahierte Funktion. Diese abstrahierte Funktion wird als vorhandenes Muster in den Daten bezeichnet, welches in der Regel zu einem diskreten Wert führt.

Zum Finden der Muster oder Funktionen gibt es verschiedene Verfahren, welche je nach vorhandenen Daten ausgewählt werden müssen. Zum Finden des Musters oder der Funktion werden Methoden des [[http://wissensexploration.de/datamining-kdd-machine-learning.php|Maschinelles Lernen]] verwendet. Dabei finden bei Data-Mining nur Methoden des induktiven Maschinellen Lernens Anwendung (deduktives Maschinelles Lernen wird für Automatisierungen verwendet).

== Vorgehensweise ==

Teilweise werden zuerst verschiedene [[Datenbanken]] zu einer zusammengeführt, dieses wird WikiPedia:Datawarehouse genannt. Als nächstes werden die Daten von Fehlern bereinigt. Dann werden aus den Datensätzen sogenannte Trainingsdaten ausgewählt. Dieses ist ein heikler Punkt, da bei falscher Auswahl der Trainingsdaten es zur Überanpassung und somit fehlerhaften Ergebnissen führen kann. Dann wird eine der Methoden des Data-Mining ausgewählt um die abstrahierte Funktion oder das Muster zu finden, wobei hier das Problem ist die passende Methode auszuwählen. Dafür bedarf es theoretisch die genaue Kenntnis der Methoden. Anschließend wird die gefundene Funktion mit Testdaten, die ebenfalls aus den Datensätzen stammen, überprüft und notfalls verbessert.

== Einige Methoden des Data-Mining ==

=== Klassifizierung ===

Bei der WikiPedia:Statistical_classification wird versucht in den vorhandenen Daten allgemeinen Strukturen zu finden, so dass Rückschlüsse neuen Daten gezogen werden können, d.h. es wird eine Abbildungsfunktion auf einen diskreten Wert (Klasse) oder einen numerischen Funktionswert gesucht. Methoden der Klassfizierung sind u.a. Nearest Neigbour Methode, Entscheidungsbäume und Neuronale Netze.

==== K-Nearest Neighbor Methode ==== Die WikiPedia:K-nearest_neighbor_algorithm wird auch als Lazy Learning bezeichnet, d.h. von allem vorhandenen Datensätzen werden K-Datensätzen gesucht, die den neuen Daten am ähnlichsten sind. Dabei werden bei nicht numerischen Daten abstrahierte Abstände verwendet. Die gesuchten Werte des neuen Datensatzes werden dann mit Gewichtung 1/Abstand aus den nächsten K-Datensätzen bestimmt.

==== Entscheidungsbäume ====

Beim [[Lernen aus Entscheidungsbäumen]] (WikiPedia:Decision_tree_learning) werden die Daten in einem abstrakten Baum dargestellt, wobei jede Verästelung eine Entscheidung darstellt. Es wird dabei von der Wurzel angefangen und bei jeder Verästelung wird ein Attribut abgefragt und die nächste Verästelung ausgewählt. Diese Prozedur wird so lange fortgesetzt, bis das Baumende erreicht ist. Da es verschieden Möglichkeiten für die Bäume gibt wird per Algorithmus versucht den Kürzesten und Optimalsten (d.h. mit dem kleinsten Fehler) zu finden. [[Lernen aus Entscheidungsbäumen]] ist eines der am häufigsten benutzten Verfahren beim Data Mining.

==== Naive Bayes ====

Bei Naive_Bayes_classifier wird vereinfacht (d.h. naiv) angenommen, dass alle Attribute der Datensätze mit einer von einander unabhängigen Wahrscheinlichkeit ein Klasse oder einen Funktionswert ergeben (D.h. die Wahrscheinlichkeit mit der jeweils islamischer Glauben, technischer Studiengang, ... einen Terroristen ausmacht). So kann das Bayes Theorem in einer vereinfachten Version verwendet werden)Wenn diese vereinfachende Annahme nicht gemacht würde, wären die Ergebnisse nur durch erheblichen rechnerischen Aufwand zu erzielen). Die Werte für die unabhänig angenommenen Wahrscheinlichkeiten lassen sich dann mit Hilfe der Trainingsdaten und dem Bayes Theorem berechnen.

==== Künstliche Neuronale Netze ====

Mit neuronalen Netzen (WikiPedia:Artificial_neural_networks)versucht man laut einer [[http://www.ai.wu.ac.at/~koch/courses/wuw/archive/inf-sem-ws-00/nentwich/index.htm#413|Seminararbeit an der Wiener Uni]] die Vorgänge im menschlichen Gehirn nachzubilden. Das Wissen zur Lösung einer Aufgabe wird in den Neuronen (den Knoten) eines Netzes abgelegt, zwischen denen dann Verbindungen (links) hergestellt werden. Die Knoten entsprechen dabei einem Neuron des menschlichen Gehirns, die Kanten stellen Verbindungen zwischen Neuronen im menschlichen Gehirn dar. Es ist dem Entscheidungsbaumverfahren sehr ähnlich allerdings erweitert es seine Parameter selbstständig, um genauere Schlüsse zu ziehen.

=== Assoziationsregeln === Beim WikiPedia:Association_rule_learning werden die gegenseitige Abhängigkeit von Attributen mit Hilfe von Wahrscheinlichkeitsrechnung bestimmt. (wie z.B. dicke Jacke, Selbstmordattentäter)

=== Clustering ===

WikiPedia:Cluster_analysis ist das Einteilen der Datensätze in verschiedene Mengen (d.h. Clustern) mit ähnlichen Eigenschaften.

==== k-means Clustering ==== Beim k-means Clustering werden die Daten in eine vorher festgelegt Anzahl von Clustern (nämlich k) eingeteilt.

==== Hierachisches Clustering ==== Beim Hierachischen Clustern wird zuerst jeder Datensatz als ein Cluster angenommen um dann suksessiv die Anzahl der Cluster zu verkleinern.

=== Regression ===

Die WikiPedia:Regression_analysis wird mit Hilfe von numerischen Methoden eine mehrdimensionale Funktion für die Daten approximiert.

== Anwendungen von Data-Mining ==

=== SPAM Erkennung ===

Spam_(electronic) Erkennungsprogramm verwenden Data-Mining, dabei wird [[http://de.wikipedia.org/wiki/Bayes-Klassifikator#Beispiel|Naive Bayes]] verwendet. SPAM-Erkennung ist ein gutes Beispiel für das was Data Mining kann und was nicht. Es ist kein Problem für einen Menschen Mails in Spam und Nicht-Spam einzuteilen, Merkmale die in allen Spams vorkommen rauszufinden ist dagegen schwieriger. Bei Anwendung von Data Mining zur SPAM-Erkennung wurde entdeckt, dass die meisten SPAM-Mails die Zeichenfolge FF0000 enthielten, da wäre manuell niemand drauf gekommen, es ist aber logisch da es in WikiPedia:HTML zur Erzeugung von roter Schrift dient.

=== Scooring ===

Beim [[Private Datenbanken#Scoring_Datenbanken|Scooring]] werden Entscheidungsbäume und andere Methoden der Klassifizierung verwendet um die Menschen in kredit- und nicht kreditwürdig einzuteilen.

=== Operative Fallanalyse ===

Zur Unterstützung der [[operativen Fallanalyse]] werden Data-Mining Programme verwendet. In der BRD werden dafür meist, die bei der Polizei vorhandenen Daten und Vekehrsdaten verwendet.

'''vgl [[Länderübergreifende Software#Data-Mining_Software]]'''

=== Vorratsdatenspeicherung ===

Es gibt laut [[http://www.heise.de/newsticker/meldung/ETSI-legt-Standards-zum-Data-Mining-bei-der-Vorratsdatenspeicherung-fest-178769.html|Heise-Newsticker]] eine WikiPedia:ETSI Richtlinie, die die Anwendung von Data Mining auf die [[Vorratsdatenspeicherung]] regelt. In einem [[http://derstandard.at/1297216314225/WebStandard-Interview-Wir-steuern-einem-Paranoiastaat-entgegen|Interview mit dem Standard]] kritisiert der Obmann der [[Österreich|österreichischen]] Bürgerrechtsorganisation [[http://www.argedaten.at/|Arge-Daten]], das die Anwendung von Data Mining auf die Vorratsdatenspeicherung dazu führen würde das wir uns als Bürger immer öfter für unsere Taten rechtfertigen müssen, aber nicht weil sie illegal sind, sondern bloß weil sie einem verdächtigen Muster entsprechen würden.

=== Überwachung von Beschäftigen ===

Laut einem [[https://archiv.foebud.org/bba/docs/bba_ts021023_lischka_kundeDiebUndCDU-Waehler.html|Tagespiegelartikel]] vom Oktober 2002 werden Data Mining Tools des Stinnes-Tochterunternehmen Logware verwendet um aufspüren, welcher Angestellte an der Kasse unterschlägt. Alle Vorgänge wie Mitarbeiterkauf, Korrektur, Umtausch, Storno, Rückgabe, Öffnung der Kassenschublade werden an eine [[Datenbank]] übermittelt und zu *Kassierprofilen* verdichtet. Wer von diesen abweicht, macht sich verdächtig. Die Ketten Edeka, Kaufhof und Toom würden das Programm bereits einsetzen.

=== Verbrechensvorhersage (Predictive Policing) ===

Die Polizei in Chicago hat die Daten von Verbrechensmeldungen mit Wetterdaten, geographischen Daten, Verkehrsaufkommen verknüpft und somit Tatort und Tatzeit von möglichen Verbrechen vorhergesehen. Laut einem [[http://www.schneier.com/blog/archives/2007/08/police_data_min.html|Blog-Artikel]] von 2007 des Computer-Sicherheitsexperten WikiPedia:Bruce_Schneier ist das ein Beispiel für eine sinnvolle Anwendung von Data Mining durch die Polizei. Wogegen in einem [[http://www.zeit.de/digital/datenschutz/2011-08/predictive-policing/komplettansicht|Zeit-Artikel]] von 2011 die Gefahr gesehen wird, dass durch diese Verbrechensvorhersagen ein "diffuses Gefühl des Beobachtetseins" enstehen könnte, weil man befürchtet, verdächtigt zu werden, nur weil man aus irgendeinem Grund in Gebieten mit hoher Kriminalitätswahrscheinlichkeit unterwegs ist. Dann meidet man diese Gebiete künftig möglicherweise lieber, selbst wenn man sich nichts hat zu Schulden kommen lassen. Dazu müsse man nicht einmal wissen, ob ein Gebiet wirklich ein Ort häufiger Verbrechen ist – für das subjektive Empfinden reiche es, wenn man nur glaubt, an einem solchen Ort zu sein.

=== Total Information Awarness Program ===

Im Zuge der Terror-Hysterie nach dem 9/11 Anschlag auf das World Trade Center, starteten die amerikanischen Sicherheitsbehörden das TIA-Programm. Bekannt wurde es 2002 und im Jahre 2003 auf Grund von öffentlicher Proteste dann beendet. Laut einem [[http://www.schneier.com/essay-163.html|Artikel]] von WikiPedia:Bruce_Schneier werden zahlreiche andere Data-Mining Projekte von den [[USA|amerikanischen]] Sicherheitsbehörden weiterbetrieben.

=== INDECT ===

Im Rahmen der [[Datenbanken EU|EU]] gibt es ein Projekt namens [[INDECT]] indem alle Datenbanken, alle durch [[Überwachungstechnik]] aufgezeichneten Daten und durch Software Agenten im Netz gefundenen Daten mit Hilfe von Data-Mining geplante Verbrechen vorhergesehen werden sollen.

== Kritik ==

Von Data Mining wird laut der [[http://www.ai.wu.ac.at/~koch/courses/wuw/archive/inf-sem-ws-00/nentwich/index.htm#A|Kritik in einer Seminarbeit an der Wiener Uni]] häufig geglaubt, es diene dazu, Zusammenhänge automatisch zu entdecken, an die bisher noch nicht einmal jemand gedacht hat, und Fragen zu beantworten, die nicht einmal noch jemand gestellt hat. Diesem wird dort wiedersprochen. "Schlaue" Data Mining Tools können danach kein profundes Know How ersetzen.

In einem [[http://www.faz.net/s/Rub117C535CDF414415BB243B181B8B60AE/Doc~E38A2F6DD0A734EB789AAD27EDE6F9A35~ATpl~Ecommon~Scontent.html|FAZ-Artikel über Data-Mining]] von [[http://frank.geekheim.de|Frank Rieger]] wird die Problematik von Data-Minig sehr anschaulich beschrieben:

{{{#!blockquote Die Profile sind nützlich, um uns gezielt zum Kauf von mehr nutzlosem Tand oder interessanteren Büchern zu verleiten, uns effizienter zu verwalten und zukünftiges Verhalten zu prognostizieren. Und um Menschen unter präventive Überwachung zu stellen, deren Profil sich bedenklich dem von Straftätern nähert. Dabei geht es nicht um hundertprozentige Präzision der Vorhersage. Wahrscheinlichkeiten, Neigungen, Tendenzen, Zugehörigkeit zu Kohorten sind die Währungen der algorithmischen Orakel. }}}

''Anmerkung: Bei falscher zielgruppenorientierter Werbung ist das für den Beworbenen erstmal lustig oder auch nur nervig. Bei negativen [[Private Datenbanken#Scoring_Datenbanken|Scooring]] der Kreditwürdigkeit ist das ärgerlich, bei der falscher Verdächtigung von der [[Datenbanken der Bundespolizeien|Polizei]] kann es dann allerdings richtig unangenehm werden.''

=== Problematik bei der Anwendung im Polizei-Bereich ===

In einem [[http://www.nytimes.com/2006/05/16/opinion/16farley.html|Artikel der New York Times vom 16.5.2006]] kritisiert Jonathan Farley, dass graphentheoretische Methoden (d.h. die Analyse von Telekommunikationsnetzen) zur Identifikation von Terroristen ungeeignet seien, einerseits weil, wie im [[http://de.wikipedia.org/wiki/Kleine-Welt-Ph%C3%A4nomen|kleine Welt-Experiment]] von Stephan Milgram gezeigt, andererseits weil "Schläfer" ohnehin ganz normale Kommunikationsprofile haben.

Zu ähnlichen Schlüssen ist der [[USA|US-amerikanische]] National Research Council nach einem [[http://news.cnet.com/8301-13578_3-10059987-38.html?part=rss&subj=news&tag=2547-1_3-0-20|Artikel von Cnet news]] gekommen:

{{{#!blockquote But the authors conclude the type of data mining that government bureaucrats would like to do--perhaps inspired by watching too many episodes of the Fox series 24--can't work. "If it were possible to automatically find the digital tracks of terrorists and automatically monitor only the communications of terrorists, public policy choices in this domain would be much simpler. But it is not possible to do so." }}}

Krytographieguru WikiPedia:Bruce_Schneier kritisiert in einem [[http://www.wired.com/politics/security/commentary/securitymatters/2006/03/70357|Wired-Artikel von 2006]], die zu erwartende Zahl falscher Positiver sei groß. Terroristen-Plots seien nicht so simpel, wie z.B. die Identifizierung von gestolenen Kreditkarten, welche sich gut durch Data-Mining identifizieren ließen:

{{{#!blockquote Terrorist plots are different. There is no well-defined profile and attacks are very rare. Taken together, these facts mean that data-mining systems won't uncover any terrorist plots until they are very accurate, and that even very accurate systems will be so flooded with false alarms that they will be useless. }}}

''Anmerkung: Gerade im Sicherheitsbereich, wenn [[Länderübergreifende Software#Data-Mining_Software|Data-Mining Software bei der Polizei]] dazu benutzt wird aus polizeiliche Datenbanken neue Erkenntnisse zu gewinnen, kann es dazu führen, dass nicht mehr ergebnissoffen in alle Richtungen ermittelt wird. Bei einem Serientäter kann es so ersten zu Problem für die Allgemeinheit werden, da so der oder die Täter_in nicht gefunden wird und zweitens kann es für die eventuell Falschverdächtigten zu den üblichen Nachteilen führen. Ganz problematisch wird es allerdings, wenn geglaubt wird mittels Data Mining Prognosen über zukünftige Straftäter anstellen zu können.''

=== Data Mining bei der Anti-Terror-Datei ===

In einem [[http://www.heise.de/tp/artikel/37/37967/1.html|Telepolis-Artikel]] zur Klage vorm Bundesverfassungsgericht wegen der [["Anti-Terror-Datenbank"]] wird beschrieben wie durch Data Mining aus einer zur Unrecht in der Datenbank gelandeten Person ein Verdächtiger wird.

== Weitere Infos ==

- [[http://www.hs-weingarten.de/~ertel/index.php?page=buch-ki|Grundkurs Künstliche Intelligenz von Wolfgang Ertel, Vieweg-Verlag]] -- bietet eine einfache und theoretisch fundierte Einführung in Künstliche Intelligenz; es gibt auch ein Kapitel zu Data Mining
- [[http://www.ai.wu.ac.at/~koch/courses/wuw/archive/inf-sem-ws-00/nentwich/index.htm|Seminararbeit zu Data-Mining]] -- bietet eine Übersicht
- [[http://dbs.informatik.uni-halle.de/Lehre/KDD_SS09_web/dm_skript.pdf|Data-Mining Skript zu einer Vorlesung in Halle]] (pdf)
- <<Rellink(/gc/html/datamining.html,Artikel in der RHZ zu Data Mining)>>
- [[http://www.datenminen.net|Blog zur Kritik von Data Mining]]

Version #1
Erstellt: 2025-10-27 22:35:59 UTC von Datenschmutz Migration Bot
Zuletzt aktualisiert: 2025-10-27 22:35:59 UTC von Datenschmutz Migration Bot