Data Mining –
Suche nach verborgenen Mustern

Hochschulreihe FH Brandenburg

Dipl.-Inform. I. Boersch
FB Informatik und Medien

Januar 2009

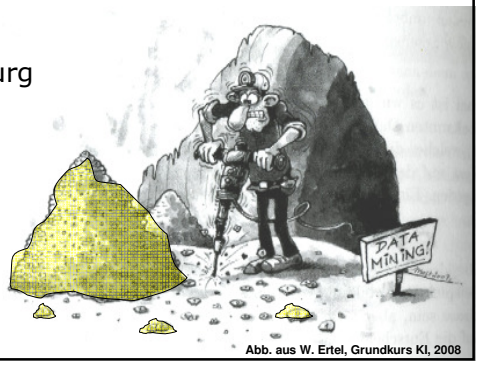


Abb. aus W. Ertel, Grundkurs KI, 2008

Ausgangsfrage

Ist es möglich, die angefallenen und weiter anfallenden riesigen Datenbestände in nützliche Informationen oder sogar Wissen umzuwandeln?

Falscher Alarm auf der Intensivstation

- Monitore überwachen Patientenzustand
- Schwellwerte -> Alarm in kritischen Situationen

- Problem: viele Fehlalarme -->
  Aufmerksamkeit sinkt, Schwellwerte werden erhöht,
  Messungen werden deaktiviert

- Ziel: Erkennen, wann ist ein Alarm wichtig,
  NB: max. 2% echte Alarme verpassen (sens=0.98)

- Ergebnis des Data Mining: 33% weniger Fehlalarme
- TU Dortmund, Universitätsklinikum Regensburg, 2007

[SG07]

Heute

- DM und KDD, Phasen
- Aufgabenstellungen des Data Mining
- Eine Methode – Entscheidungsbaumlernen
  - Weka
  - Spongebob und Crabs
- DM und Ethik

- Vortrag im Netz: http://ots.fh-brandenburg.de/dm.pdf
- Kontakt: boersch(at)fh-brandenburg.de

Einordnung und Begriff KDD und DM

„Knowledge Discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. …

gültige

neue

Data mining is a step in the KDD process …
[FPSS96]

nützliche

verständliche Muster

- Heute oft vereinfacht KDD = DM

Definitionsüberblick in Wiedmann2001

Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic

Phasen des KDD

Datenselektion / -extraktion
- Welche Daten notwendig und verfügbar?

Datenreinigung und Vorverarbeitung
- Fehlende Werte, Ausreißer, Inkonsistenzen

Datentransformation
- Format für DM (einzelne Tabelle), Aggregation, Aufteilung in Trainings- und Testdaten

Data Mining (10 .. 20% Zeitaufwand)
- Finden von Mustern

Interpretation und Evaluation
- Präsentation
- Bewertung mit Testdaten

Data Mining „Finden von Mustern“

- ein Prozess erzeugt Daten

⇒ Beschreibung, Vorhersage gewünscht

⇒ Modelle / Konzepte / Wissensrepräsentationen sind zu finden / zu konstruieren / zu optimieren

⇒ interdisziplinär: Maschinelles Lernen, Mustererkennung, Statistik, Datenbanken, ...

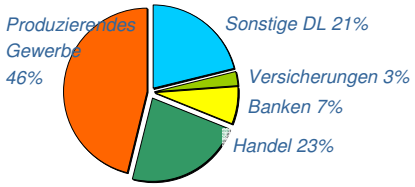
⇒ Querschnittstechnologie

Anwendungsbeispiele

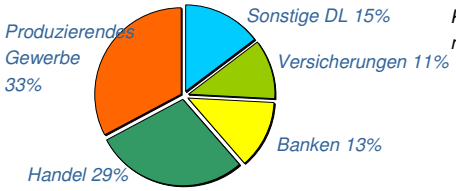
- Wie erkennt der Händler, ob es sich bei einer Bestellung um einen zahlungswilligen Kunden handelt, der letztendlich die Ware auch bezahlt?
- Wovon hängt es bei der Evaluierung im FBI ab, ob ein Student in der Veranstaltung subjektiv „viel gelernt“ hat?
- Wie beeinflussen Einstellungen am Produktionsprozess die Qualität des Produktes?
- Wie erkennt man anhand der Nervenfaserschichtdicken der Retina eine Augenkrankheit?
- ...

Befragung 2002
Analysieren Sie Ihre Kunden?

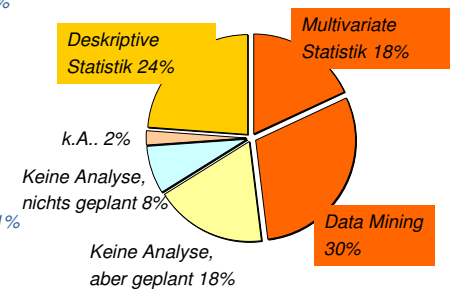
- 734 Unternehmen kontaktiert*



- 103 haben geantwortet



- Art der Kundenanalysen:



* die 734 größten deutschen Unternehmen
Hippner, H., Merzenich, M. und Stolz, C. Data Mining: Anwendungspraxis in deutschen Unternehmen. In: Wilde, K.D., Data Mining Studie, absatzwirtschaft, 2002

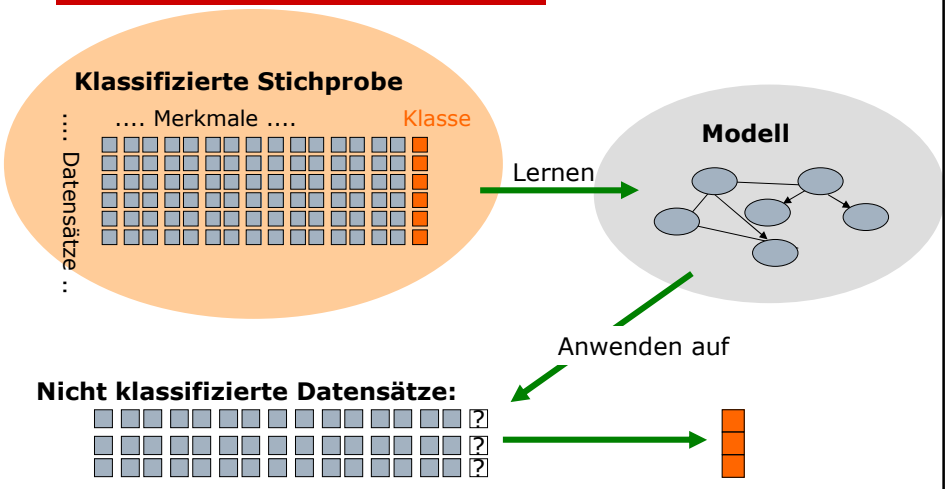
Aufgabenstellungen des Data Mining

- Klassifikation
- Numerische Vorhersage
- Abhängigkeitsanalyse
- Clustering
- Abweichungsanalyse

Klassifikation

- Lernverfahren nimmt eine Menge klassifizierter Beispiele entgegen, aus denen es lernen soll, unbekannte Beispiele zu klassifizieren.
- Gegeben: Klassifizierte Stichprobe
- Gesucht: Modell zum Beschreiben und Vorhersagen von Klassen
- Gut lesbare Modelle:
  - Entscheidungsbäume, Regelmengen, Bayessche Netze, Fuzzy-Systeme …
- Andere:
  - Neuronale Netze, logische Ausdrücke, SVM …

Ablauf



Klassifikation – Beispiel (DMC06)

- Vorhersage - erzielt eine ebay-Auktion einen überdurchschnittlichen Verkaufserlös?

Stichprobe: 8.000 Online-Auktionen der Kategorie "Audio&Hi-Fi:MP3-Player:Apple iPod„

Merkmale: Titel, Untertitel, Zeitpunkte, Dauer, Rating, Startpreis, Sofortkaufen, Galerie, Fettschrift ...

Klasse: Hochpreis oder nicht

- Anwenden auf 8000 unklassifizierte Auktionen -> Ergebnis einsenden, Ranking

file:///imVortrag/dmc2006_train.txt

DM-Cup 2008 - Kündigerprävention

- 121. SKL: Kündigungskandidaten identifizieren
- 113.477 Datensätze
  - Personendaten (Alter, Geschlecht)
  - abgeleitete Personendaten (Bankart, Telefonart)
  - Marketingdaten (Werbeweg, Responseweg)
  - Spieldaten (gewünschter Einsatz, div. Spielparameter)
  - Kontaktinformationen (Kategorien: Information, Reklamation)
  - 50 Variablen, wie z.B. Altersverteilung, Kfz-Typen und -verteilung, Kaufkraft, Gebäudetypen, Konsumaffinitäten.
- Klasse:
  - Bezahlt gar nicht
  - Bezahlt nur die 1. Klasse
  - Bezahlt bis einschließlich 2. Kl
  - Bezahlt bis einschließlich 6
  - Bezahlt mind. bis einschlie

Rückblick DMC Wettbewerb 2008
Anmeldungen: 618
Beteiligte Universitäten: 164
Beteiligte Länder: 42
Eingereichte Lösungen: 231

Herzinfarkterkennung in der Notaufnahme

- Patienten mit Brustschmerz in Notaufnahme in *Edinburgh* (1252) *und Sheffield (500)*
- 45 Merkmale:
  age, smoker, ex-smoker, family history of MI, diabetes, high blood pressure, lipids, retrosternal pain, chest pain major symptom, left chest pain, right chest pain, back pain, left arm pain, right arm pain, pain affected by breathing, postural pain, chest wall tenderness, sharp pain, tight pain, sweating, shortness of breath, nausea, vomiting, syncope, episodic pain, worsening of pain, duration of pain, previous angina, previous MI, pain worse than prev. Angina, crackles, added heart sounds, hypoperfusion, heart rhythm, left vent. hypertrophy, left bundle branch block, ST elevation, new Q waves, right bundle branch block, ST depression, T wave changes, ST or T waves abnormal, old ischemia, old MI, sex
- Woran erkennt man den Herzinfarkt (MI)?
- -> Entscheidungsbaum [TFLK98]

Ergebnisbaum Herzinfarkterkennung


```
ST elevation = 1: 1
ST elevation = 0:
| New Q waves = 1: 1
| New Q waves = 0:
| | ST depression = 0: 0
| | ST depression = 1:
| | | Old ischemia = 1: 0
| | | Old ischemia = 0:
| | | | Family history of MI = 1: 1
| | | | Family history of MI = 0:
| | | | | age <= 61 : 1
| | | | | age > 61 :
| | | | | | Duration of pain (hours) <= 2 : 0
| | | | | | Duration of pain (hours) > 2 :
| | | | | | | T wave changes = 1: 1
| | | | | | | T wave changes = 0:
| | | | | | | | Right arm pain = 1: 0
| | | | | | | | Right arm pain = 0:
| | | | | | | | | Crackles = 0: 0
| | | | | | | | | Crackles = 1: 1
```

Sensitivity = 81.4%
Specificity = 92.1%
PPV = 72.9%
Accuracy = 89.9%

Nur 10 Merkmale

Numerische Vorhersage

- Variante der Klassifikation
- Das vorhergesagte Ergebnis ist keine diskrete Klasse, sondern eine numerische Größe.

- Modelle:
  - Formeln und Parameter (z.B. Evolutionäre Algorithmen: GA, GP)
  - Regressionsbäume, Modellbäume
  - Modelle der Klassifikation
  - Nächster Nachbar …

Numerische Vorhersage - Beispiel

- Sagen Sie aus den Konfigurationsdaten eines PCs (cycle time [ns], main memory [KB, min, max], cache size [KB, min, max], channels) die Performance vorher.

→ Lineares Regressionsmodell

$$p = 0.0661 * MYCT + 0.0142 * MMIN + 0.0066 * MMAX + 0.4871 * CACH + 1.1868 * CHMAX + -66.5968$$

- Gegeben sind Studentendatensätze mit Notenprofil. Erstellen Sie ein Modell zur Vorhersage der Note der Abschlussarbeit.

  Möglich?

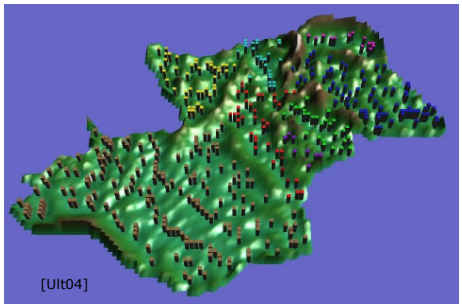
Clustering
(auch Segmentierung, Gruppenbildung, Clusteranalyse)

- Aufspaltung unklassifizierter Daten in interessante und sinnvolle Teilmengen / Klassen
- Innerhalb der Klasse möglichst ähnlich, zwischen den Klassen möglichst unähnlich
- Modelle:
  - Cluster-Repräsentanten k-Means, EM
  - Selbstorganisierende Karten
  - Hierarchisches Clustering
  - ...

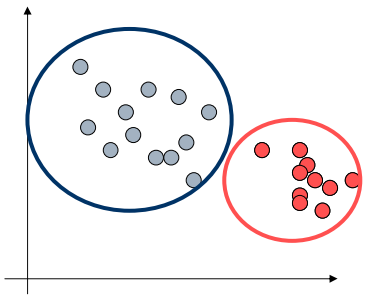
Clustering - Beispiel

Kundensegmentierung

„Die Anwendung auf das Kundenverhalten einer Mobilfunk Gesellschaft ermöglichte eine Klassifizierung der Kunden in verschiedene Gruppen. Hierin konnte neues Wissen über das Verhalten von Kundensegmenten gewonnen werden. Von besonderem Interesse waren die Kunden, die mit hoher Wahrscheinlichkeit bald den Vertrag kündigen. „



- [Ult04]



Abhängigkeitsanalyse

- Finden von Abhängigkeiten zwischen Attributen
- Modelle:
  - Assoziationsregeln
- Assoziationsregeln können
  - jedes einzelne Attribut vorhersagen, nicht nur das Klassenattribut
  - mehrere Attributwerte gleichzeitig vorhersagen

Abhängigkeitsanalyse - Beispiel

- Einkaufswagenanalyse
- (Windeln, Freitag) Bier

Deutung:

- Schatz, bring noch Windeln mit.
- Pub schaffe ich nicht mehr, da nehm‘ ich gleich Bier mit.
- Vermutlich moderne Sage

Abweichungsanalyse

- Ursachen für untypische Merkmalsausprägungen der Ausreißer finden
- Liegen falsche (oder alte) Grundannahmen über den Datengenerierungsprozess vor?
- Sonst entfernen oder reparieren.

Des einen Ausreißer ist des anderen Datenpunkt.

Aufgabenstellungen des Data Mining

Klassifikation
Numerische Vorhersage
Abhängigkeitsanalyse
Clustering
Abweichungsanalyse

Wie funktioniert es?

- Lösungsansätze

Anwendbarkeit beim Data Mining

im Vortrag

Abb. aus W. Ertel, Grundkurs KI, 2008

Anwendbarkeit beim Data Mining

im Vortrag

27

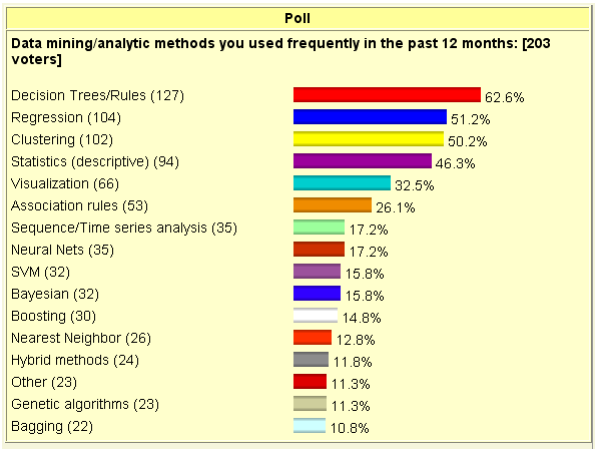
Abb. aus W. Ertel, Grundkurs KI, 2008

Heute

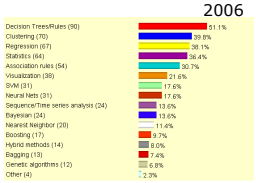
- DM und KDD, Phasen
- Aufgabenstellungen des DM, Beispiele
- **Eine Methode – Entscheidungsbaumlernen**
  - Weka
  - Spongebob und Crabs
- DM und Ethik

- Vortrag im Netz: http://ots.fh-brandenburg.de/dm.pdf
- Kontakt im WWW

Data mining/ analytic methods you used frequently in the last year (Umfrage KD-Nuggets März 2007)



2006



[http://www.kdnuggets.com/polls/index.html]

Ein Spiel an frischer Luft

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

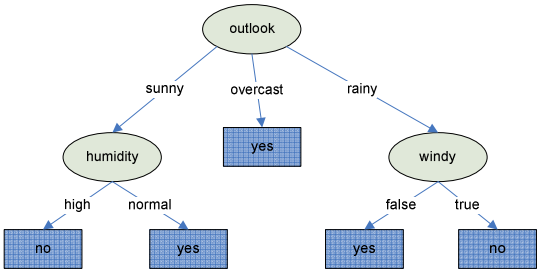
ARFF-Datenformat

```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```


Ein Entscheidungsbaum

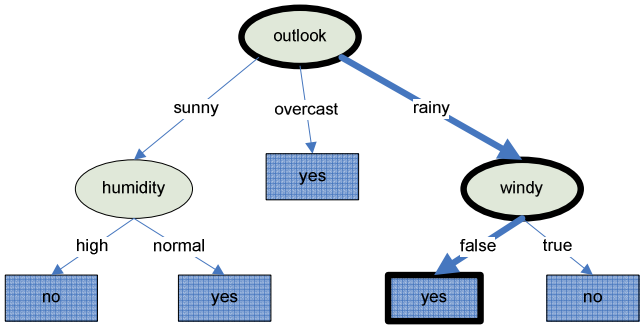


Knoten: Testfunktion für ein Attribut,
für jedes Ergebnis ein Nachfolgerknoten
Blatt: Klassifikation

Verschiedene Bäume möglich –
Welcher ist der beste (einer der besten)?

Wie wird klassifiziert?

- Welche Klasse hat der Datensatz Nr. 5?
  {Outlook=rainy, temperature=cool, humidity=normal, windy=false}



Lernen von Entscheidungsbäumen

- Der beste Baum?
  - Gültig, flach, geringe Knotenanzahl, gleichmäßige Datendichte, übersichtlich …=> **„kleiner Baum"**
- Anzahl der möglichen Entscheidungsbäume gigantisch
- Durchprobieren aller Bäume undurchführbar
- **=> Ein Suchproblem!**
- Gierige (greedy) Strategie ähnlich Bergsteigen – bergauf losmarschieren und niemals umdrehen

Top-Down Induction of DecisionTrees (TDIDT)

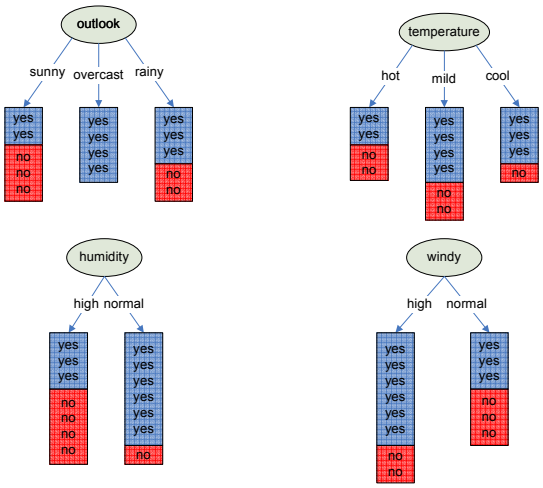
Grundidee TDIDT

- TDIDT reduziert die Suche nach dem besten Baum auf die Bestimmung des besten Attributes

1. Wähle das **beste Attribut A** für den aktuellen Knoten.
2. Für jeden Wert von A erzeuge einen Nachfolgeknoten und markiere die Kante mit dem Wert.
3. Verteile die aktuelle Beispielmenge auf die Nachfolgeknoten, entsprechend den jeweiligen Werten von A.
4. Wende TDIDT auf die neuen Blattknoten an (Rekursion)

- Heuristische Suche
- Keine Optimalitätsgarantie

Das beste Attribut?

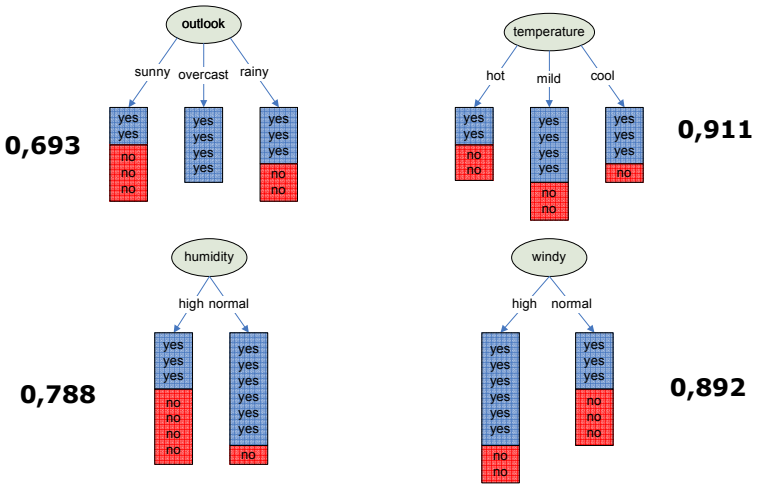


Welches ist das beste Attribut?

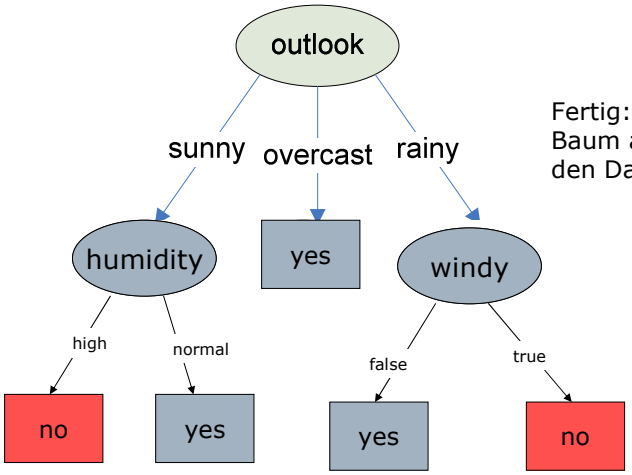
- Welche Frage stellt der Arzt / Kfz-Mechaniker / PC-Experte / Kundenberater als **erste** (, um möglichst schnell zum Ziel zu kommen)?
- Erinnerung:
  - Ziel: möglichst kleiner Baum
  - Bei Knoten mit Beispielen einer Klasse (reiner Knoten) ist die Zerlegung beendet

=> **<u>Gierige Heuristik</u>:**
Wähle das Attribut, das die "reinsten" Knoten erzeugt

Unreinheit der Zerlegung E(H)



Aufbau des Baumes



Fertig:
Baum automatisch aus den Datensätzen erstellt

Das Wetterbeispiel in WEKA

Weka:

- Data Mining Software
- Waikato Environment for Knowledge Analysis
- University of Waikato (Neuseeland)
- www.cs.waikato.ac.nz/ml/weka
- Java, GUI, GPL, erweiterbar
- ARFF-Files als Dateneingabe

Der Vogel Weka
http://danny.oz.au/travel/new-zealand/p/1289c-weka.jpg

Spongebob und Crabs

- Lernen von Objektbeschreibungen durch Zeigen von Beispielen
- Diplomarbeit **Benjamin Kieper**: *Entwurf und Implementierung einer Anwendung zum dialogbasierten, überwachten Lernen von Objektmodellen aus Bildern*

Clustering - Selbstorganisierende Karten

- gegeben: Tiere durch Eigenschaften beschrieben
- gesucht: Anordnung auf einer 2D-Karte, so dass ähnliche Tiere benachbart sind
- Diplomarbeit von **Benjamin Hoepner**: *Entwurf und Implementierung einer Applikation zur Visualisierung von Lernvorgängen bei Selbstorganisierenden Karten*

- Sombrero80x90, torus
- a=1; exp=1; r=50; exp=0.995; s=20000

Data Mining und Ethik

- *„Die Nutzung von Data Mining-Techniken bedeutet, dass die Nutzung der Daten weit über das hinausgehen kann, was bei der Aufnahme der Daten ursprünglich bekannt war"*

  [WF01 S.36]

- d. h. für personenbezogene Daten:

  …, was dem Erfassten bei der Aufnahme der Daten bewusst und von ihm geduldet war.

Gesetzgeber verlangt Einwilligung zum Datenschürfen

Interview von www.it-business.de mit Sabine Heukrodt-Bauer am 12.04.2007
Rechtsanwältin und Fachanwältin für IT-Recht

Personenbezogene Daten bspw. Kundendaten:

- Datenverarbeitung erlaubt durch Gesetz oder Einwilligung, sonst verboten
- Bspw. erlaubt im Online-Shop:
  - für die Abwicklung des Kaufes, oder wegen der steuerrechtlicher Aufbewahrungspflichten
  - Keine andere Verwendung ohne die Einwilligung des Betroffenen erlaubt. (Gebot der Zweckbindung)
- Einwilligung zur Werbung und Marktforschung **aktiv**: z. B. nicht vorgewählte Checkbox

  „Mit der Erhebung und Verwendung meiner Daten zu Werbezwecken bin ich einverstanden. Ich weiß, dass ich mein Einverständnis jederzeit widerrufen kann."
- -> dann Weitergabe von Beruf, Geburtsjahr, Grad, Anschrift … erlaubt

Gesetzgeber verlangt Einwilligung zum Datenschürfen

Interview von www.it-business.de mit Sabine Heukrodt-Bauer am 12.04.2007
Rechtsanwältin und Fachanwältin für IT-Recht

- **ITB: In der Praxis dürften die Betroffenen aber selten merken, was mit ihren Daten geschieht, und wie heißt es so schön: »Wo kein Kläger, da kein Richter«?**
- Heukrodt-Bauer:
  Das ist richtig. Aber in dem Maße, wie sich Data Mining durchsetzt, werden die Betroffenen wahrscheinlich die **Auswirkungen spüren, ihr Recht auf Auskunft** geltend machen und Aufklärung verlangen über die über sie gespeicherten Daten.

Daten
erheben, speichern, übermitteln und nutzen

- BDSG: ***Datensparsamkeit, Datenvermeidung***
- Verwendung personenbezogener Daten in D. restriktiv geregelt
  - Schutz der Daten, Recht auf Auskunft, Berichtigung, Löschung, Verwendungsbeschränkung …

Aber

- **Kontroll-Problem**: Datenerheber hat ein massives (existentielles?) Interesse an abgeleitetem (personenbezogenem) Wissen. Wer kontrolliert ihn?
- **Mehr und bessere Daten**: wachsende Speicher, Vernetzung von Datenbanken, mikrogeografische Daten, Kopierbarkeit, das Internet vergisst nichts – die Bahn* aber auch nicht
- Es gibt keine ‚belanglosen' Daten mehr

*- Amazon, Payback, DAK …

Mikrogeografische Daten

Kaufkraft nach Altersgruppen
40- bis unter 50-Jährige

		unter	72
72	bis unter	80	
80	bis unter	88	
88	bis unter	96	
96	bis unter	104	
104	bis unter	112	
112	bis unter	120	
120	bis unter	128	
128	und mehr		

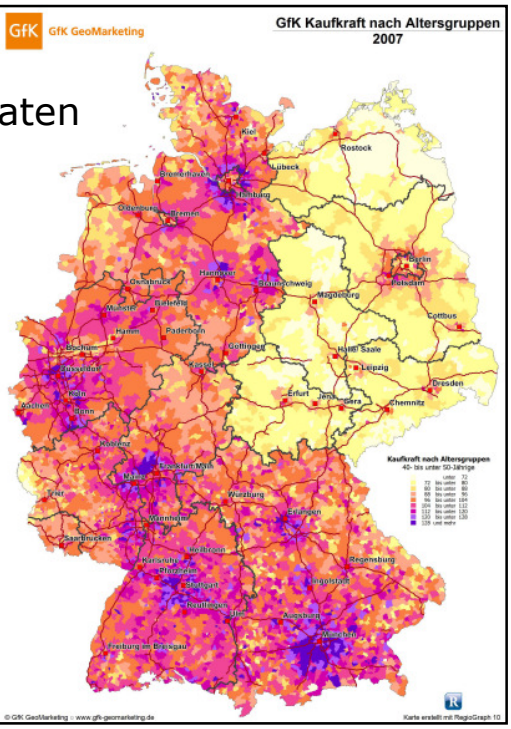


Foto: GfK GeoMarketing

Januar 2009
I. Boersch

Speichern und Übermitteln von Daten
Sexuelle Vorlieben im BDSG

- Anonyme Daten dürfen immer gespeichert werden
- Allgemein zugängliche pb-Daten im Prinzip auch
- Weitergabe von Listen erlaubt – demnächst nicht mehr

- **Besondere Arten personenbezogener Daten**: rassische und ethnische Herkunft, politische Meinungen, religiöse oder philosophische Überzeugungen, Gewerkschaftszugehörigkeit, Gesundheit oder Sexualleben: **§ 28**
- es sich um Daten handelt, die der Betroffene offenkundig öffentlich gemacht hat
  - Forschung, Medizin (inkl. Beauftragte)
  - Abwehr von erheblichen Gefahren, Verfolgung von Straftaten von erheblicher Bedeutung
  - Organisationen ohne Erwerbszweck: Mitglieder oder Personen, die im Zusammenhang mit deren Tätigkeitszweck regelmäßig Kontakte mit ihr unterhalten…….

Data Mining – Die Automatisierte Einzelentscheidung

(§ 6a BDSG, A15 Europäische Datenschutzrichtlinie)

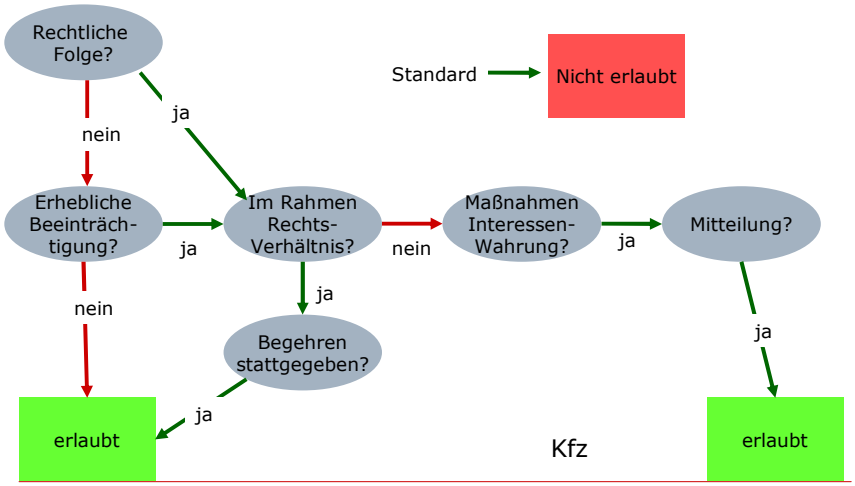
(1) Entscheidungen, die für den Betroffenen eine **rechtliche Folge** nach sich ziehen oder ihn **erheblich beeinträchtigen**, dürfen nicht **ausschließlich auf eine automatisierte Verarbeitung** personenbezogener Daten gestützt werden, die der Bewertung einzelner Persönlichkeitsmerkmale dienen.

(2) **Dies gilt nicht**, wenn

1. die Entscheidung im Rahmen des Abschlusses oder der Erfüllung eines **Vertragsverhältnisses** oder eines sonstigen **Rechtsverhältnisses** ergeht und dem Begehren des Betroffenen stattgegeben wurde oder

2. die **Wahrung der berechtigten Interessen** des Betroffenen durch geeignete Maßnahmen gewährleistet und dem Betroffenen von der verantwortlichen Stelle die Tatsache des Vorliegens einer Entscheidung im Sinne des Absatzes 1 **mitgeteilt** wird. Als geeignete Maßnahme gilt insbesondere die Möglichkeit des Betroffenen, seinen Standpunkt geltend zu machen. Die verantwortliche Stelle ist verpflichtet, ihre Entscheidung erneut zu prüfen.

Automatisierte Einzelentscheidung (ausschließliche) § 6a BDSG



Aspekte

- Wann wird Diskrimination zu Diskriminierung?
- Vorteils-Auswahl: Kündigerprävention, besten Studenten
- Nachteils-Auswahl: Zahlungsart, Kreditvergabe
- Attribute (z.B. Rasse, Religion) auch anonym problematisch
  - Kreditanträge: unethisch
  - in der Medizin: akzeptabel
- Triebkräfte: DM als Geschäftsvorteil, Sicherheitsbedürfnis, Neugier ..
- Globaler Trend
- Jeder will den gläsernen Kunden, keiner will es sein.

Informatiker am Hebel?

DM-Resultate kritisch betrachten

- Schätzung des Generalisierungsfehlers – schwierig!
- Prozes
- Daten
  Zusam
- Man fi

⇒ Herausfor
⇒ Anonymisi
⇒ Recht des
  Perspektivwechsel
⇒ Informieren über Potentiale und Grenzen des Data Mining

Prediction is very difficult, especially about the future.

Niels Bohr

⇒ Aktuelles?

Aktuell in Brandenburg

- **Wolfgang Neskovic**, ehemaliger Bundesverfassungsrichter, MdB

27. Januar, 19 Uhr, Bürgerhaus der Altstadt (Bäckerstraße)

Vortrag: Überwachungsstaat BRD - Vom Verlust der Freiheitsrechte

- **Securityforum 2009**

29. Januar, ganztägig, FH Brandenburg

Unternehmenssicherheit, Entwicklungstrends und Perspektiven in der Sicherheitsbranche

Ausgangsfrage

Ist es möglich, die angefallenen und weiter anfallenden riesigen Datenbestände in nützliche Informationen oder sogar Wissen umzuwandeln?

Ja.

(Vielleicht mehr, als Sie sich vorstellen können.)

Fliege in Borneo

- In Borneo gibt es ein Spiel, bei dem alle Mitspieler im Kreis sitzen und ihren Einsatz in die Mitte werfen. Jeder erhält vor sich eine Karte. Nach einer Weile entsteht ein Tumult, und ein Mitspieler gewinnt alles Geld, die Karten werden neu gemischt.

- Es war dem europäischen (sprachunkundigen) Gast auch nach langem Zusehen nicht möglich, die Regel zu erkennen, nach der die Gewinnerkarte bestimmt wurde. Keine Theorie hielt lange – rot vor schwarz, Bilder vor Zahlwerten ...

Ein Gastgeber klärte ihn auf – **es gewinnt die Karte, auf die sich als Erstes eine Fliege setzt.**

Roger Willemsen. ... und Du so? Soloprogramm, 2006

Literatur

Literaturempfehlung

- [WF01] Witten, Ian H. ; Frank, Eibe: **Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen**. Hanser Fachbuch,2001 (12 Exemplare in der Bibliothek)
- [HK01] Han, Jiawei ; Kamber, Micheline: **Data Mining. Concepts and Techniques**. Morgan Kaufmann Publishers, 2001 (1 Exemplar in der Bibliothek)
- http://www.datenschutzverein.de/ und
  ips-Kriterien, Teil Rechtliche Grundlagen unter http://www.datenschutz-cert.de/kriterien/Modul_MI_Rechtliche_Grundlagen.pdf

Algorithmen von Quinlan

- [Qui86] Quinlan, J.R.: **Induction of decision trees**. In Machine Learning 1986 Volume 1 Number 1. Springer Netherlands. Seiten 81-106: online unter http://www.springerlink.com/content/ku63wm5513224245/fulltext.pdf
- [Qui93] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

Quellennachweis

- [Ult04] Ultsch, Alfred: **Anwendungen Emergenter SOM**. 2004. –http://www.mathematik.uni-marburg.de/-databionics/de//?q=app Stand 18.05.2006
- [FPSS96] Fayyad, Usama ; Piatetsky-Shapiro, Gregory ; Smyth, Padhraic: **From Data Mining to Knowledge Discovery in Databases**. In: AI Magazine 17(3) (1996), S. 37–54. – http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf Stand 18.05.2006
- [AG05] prudsys AG: **Data Mining Cup**. Webseite. 2005. – http://www.data-mining-cup.de/ Stand 18.05.2006
- [TFLK98] Tsien, C., Fraser, H., Long, W. and Kennedy, R.: **Using classification tree and logistic regression methods to diagnose myocardial infarction**. Medinfo. v9. 493-497., 1998
  online unter: http://groups.csail.mit.edu/medg/people/hamish/medinfo-chris.pdf
- [SG07] Sieben, W., Gather, U.: **Classifying Alarms in Intensive Care - Analogy to Hypothesis Testing**, in Springer's Lecture Notes of Computer Science Series: Artificial Intelligence in Medicine. Proceedings of the 11th Conference on Artificial Intelligence in Medicine, Vol. 4594/2007, eds. R. Bellazzi, A. Abu-Hanna, J. Hunter, Berlin / Heidelberg: Springer, 130-138, 2007

Data Mining bekämpft Abrechnungsmanipulation effizient

Fahndung nach den Kassenplünderern

Krankenkassen finden betrügende Ärzte

Gerade in Zeiten steigender Beiträge fordern Versicherte von den Krankenkassen einen effizienten Einsatz der gezahlten Beitragsgelder. Ebenso erwartet die Öffentlichkeit den konsequenten Kampf gegen jeglichen Missbrauch. Stichproben wirken dabei kaum mehr als ein Tropfen auf dem heißen Stein. Wirklich systematische Methoden, um betrügerischen Ärzten und Kassenplünderern auf die Spur zur kommen, bietet dagegen Data Mining.

Autor: Dr. Jörg Reinnarth, Business Manager, Altran CIS, (j.reinnarth@altran-cis.de).

Bild: Altran CIS

Ein Schaden von rund einer Milliarde Euro entsteht jedes Jahr durch Abrechnungsbetrug und Manipulation im Gesundheitswesen. Diese drastische Zahl vertrat

ßigkeiten wittern. Typische Indizien stellen etwa überdurchschnittlich häufige Abrechnungen von teuren oder mit hohen Margen verbundenen Behandlungsmethoden dar. Hierzu zählen Magen-Darm-Spiegelungen, Ultraschallanalysen oder Röntgenaufnahmen, die in Wirklichkeit nie stattgefunden haben. In manchen Fällen arbeiteten Ärzte mit Pflegern zusammen, die ihnen heimlich Krankenkassenkarten aus einem Altenheim besorgten. Somit kann der Arzt Behandlungen für Heimbewohner abrechnen, obwohl sie nie bei ihm waren. Besonders delikat waren auch Fälle, wo bei älteren Damen die Behandlung von Erektionsstörungen abgerechnet

gruppen erlauben. Die immer noch erforderlichen manuellen Prüfungen würden auf diese Weise erheblich höhere Erfolgsquoten erzielen. Und gerade bei Krankenversicherungen findet Data Mining aufgrund der hohen Abrechnungsmengen gute Rahmenbedingungen vor. Denn das Verfahren entwickelt seine Stärke genau dann, wenn nicht detaillierte Einzelfälle zu untersuchen, sondern vielmehr ein Muster von auffälligen Daten zu identifizieren ist. Da die betreffenden Ärzte in der Regel nicht nur bei einem Patienten, sondern kontinuierlich betrügen, trifft genau dies zu.

Ein Rechenbeispiel zeigt die Effi-

Quelle: „Versicherungsbetriebe 4/2007“

Richtlinie 95/46/EG (Europäische Datenschutzrichtlinie)

Artikel 15

Automatisierte Einzelentscheidungen

(1) Die Mitgliedstaaten räumen jeder Person das Recht ein, keiner für sie rechtliche Folgen nach sich ziehenden und keiner sie erheblich beeinträchtigenden Entscheidung unterworfen zu werden, die ausschließlich aufgrund einer automatisierten Verarbeitung von Daten zum Zwecke der Bewertung einzelner Aspekte ihrer Person ergeht, wie beispielsweise ihrer beruflichen Leistungsfähigkeit, ihrer Kreditwürdigkeit, ihrer Zuverlässigkeit oder ihres Verhaltens.

(2) Die Mitgliedstaaten sehen unbeschadet der sonstigen Bestimmungen dieser Richtlinie vor, daß eine Person einer Entscheidung nach Absatz 1 unterworfen werden kann, sofern diese

a) im Rahmen des Abschlusses oder der Erfuellung eines Vertrags ergeht und dem Ersuchen der betroffenen Person auf Abschluß oder Erfuellung des Vertrags stattgegeben wurde oder die Wahrung ihrer berechtigten Interessen durch geeignete Maßnahmen - beispielsweise die Möglichkeit, ihren Standpunkt geltend zu machen - garantiert wird oder

b) durch ein Gesetz zugelassen ist, das Garantien zur Wahrung der berechtigten Interessen der betroffenen Person festlegt.

Gesetze zum Datenschutz

- Richtlinie 95/46/EG des Europäischen Parlaments
- Bundesdatenschutzgesetz
- Datenschutzgesetze der Länder (vor BDSG)
- Datenschutznormen des Sozialgesetzbuchs
- Informationsfreiheitsgesetz
- Telemediengesetz (TMG)
- Telekommunikationsgesetz (TKG)
- BKA-Gesetz
- Gesetz zur Änderung seeverkehrsrechtlicher, verkehrsrechtlicher und anderer Vorschriften mit Bezug zum Seerecht
- Gesetz zu dem Abkommen vom 26. Juli 2007 zwischen der Europäischen Union und den Vereinigten Staaten von Amerika über die Verarbeitung von Fluggastdatensätzen (Passenger Name Records – PNR) und deren Übermittlung durch die Fluggesellschaften an das United States Department of Homeland Security (DHS) (PNR-Abkommen 2007)
- .....

[Benjamin Franklin 1759]

- They who can give up essential liberty to obtain a little temporary safety, deserve neither liberty nor safety
- Wer grundlegende Freiheiten für ein wenig vorübergehende Sicherheit aufgeben kann, verdient weder Freiheit noch Sicherheit