Data Mining in der Wirtschaft und durch Geheimdienste

Methoden und Anwendungen

Balázs Bárány

Rapid-I GmbH

Club IT, 11. 9. 2013

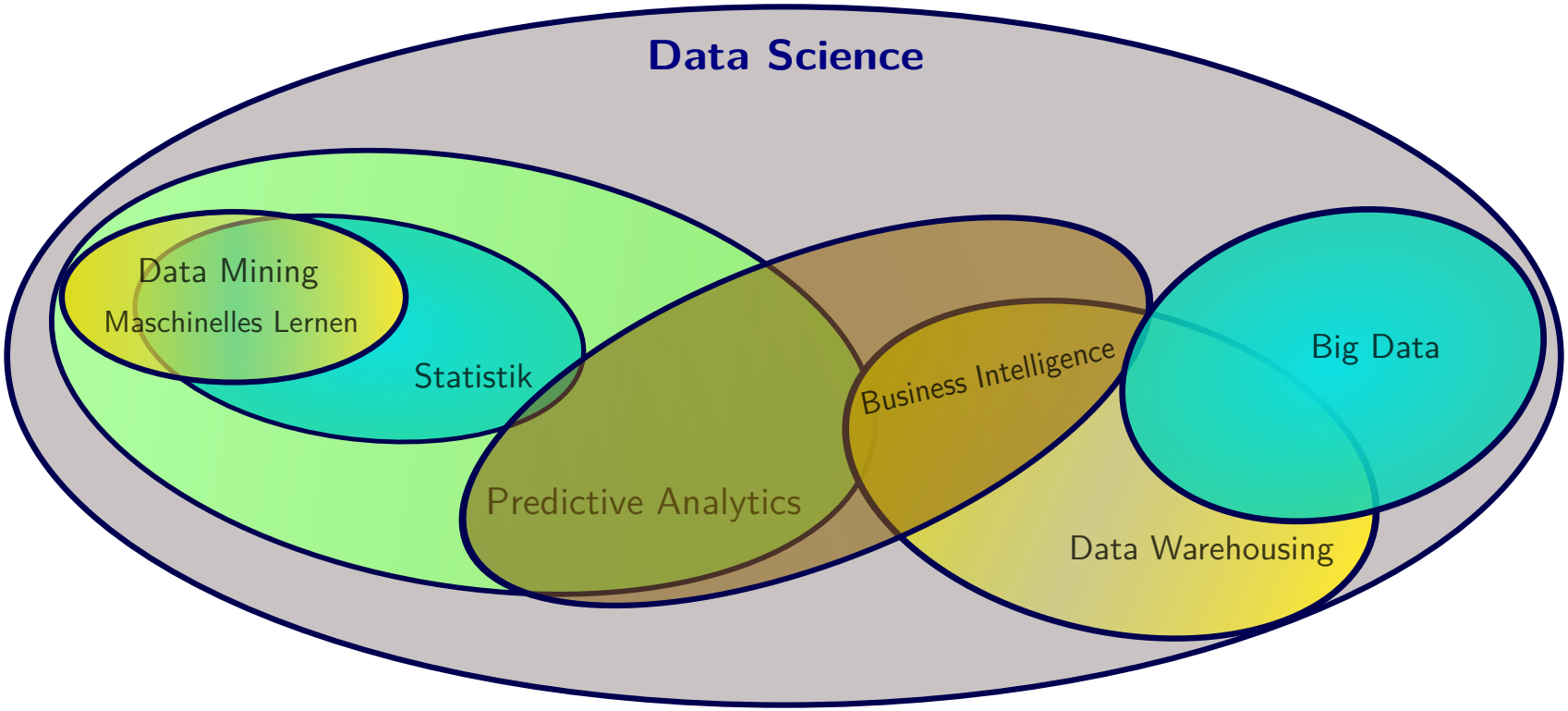
Inhalt

Einführung und Überblick
- Data Mining im Kontext

Methoden und Technik
- Einteilung der Verfahren
- Beispielhafte Lernalgorithmen
- Erweiterte Anwendungen
- Umsetzung in die Praxis
- Werkzeuge

Anwendungen in der Wirtschaft (und durch die Geheimdienste?)

Wichtige Begriffe



Ethische Fragen

- Werkzeug-Aspekt

Ethische Fragen

- Werkzeug-Aspekt
- Datenschutz

Ethische Fragen

- Werkzeug-Aspekt
- Datenschutz
- Aussagekraft; Umgang mit falschen Vorhersagen

Methoden und Technik

Methoden: Einteilung

Einteilung der Verfahren

- informierte Verfahren („supervised”): Zielvariable bekannt

Einteilung der Verfahren

- informierte Verfahren („supervised”): Zielvariable bekannt
- Aufgaben:
  - Klassifikation
  - Regression
  - Assoziation

Einteilung der Verfahren

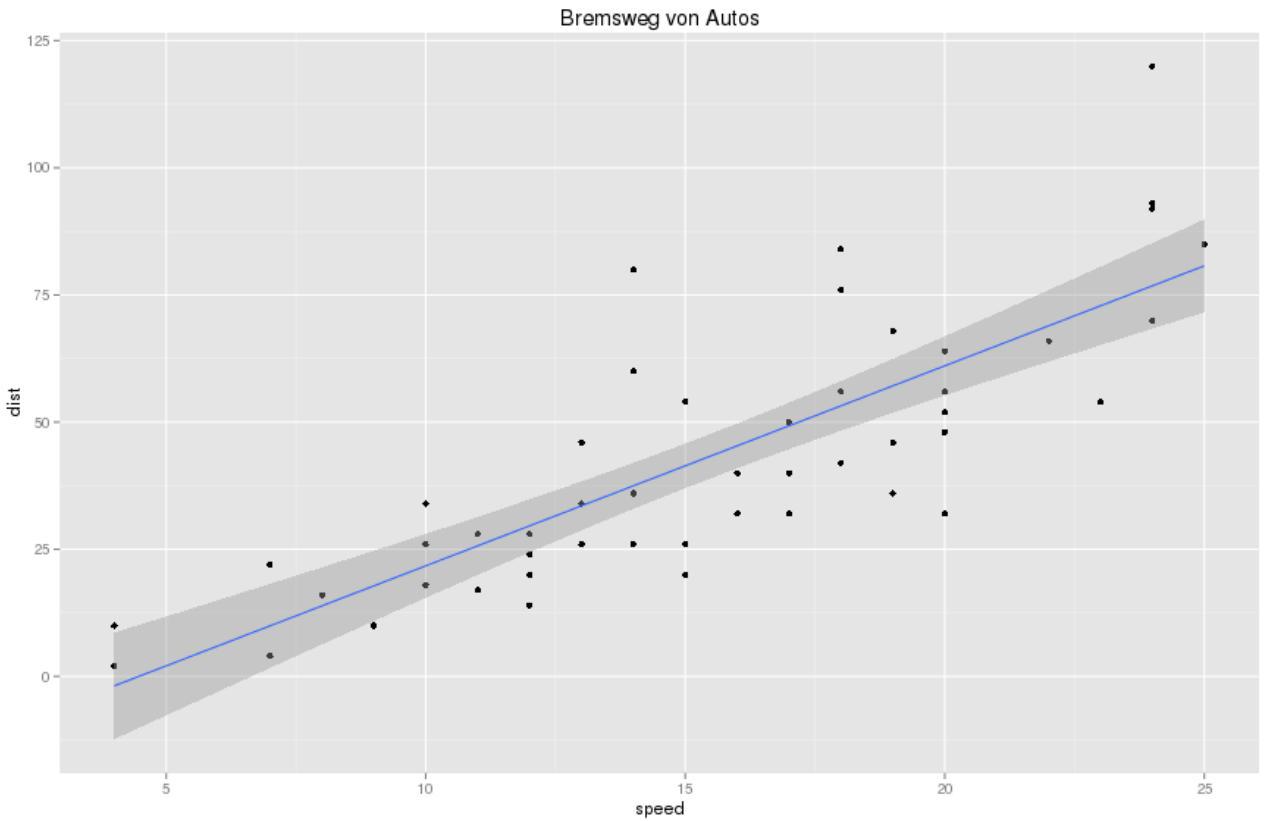
- informierte Verfahren („supervised"): Zielvariable bekannt
- Aufgaben:
  - Klassifikation
  - Regression
  - Assoziation
- uninformierte Verfahren („unsupervised"): Zielvariable nicht vorhanden oder nicht bekannt
  - Clustering
  - Ausreißererkennung

Methoden und Technik

Lernalgorithmen

(Lineare) Regression

- Erstellt Gleichungen der Form $y = ax + b$

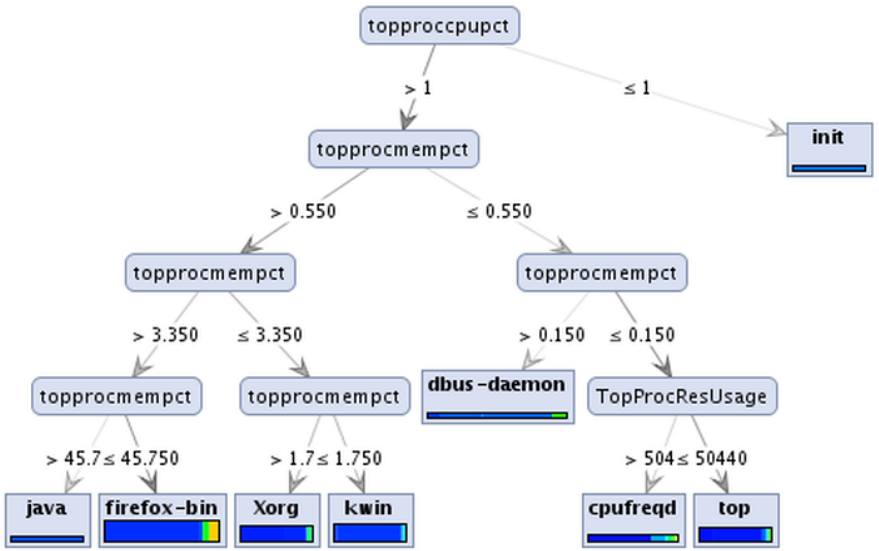


Quelle: R-Projekt, Datensatz „cars"

Regression: Erweiterungen

- mit mehreren Variablen
- nichtlineare Funktionen
- andere Verteilungen

Entscheidungsbäume



Entscheidungsbaum (Quelle: RapidMiner)

Entscheidungsbäume: Algorithmus

- Attribut und Wert suchen, der die Daten am besten aufteilt
  - Rekursiv auf die Teilmengen anwenden, ...
  - ... bis irgendeine Abbruchbedingung erfüllt ist: zu kleine Gruppen, Baum zu „tief" usw.

Entscheidungsbäume: Algorithmus

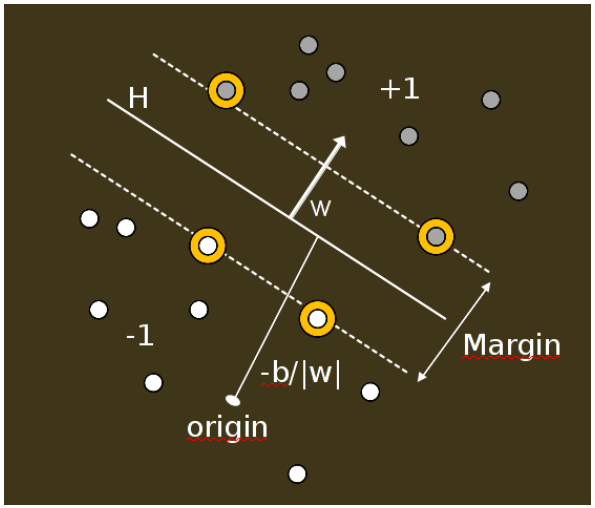
- Attribut und Wert suchen, der die Daten am besten aufteilt
  - Rekursiv auf die Teilmengen anwenden, ...
  - ... bis irgendeine Abbruchbedingung erfüllt ist: zu kleine Gruppen, Baum zu „tief" usw.
  - Mögliche Kriterien für Aufteilung: Informationsgewinn (Gain); Gain ratio; Gini-Koeffizient

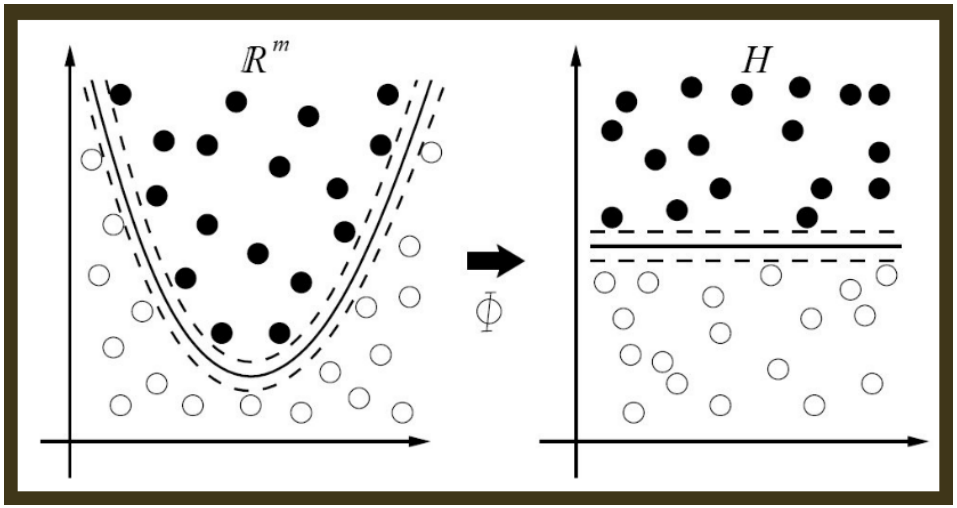
Entscheidungsbäume: Algorithmus

- Attribut und Wert suchen, der die Daten am besten aufteilt
  - Rekursiv auf die Teilmengen anwenden, ...
  - ... bis irgendeine Abbruchbedingung erfüllt ist: zu kleine Gruppen, Baum zu „tief" usw.
  - Mögliche Kriterien für Aufteilung: Informationsgewinn (Gain); Gain ratio; Gini-Koeffizient
- Viele Varianten und Erweiterungen

Support Vector Machine

- Versucht, die größtmögliche Distanz zwischen Kategorien zu finden
- Transformation der Attribute, um „lineare Trennung" zu erzielen

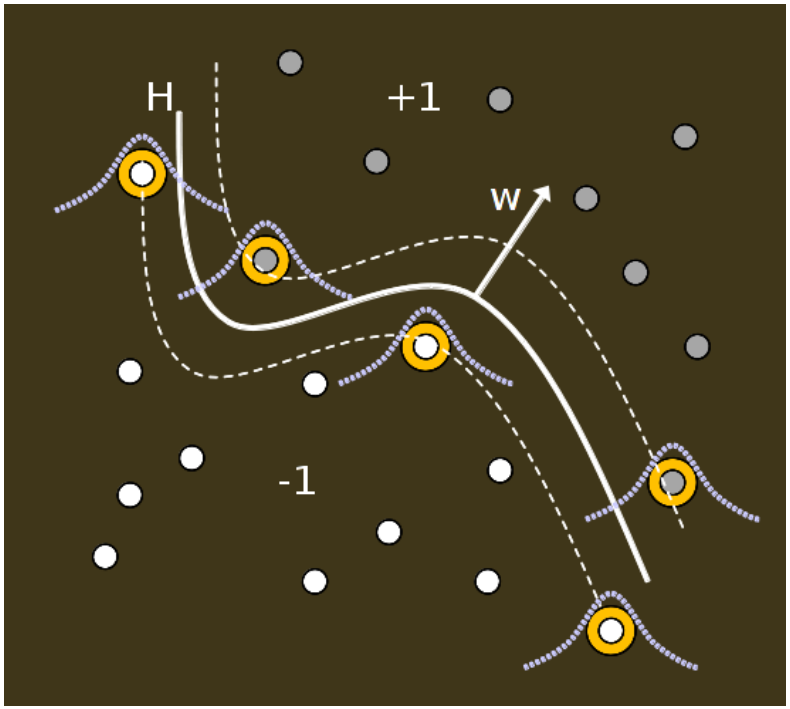




Quelle: RapidMiner-Schulungsunterlagen

Support Vector Machine

- Nur einige Punkte („support vector") definieren die Klassengrenzen



Quelle: RapidMiner-Schulungsunterlagen

Weitere Algorithmen

- Naive Bayes
  - „bedingte Wahrscheinlichkeiten”

Weitere Algorithmen

- Naive Bayes
  - „bedingte Wahrscheinlichkeiten"
- Neuronale Netze
  - simples Modell menschlicher Nervenzellen

Weitere Algorithmen

- Naive Bayes
  - „bedingte Wahrscheinlichkeiten"
- Neuronale Netze
  - simples Modell menschlicher Nervenzellen
- Nearest Neighbors
  - Suche nach ähnlichen Attributwerten

Weitere Algorithmen

- Naive Bayes
  - „bedingte Wahrscheinlichkeiten"
- Neuronale Netze
  - simples Modell menschlicher Nervenzellen
- Nearest Neighbors
  - Suche nach ähnlichen Attributwerten
- Ensemble-Algorithmen u. v. m.

Methoden und Technik

Erweiterungen für unstrukturierte Daten

Text mining

- Tabellenstruktur aus unstrukturierten Texten/Dokumenten

ExampleSet (270 examples, 2 special attributes, 203 regular attributes) View Filter (270 / 270): all

Row...	Wertung	ID	Datum	a	abgehol	abhol	abholberei	abholung	abwicklung	aktuell	alt	angebo	anruf	antwor	artikel
60	+	160	2005.01.20.	0	0	0	0	0	0	0	0	0.210	0	0	0
61	-	167	2005.05.10.	0	0	0	0	0	0	0	0	0	0	0	0.485
62	-	180	2005.05.26.	0	0	0	0	0	0	0	0	0	0	0	0
63	+	190	2005.06.10.	0	0	0	0	0	0	0	0	0.434	0	0	0
64	+	191	2005.06.15.	0	0	0	0	0	0	0	0	0	0	0	0
65	+	238	2006.02.01.	0	0	0	0	0	0	0	0	0	0	0	0
66	+	239	2006.02.01.	0	0	0	0	0	0	0.550	0	0	0	0	0
67	+	240	2006.02.01.	0	0	0	0	0	0	0	0	0	0	0	0
68	+	241	2006.02.01.	0	0	0	0	0	0	0	0	0	0	0	0
69	-	248	2006.02.01.	0	0	0	0	0	0	0	0	0	0.684	0	0
70	+	252	2006.02.01.	0	0	0	0	0	0	0	0	0	0	0	0
71	+	253	2006.02.01.	0	0.297	0	0	0	0	0	0	0	0	0	0
72	-	254	2006.02.01.	0	0	0	0	0	0	0	0	0	0	0	0
73	+	278	2006.02.02.	0	0.365	0	0	0	0	0	0	0	0	0	0
74	+	279	2006.02.02.	0.829	0	0.137	0	0	0	0	0	0	0	0	0
75	+	280	2006.02.02.	0	0	0	0	0	0	0	0	0	0	0	0
76	+	281	2006.02.02.	0	0	0	0	0.481	0	0	0	0	0	0	0
77	+	282	2006.02.03.	0	0	0	0	0	0	0	0	0	0	0	0

Quelle: RapidMiner, Händlerbewertungen auf geizhals.at

Vorgehensweise bei Text Mining

- Berechnung: Term-Frequenz, TF/IDF

Beispiel

$$TF/IDF = \frac{term\ frequency}{document\ frequency}$$

Je häufiger ein Term in einem Dokument vorkommt, aber je seltener in anderen Dokumenten, umso relevanter ist er für dieses Dokument.

Image mining

- Segmentierung und Berechnung von Kennzahlen
  - Farbverteilung, Komplexität, Helligkeit, ...

Image mining

- Segmentierung und Berechnung von Kennzahlen
  - Farbverteilung, Komplexität, Helligkeit, ...
- Erkennung von Formen und Figuren (Anzahl, Position)

Image mining

- Segmentierung und Berechnung von Kennzahlen
  - Farbverteilung, Komplexität, Helligkeit, ...
- Erkennung von Formen und Figuren (Anzahl, Position)
- Ähnlichkeit ganzer Bilder oder Bildelemente (Personen, ...)

Image mining

- Segmentierung und Berechnung von Kennzahlen
  - Farbverteilung, Komplexität, Helligkeit, ...
- Erkennung von Formen und Figuren (Anzahl, Position)
- Ähnlichkeit ganzer Bilder oder Bildelemente (Personen, ...)
- Erweiterung: Videoanalyse

Netzwerkanalyse

- Netzwerkstruktur: Graph im mathematischen Sinne
  - Besteht aus Knoten und Kanten

Netzwerkanalyse

- Netzwerkstruktur: Graph im mathematischen Sinne
  - Besteht aus Knoten und Kanten
- Kennzahlen: Zentralität, Intensität und Richtung der Beziehungen
  - Entfernung vom „Zentrum" oder definierten Knoten

Methoden und Technik

Umsetzung in die Praxis

Algorithmen sind nicht alles

- Vorverarbeitung fast immer notwendig

Algorithmen sind nicht alles

- Vorverarbeitung fast immer notwendig
- Behandlung fehlender Werte

Algorithmen sind nicht alles

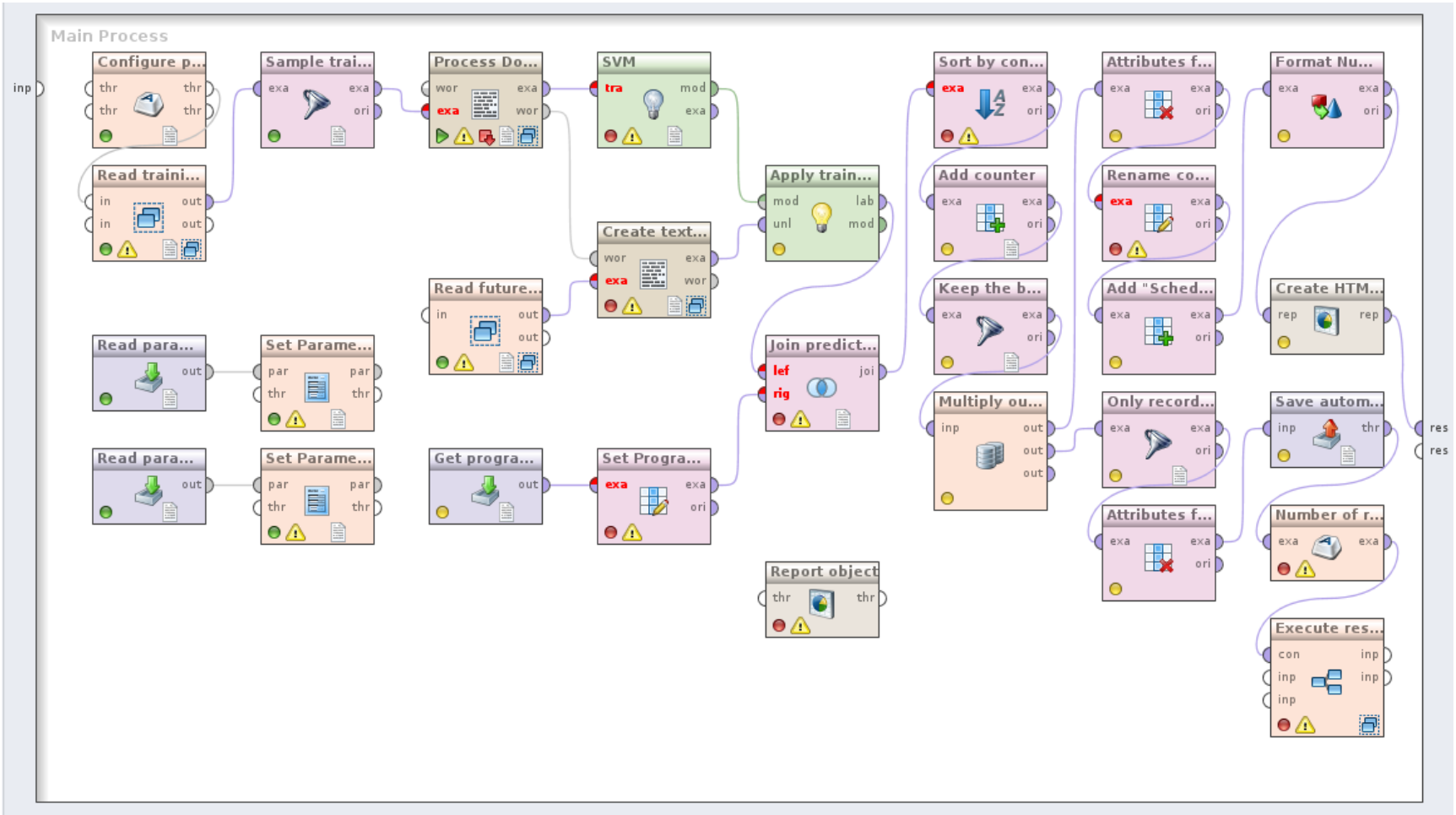
- Vorverarbeitung fast immer notwendig
- Behandlung fehlender Werte
- Attributkonstruktion

Algorithmen sind nicht alles

- Vorverarbeitung fast immer notwendig
- Behandlung fehlender Werte
- Attributkonstruktion
- Parameteroptimierung

Umsetzung

- Lernalgorithmus nur kleiner Teil der Gesamtlösung!



Quelle: RapidMiner; MythMiner-Projekt des Vortragenden

Methoden und Technik

Data-Mining-Werkzeuge

Werkzeuge

- Komplettlösungen für Data Mining und Analytik, GUI
  - Open source: RapidMiner, KNIME, Weka, Orange usw.

Werkzeuge

- Komplettlösungen für Data Mining und Analytik, GUI
  - Open source: RapidMiner, KNIME, Weka, Orange usw.
  - Kommerziell: SAS, IBM SPSS, Statistica usw.

Werkzeuge

- Komplettlösungen für Data Mining und Analytik, GUI
  - Open source: RapidMiner, KNIME, Weka, Orange usw.
  - Kommerziell: SAS, IBM SPSS, Statistica usw.
- Statistische Programmiersprache R (open source)
  - GUI für Data Mining: Rattle

Anwendungen

Anwendungen in Wirtschaft, Wissenschaft und durch die Geheimdienste

Kündigerprävention

- In der Wirtschaft: Daten wie Vertragsdauer, Nutzungsintensität, Konkurrenzsituation usw.
  - Daraus Erkennung kündigungswilliger Kunden, Reaktion

Kündigerprävention

- In der Wirtschaft: Daten wie Vertragsdauer, Nutzungsintensität, Konkurrenzsituation usw.
  - Daraus Erkennung kündigungswilliger Kunden, Reaktion
- Durch Geheimdienste:
  - Änderungen in der Lebenssituation, Kommunikationsverhalten
  - Vorhersage: Gefahr der „Radikalisierung"
  - oder: bekannte „Zelle" wird jetzt aktiv
  - oder: welche Mitglieder sind unglücklich und können „umgedreht" werden

Betrugserkennung

- In der Wirtschaft (z. B. Kreditkartenfirma): Daten wie Kaufhäufigkeit, Ort der Transaktion, Länge der Geschäftsbeziehung, Art der bezahlten Waren/Dienstleistungen usw.
  - Daraus Erkennung „verdächtiger" Transaktionen und eventuell Zurückweisung

Betrugserkennung

- In der Wirtschaft (z. B. Kreditkartenfirma): Daten wie Kaufhäufigkeit, Ort der Transaktion, Länge der Geschäftsbeziehung, Art der bezahlten Waren/Dienstleistungen usw.
  - Daraus Erkennung „verdächtiger" Transaktionen und eventuell Zurückweisung
  - Bei falscher Anwendung zu hohe Rate „Falscher Positive"

Betrugserkennung

- In der Wirtschaft (z. B. Kreditkartenfirma): Daten wie Kaufhäufigkeit, Ort der Transaktion, Länge der Geschäftsbeziehung, Art der bezahlten Waren/Dienstleistungen usw.
  - Daraus Erkennung „verdächtiger" Transaktionen und eventuell Zurückweisung
  - Bei falscher Anwendung zu hohe Rate „Falscher Positive"
- Durch Geheimdienste:
  - Auffällige, „verdächtige" Kommunikation oder Lebensführung oder Änderungen dieser Aspekte
  - Auffällige, unübliche Geldflüsse (SWIFT-Datentransfer in die USA!)
  - Identifiziert Verdächtige für genauere Überprüfung oder für schwarze Listen („no-fly list")

Regressionsmethoden

- In der Wirtschaft:
  - Absatzplanung, Bedarfsplanung

Regressionsmethoden

- In der Wirtschaft:
  - Absatzplanung, Bedarfsplanung
  - Optimierung von Abläufen und technischen Prozessen
    - Maschinenwartung, Fertigungsprozesse, Unternehmensprozesse

Regressionsmethoden

- In der Wirtschaft:
  - Absatzplanung, Bedarfsplanung
  - Optimierung von Abläufen und technischen Prozessen
    - Maschinenwartung, Fertigungsprozesse, Unternehmensprozesse
- Durch Geheimdienste:
  - Analyse wirtschaftlicher Kennzahlen, daraus Vorhersagen über Handlungen von Staaten oder Gruppen
  - Bedarfsplanung (z. B. benötigte Rechenzentrumskapazität)

Text mining, Textanalyse

- In der Wirtschaft:
  - Spamfilter, Vorsortierung des Maileingangs
  - Stimmung, Einstellung zur Marke oder Firma in Online-Bewertungen, Foren- und Facebook-Postings, Tweets
  - Dokumentenkategorisierung

Text mining, Textanalyse

- In der Wirtschaft:
  - Spamfilter, Vorsortierung des Maileingangs
  - Stimmung, Einstellung zur Marke oder Firma in Online-Bewertungen, Foren- und Facebook-Postings, Tweets
  - Dokumentenkategorisierung
- Durch Geheimdienste:
  - Erkennung von diskutierten Themen (Anschlagspläne, radikale Einstellungen, ...)
    - Fall Andrej Holm

Text mining, Textanalyse

- In der Wirtschaft:
  - Spamfilter, Vorsortierung des Maileingangs
  - Stimmung, Einstellung zur Marke oder Firma in Online-Bewertungen, Foren- und Facebook-Postings, Tweets
  - Dokumentenkategorisierung
- Durch Geheimdienste:
  - Erkennung von diskutierten Themen (Anschlagspläne, radikale Einstellungen, ...)
    - Fall Andrej Holm
  - Erkennung ungewöhnlicher Kommunikation (Thema paßt nicht zum sozialen Umfeld: kodierte Sprache?)

Image Mining

- In der Wirtschaft:
  - (Facebook, Google): Erkennung von Menschen auf Bildern
  - Landwirtschaft, Bergbau: Verarbeitung von Satellitenbildern

Image Mining

- In der Wirtschaft:
  - (Facebook, Google): Erkennung von Menschen auf Bildern
  - Landwirtschaft, Bergbau: Verarbeitung von Satellitenbildern
- Durch Geheimdienste:
  - Auswertung von Bildern aus Überwachungskameras, -satelliten
  - Erkennung von Verdächtigen (z. B. am Flughafen)

Netzwerkanalyse

- In der Wirtschaft:
  - Kündigerprävention
  - Empfehlung neuer Waren oder Dienstleistungen

Netzwerkanalyse

- In der Wirtschaft:
  - Kündigerprävention
  - Empfehlung neuer Waren oder Dienstleistungen
- Durch Geheimdienste:
  - Identifizierung zentraler oder „interessanter" Figuren
  - Plötzliche Änderungen im Kommunikationsverhalten
  - Ausweitung der Suche auf Kommunikationsumfeld des Verdächtigen

Ausblick, abschließende Gedanken

- Data Mining wird noch zunehmen

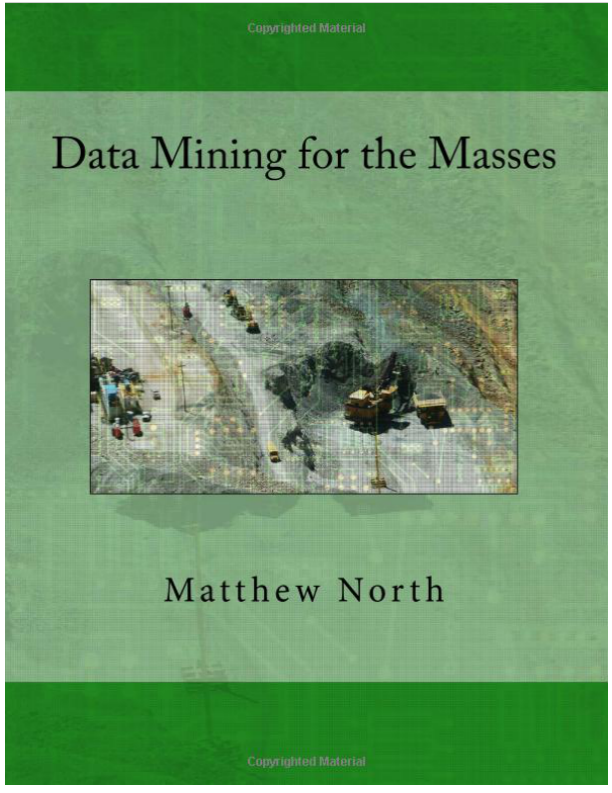
Ausblick, abschließende Gedanken

- Data Mining wird noch zunehmen
- Sinnvoll, sich damit zu beschäftigen

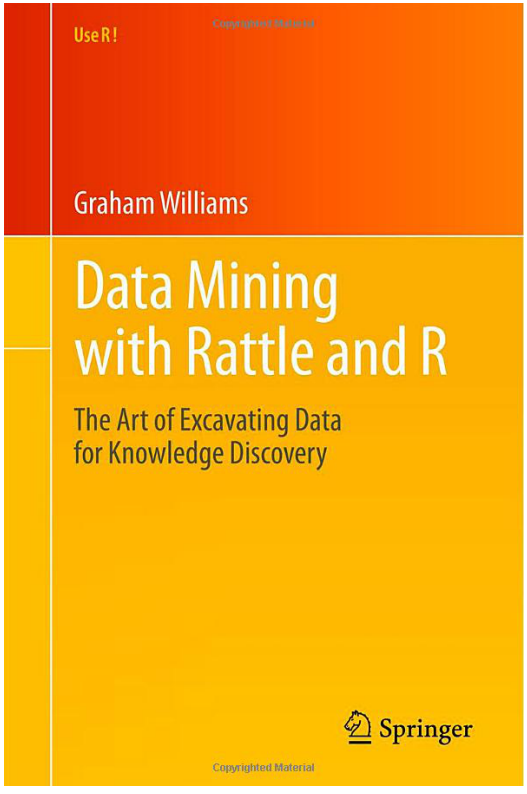
Ausblick, abschließende Gedanken

- Data Mining wird noch zunehmen
- Sinnvoll, sich damit zu beschäftigen
- Data Mining nicht per se schlimm

Literatur – anwendungsorientiert

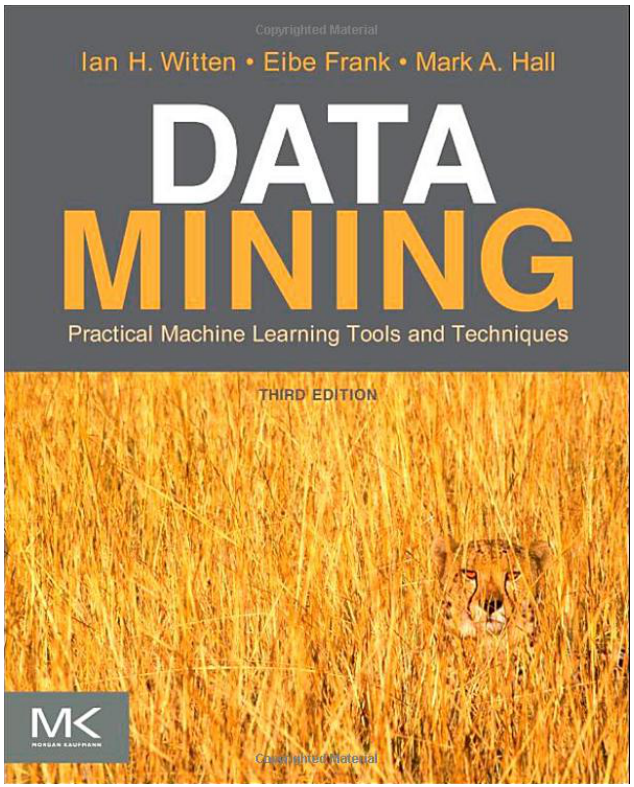


Matthew North:
Data Mining for the Masses



Graham Williams:
Data Mining with Rattle and R

Literatur – theoretische Grundlagen



Ian H. Witten, Eibe Frank, Mark A. Hall:
Data Mining (3. ed.)

Schluß

- Fragen?

Balázs Bárány <balazs@tud.at>