Applied Analytical Data Science

Teil 4: Preprocressing II

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,

Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

http://www.kietz.ch/AADS/

Heutiges Programm

- Ziele der Datenaufbereitung
- Daten Transformation
  - Erstellung abgeleiteter Attribute
  - Erstellung abgeleiteter Datensätze
- Daten Integration
- Daten Reduktion
  - Attribut Reduktion/Selektion
  - Sampling
- Ein Beispiel

Ziel der Datenaufbereitung

- Bereitstellung aller **relevanten** Daten in der **nützlichsten** Form für das DM-tool
  - ✓ Selektion der relevanten Daten, ausreichender Qualität
  - ✓ Erhöhung der Datenqualität
  - – Erstellung abgeleiteter Attribute die nützlicher sind
  - – Erstellung abgeleiteter Datensätze die nützlich/notwendig sind
  - – Erfüllen aller Formatbeschränkungen des DM-Tools
- Jegliche Manipulation der (Lern) Daten muss automatisiert sein, damit sie auch auf Test- und Anwendungsdaten angewendet werden kann.

Heutiges Programm

- Ziele der Datenaufbereitung
- Daten Transformation
  - Erstellung abgeleiteter Attribute
  - Erstellung abgeleiteter Datensätze
- Daten Integration
- Daten Reduktion
  - Attribut Reduktion/Selektion
  - Sampling
- Ein Beispiel

Attributkonstruktion

- füge neukonstruierte Attribute mit abgeleiteten Werten hinzu
- verbessert Genauigkeit und Verständnis bei hochdimensionalen Daten, z.B. Fläche = Breite * Höhe
- Operatoren z.B.: XOR (binäre Attribute)
- Recodierungen numerisch ↔ nominal
- Vorteile z.B.:
  - direkter Einbezug der zusammengesetzten Information
  - Reduktion von Mehrfachtests (z.B. Fragmentierung von Entscheidungsbäumen)

⇒ Die erwarteten Zusammenhänge in den Hypothesenraum des Data Mining Tools bringen.

Codierung nominaler Attribute

- viele Data Mining Verfahren können nur mit numerischen Werten umgehen
- nominale Wertebereiche enthalten implizite Beziehungen zwischen einzelnen Werten
- beliebige Ersetzung nominaler durch numerische Werte kann zu neuen (unerwünschten) Mustern führen
- Ersetzung kann die Beziehung verschiedener Attribute untereinander verfälschen
- die Ersetzung sollte idealerweise für viele Verfahren geeignet sein (z.B. für distanzbasierte und funktionsbasierte Verfahren)

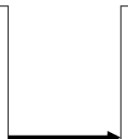
Beispiel: Codierungen

Beispiel A: Zivilstand (akzeptabel)

unverheiratet:	0
alleinstehend:	0.1
geschieden:	0.15
verwitwet:	0.65
verheiratet:	1

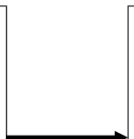
Beispiel B: Salär (schlecht)

1/2 Tag:	100
Tag:	200
1/2 Woche:	500
Woche:	1000
1/2 Monat:	2000
Monat:	4000



Codierung

Tag	1
1/2 Tag:	2
1/2 Monat:	3
1/2 Woche	4
Monat:	5
Woche:	6



1	200
2	100
3	2000
4	500
5	4000
6	1000

Techniken zur numerischen Codierung

- *One-of-n Remapping*
  kreiere binäre Pseudo-Variable für jeden Attributwert
  Beispiel: Kantone
  Vorteile: Wahrscheinlichekeit vs. Durchschnitt, Prediction
  Nachteile: Dimensionalität, Dichte
- *M-of-n Remapping*
  ersetze Attribut durch mehrere abgeleitete Variablen, deren Wertkombinationen die ursprünglichen Werte charakterisieren
  Beispiel: Lage, Bevölkerungsdichte und/oder Koordinaten von Kantonen
- *Kalibrierung* gegen (andere) numerische Attribute
- *Joint Distribution Tables* für alphanumerische Attribute

Diskretisierung skalarer Attribute

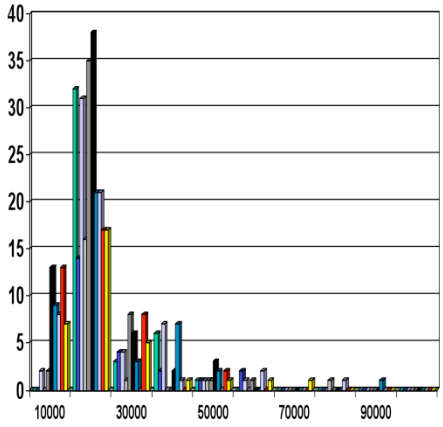
- Diskretisierung
  - unterteile den Wertebereich einer skalaren Variable in Intervalle, verwende Grenzwerte, Mittelwerte usw.
  - wird von bestimmten Algorithmen benötigt
  - ergibt auch Datenreduktion
  - dient als Vorbereitung für spätere Analysen (weitere Abstraktionen)
- Erzeugung von Abstraktionshierarchien
  - bildet Abstraktionsstufen auf elementaren Daten
  - niedrigere Granularität ist häufig aussagekräftiger im Hinblick auf das Erkennen von Gemeinsamkeiten und allgemeinen Regelhaftigkeiten
  - entsprechen dem natürlichen Verständnis/Denken

Diskretisierungmethoden

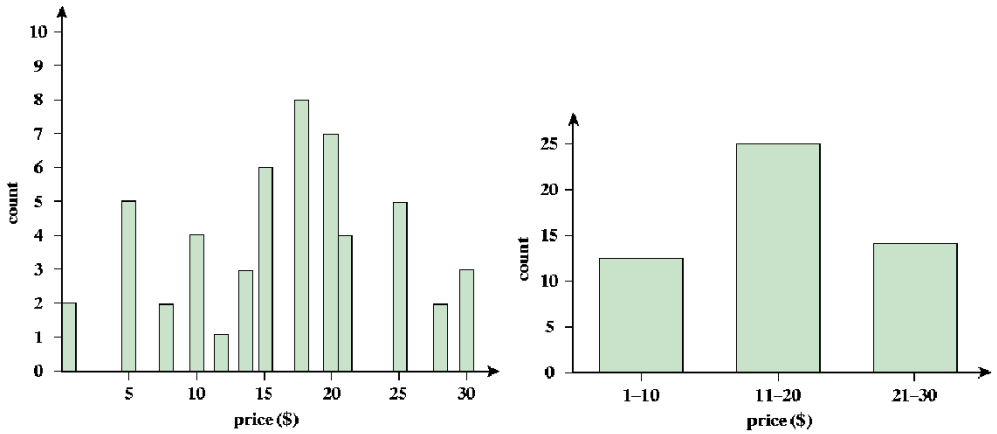
- Binning (siehe Teil 4)
- Histogramm-Analyse
- Clustering
- Entropie-basierte Diskretisierung
- Segmentbildung durch natürliche Partitionierung

Histogramme

- Populäre Datenreduktionstechnik: Approximiere Datenverteilungen
- Unterteile Daten in buckets und speichere (durchschnittliche) Häufigkeit
- Varianten
  - Equiwidth
  - Equidepth
  - V-Optimal
  - MaxDiff
  - mehrdimensional



Beispiel: Histogramme



Entropie-basierte Diskretisierung

- Informationsgewinn als allgemeines Kriterium der Attributauswahl:
  - zur grundlegenden Feature Selection im Rahmen des Preprocessing
  - zur Auswahl der Split-Kriterien bei der Induktion von Decision Trees
  - und auch zur Diskretisierung
- Voraussetzung: Information über Klassenzugehörigkeiten der Trainings-daten bzw. des Samples
- Anwendung auch bei der Diskretisierung:
  - wähle (binäre) Unterteilung eines Attributwertebereichs, die den höchsten Informationsgewinn (bzgl. einer vorgegebenen Klassifizierung) liefert
  - wende Verfahren rekursiv auf die resultierenden Partitionen an

Beispiel: Entropie-basierte Diskretisierung

- Annahme: Werte für Attribut *age* explizit als konkrete Werte gegeben
- Wähle bestimmten Schwellwert (z.B. 30) zur Partitionierung in zwei Teilmengen
- Bestimme Gain wie zuvor (Siehe Teil 6) unter Berücksichtigung der Klassenzugehörigkeiten, d.h. mit

$$E(age') = \frac{5}{14} I(2,3) + \frac{9}{14} I(7,2)$$

- Annahme: Schwellwert (30) ergibt den grössten Gain
- Dann rekursive Anwendung auf Intervall > 30 z.B. mit neuem Schwellwert 40 usw.

Heutiges Programm

- Ziele der Datenaufbereitung
- Daten Transformation
  - Erstellung abgeleiteter Attribute
  - Erstellung abgeleiteter Datensätze
- Daten Integration
- Daten Reduktion
  - Attribut Reduktion/Selektion
  - Sampling
- Ein Beispiel

Erzeugen von Datensätzen

- Erzeugen von negativen Beispielen
  Eine Kunden-DB enthält typischerweise nur wer (etwas) gekauft hat, nicht jedoch wer nichts (etwas nicht) gekauft hat.
- Balancieren des Datensatzes bei ungleich verteilter Zielvariable
  Z.B. Mailing-Responses typischerweise 95 nein : 5 ja
  duplizieren
- Änderung der Individuen, i.a. Aggregation von Datensätzen, z.B.:
  - Von der Personen zur Haushaltssicht
  - Von Tages- zu Wochen-/Monats-/Jahresumsätzen

Heutiges Programm

- Ziele der Datenaufbereitung
- Daten Transformation
  - Erstellung abgeleiteter Attribute
  - Erstellung abgeleiteter Datensätze
- Daten Integration
- Daten Reduktion
  - Attribut Reduktion/Selektion
  - Sampling
- Ein Beispiel

Daten Integration

Data integration: kombiniert/vereinigt Daten aus mehreren Quellen in eine Tabelle

- Schema integration
  - integriert Metadaten verschiedener Quellen
  - Entity identification problem: identifiziere Entitäten der realen Welt aus mehreren Datenquellen, z.B. Kundennummer: A.kd-id ≡ B.kd-#
- Entdeckung und Auflösung von Konflikten
  - für die gleiche Entität ergeben sich aus verschiedenen Quellen unterschiedliche Attributwerte
  - mögliche Gründe: unterschiedliche Repräsentation, unterschiedliche Einkeiten/Skalen
- Entdeckung der Zuordnungsvorschrift (z.B.: Namensmatching)
- Kombinieren von N:M verknüpften Tabellen

Redundante Daten

- RedundanteDaten entstehen häufig bei der Integration mehrerer Datenquellen:
  - Namensunterschiede des gleichen Attributes
  - Originäre vs. abgeleitete Attribute (Bsp. Umsatz)
- Hilfsmittel zum Aufdecken redundanter Attribute:
  Korrelationsanalyse
- Sorgfalt bei der Integration von Daten aus verschiedenen Quellen hilft, Redundanzen und Inkonsistenzen zu reduzieren/vermeiden und damit Mining-Aufwand und Qualität zu verbessern

Heutiges Programm

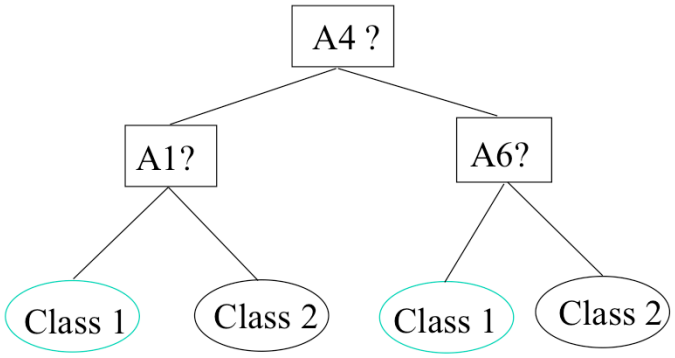
- Ziele der Datenaufbereitung
- Daten Transformation
  - Erstellung abgeleiteter Attribute
  - Erstellung abgeleiteter Datensätze
- Daten Integration
- Daten Reduktion
  - Attribut Reduktion/Selektion
  - Sampling
- Ein Beispiel

Attribut Selektion (Feature Selection)

- Attribut Selektion (d.h. Auswahl einer Teilmenge von Attributen):
  - Wähle minimale Menge von Features, so dass die Wahrscheinlichkeitsverteilung verschiedener Klassen mit den Werten dieser Features so nah wie möglich an der ursprünglichen Verteilung mit den Werten aller Features ist
  - Reduktion der Anzahl der Muster in der Datenmenge ergibt einfacheres Verständnis
- Heuristische Methoden (aufgrund exponentieller Auswahlmöglichkeiten):
  - step-wise forward selection
  - step-wise backward elimination
  - Kombination
  - decision-tree induction

Beispiel: Entscheidungsbaum Induktion

ursprüngliche Attributmenge:
{A1, A2, A3, A4, A5, A6}



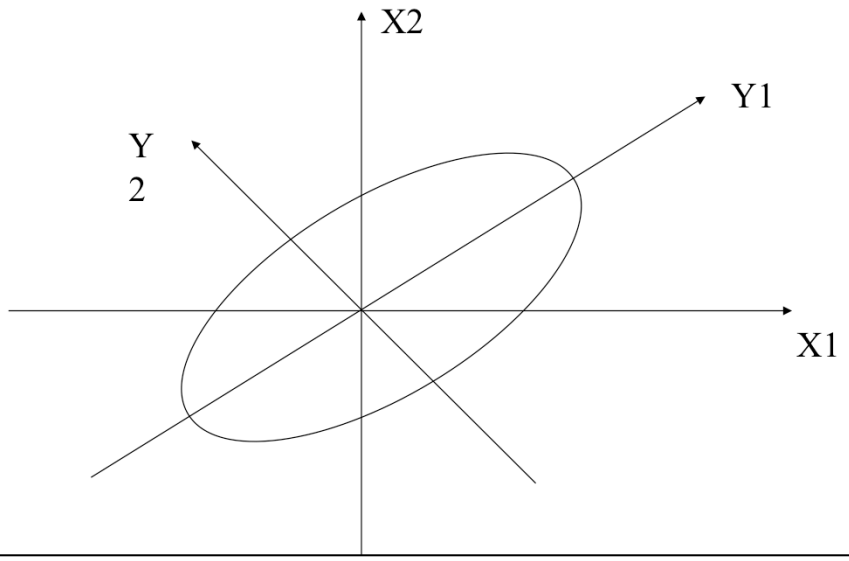
------> Reduzierte Attributmenge: {A1, A4, A6}

Attributreduktion

Attributreduktion durch Hauptkomponenten oder Faktoren-Analyse

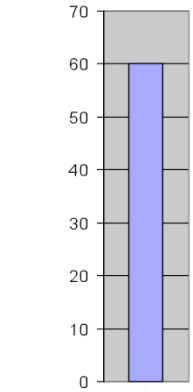
- Grundidee: Gegeben *N* Datenvektoren mit *k* Dimensionen, suche *c* <= *k* (orthogonale) Vektoren, die am besten geeignet sind, die Daten zu repräsentieren
- Dimensionsreduktion: *c* Hauptkomponenten / Faktoren, statt k Attributen
- Die Hauptkomponenten/Faktoren sind absteigend geordnet nach Signifikanz/ Varianz, Eliminierung "schwacher" Komponenten, bzw. "schwacher" Einflüsse, erlaubt weitere Dimensions-reduzierung
- insbesondere geeignet für hochdimensionale numerische Daten
- Hauptkomponenten/Faktoren (und damit die reduzierten Daten) sind nicht immer leicht zu interpretieren

Beispiel: Hauptkomponenten-Analyse

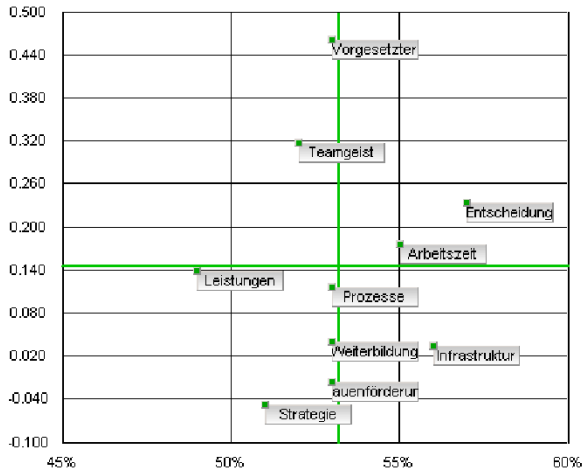


Beispiel: Faktorenanalyse

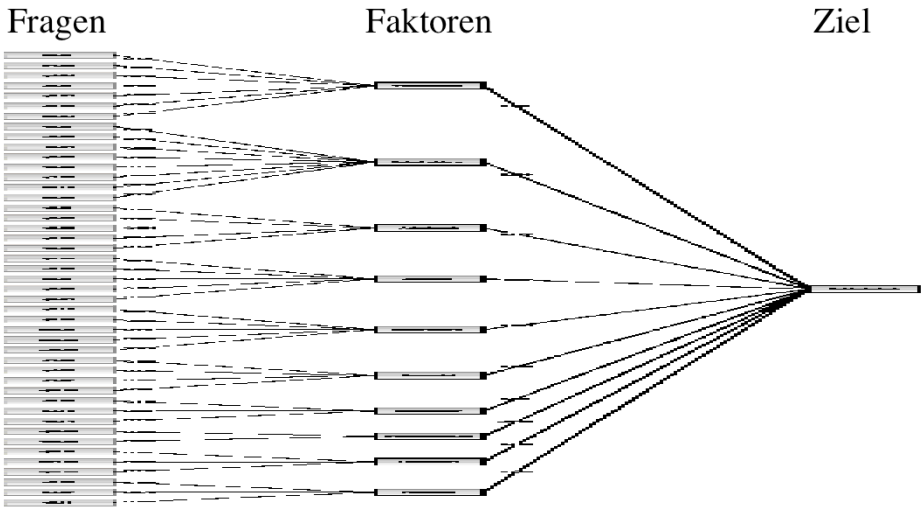
Kunden & Mitarbeiter Zufriedenheit



Gesamtzufriedenheit



Faktoren der Gesamtzufriedenheit



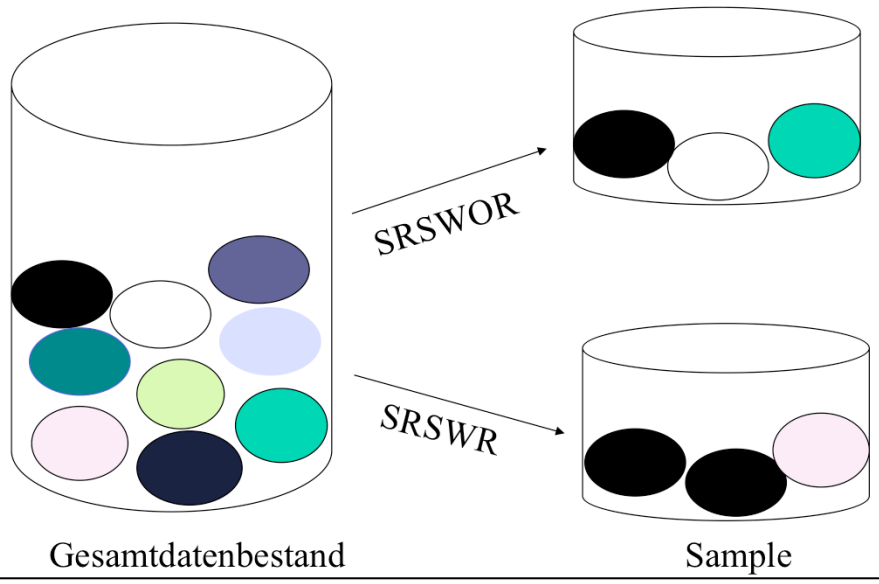
Heutiges Programm

- Ziele der Datenaufbereitung
- Daten Transformation
  - Erstellung abgeleiteter Attribute
  - Erstellung abgeleiteter Datensätze
- Daten Integration
- Daten Reduktion
  - Attribut Reduktion/Selektion
  - Sampling
- Ein Beispiel

Sampling

- Wende Data Mining Verfahren nur auf Datenausschnitte an: Performanceverbesserung vs. Kapazitätsgrenze
- Wähle (repräsentative) Teilmenge der Daten
  - vergleichsweise geringe Komplexität (O(Grösse des Samples))
  - einfaches Sampling kann schlechte Performance ergeben
- Systematischere Ansätze:
  - Cluster Sampling
    behandele ganze Cluster separat
  - Stratifiziertes Sampling
    erhalte Wertverteilungen bestimmter Attribute
    (z.B. Klassenzugehörigkeit) durch “gleichmässige” Auswahl

Beispiel: einfaches Sampling



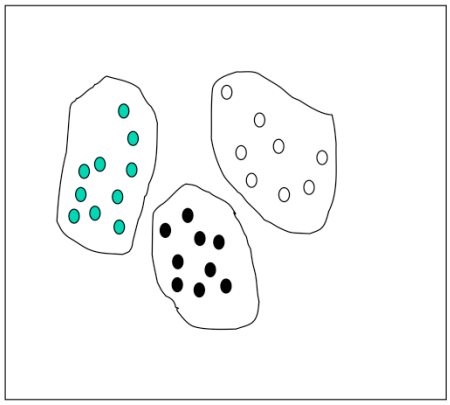
Beispiel: Stratifiziertes Sampling

T38	young
T256	young
T307	young
T391	young
T96	middle-aged
T117	middle-aged
T138	middle-aged
T263	middle-aged
T290	middle-aged
T308	middle-aged
T326	middle-aged
T387	middle-aged
T69	senior
T284	senior

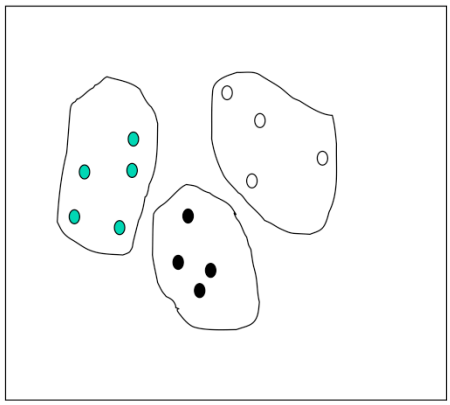
T38	young
T391	young
T117	middle-aged
T138	middle-aged
T290	middle-aged
T326	middle-aged
T69	senior

Beispiel: Cluster/stratifiziertes Sampling

Gesamtdatenbestand



Cluster/stratifiziertes Sample



Heutiges Programm

- Ziele der Datenaufbereitung
- Daten Transformation
  - Erstellung abgeleiteter Attribute
  - Erstellung abgeleiteter Datensätze
- Daten Integration
- Daten Reduktion
  - Attribut Reduktion/Selektion
  - Sampling


Ein Beispiel

Daten in mehreren Relationen

- Mögliche DM-Ziele
  - Klassifikation der Kunden in ABC-Kunden.
  - Produktvorschläge

Fazit

- Die Datenaufbereitung dient der Bereitstellung aller **relevanten** Daten in der **nützlichsten** Form für das DM-tool
- Daten Transformationen sollen helfen vermutete/interesante Zusammenhänge in den Hypothesenraum des Data Mining Tools zu bringen,
- Zur Datentransformation und Reduktion gehören:
  - Attribute Konstruktion und Selektion
  - Beispiel Konstruktion und Selektion
- Eine problembezogene Datenaufbereitung ist entscheidend für erfolgreiches Data Mining (GIGO = Garbage In ⇒ Garbage Out)

Literatur

- Pyle, D.: Data Preparation for Data Mining, Morgan Kaufmann, 1999.