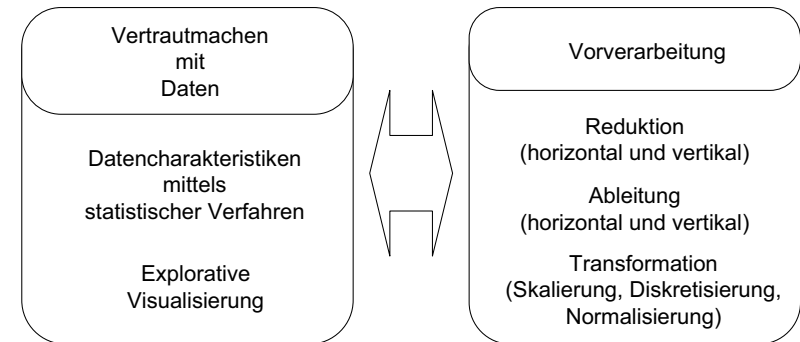


# Kapitel 6 Vorverarbeitung

## Einführung der Vorverarbeitung

### Verknüpfung von Vorverarbeitung und Datenverständnis



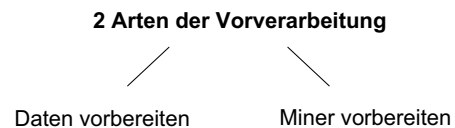
## 6 Vorverarbeitung

### 6.1 Einführung in die Vorverarbeitung

#### Zweck der Vorverarbeitung

- Transformiere die Daten so, dass sie optimal vom Miner verarbeitet werden können.

**Problem:** - lerne die „wahre“ **Beziehung**  
- lerne keinen **Lärm**



## Einführung in die Vorverarbeitung

### Weitere Aspekte der Vorverarbeitung:

Gute Vorverarbeitung benötigt das Wissen eines Domänenexperten

#### **Daten-Kontext**

- Falsche Verteilung
- nominal vs. ordinal
- Korrelation

#### **Domänen-Kontext**

- richtige Daten im Kontext?
- Repräsentieren die Daten gesuchte Zusammenhänge?
- weitere Daten notwendig?

## Einführung in die Vorverarbeitung

### Beispiel:

Repräsentieren meine Daten die gesuchten Zusammenhänge aus Sicht der Domäne?

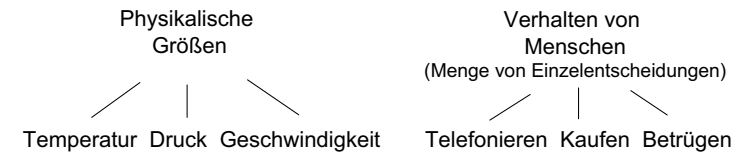
#### OLAP

- Visualisiert schnell die Zusammenhänge in den Daten
- Daten können schnell manipuliert werden

- Domänenexperte bekommt schnell einen Überblick über den aktuellen Stand der Daten und kann die gestellte Frage beantworten.

## Einführung in die Vorverarbeitung

### Prinzipielle Unterschiede bei der Vorverarbeitung



Es ergeben sich unterschiedliche Probleme bei den Daten:

- komplexe Zusammenhänge
- meist nicht-linear
- konsistent
- häufig fehlende Werte (missing values)
- sehr große Datenmenge
- häufig inkonsistent

### Beispiel:

Prozessoptimierung in einem Chemieunternehmen

Analyse des Verhaltens der Kunden einer Telefongesellschaft

## Einführung in die Vorverarbeitung

### Telekom:

Bei der Deutschen Telekom AG sammelt man „nur“ über die eigenen Kunden schon länger Informationen über das Gesprächsverhalten. Für eine wieder einmal durchzuführende Analyse stellt sich die Frage:

Kann man mit den Daten aus dem Jahr 1997 eine allgemeingültige Aussage z.B. über das Telefonverhalten der Deutschen machen? Geht dies auch noch 1999?

- 1997 ja
- aber 1999 nicht mehr, da nicht mehr alle Telefonbesitzer auch Kunden der Deutschen Telekom AG sind.
- Außerdem ist zu beachten, daß auch 1997 mit den gesammelten Daten nur Aussagen über die Telefonbesitzer gemacht werden können, nicht aber über alle Deutschen.

## Vorverarbeitungsschritte

Datenbereinigung

- **Konsistenz**
- **Detail-/Aggregationsniveau**
- **Verunreinigung**
- **Beziehungen**
- **Definitionsbereich**
- **Defaultwerte**
- **Duplikate oder redundante Variablen**
- **fehlende und leere Felder**

Datenmanipulation

- **reverse Pivotisierung**
- **Reduzierung der Dimensionalität**
- **Anstieg der Dimensionalität**
- **dünn besetzte Werte**
- **aufzählende kategorische Werte**
- **Monotonie der Daten**
- **Ausreisser**
- **Anachronismen**
- **aufzählende kategorische Werte**
- **Beziehungen zwischen den Variablen**
- **kombinatorische Explosion**

## Vorverarbeitungs-Schritte

### a) Datenbereinigung

#### Konsistenz

- verschiedene Dinge werden in verschiedenen Systemen mit dem gleichen Namen dargestellt
- gleiche Dinge werden mit verschiedenen Namen in verschiedenen Systemen dargestellt

#### Detail / Aggregationsniveau

- Transaktionsdaten (detailliert) gegenüber aggregierter Menge von Transaktionen
- Allgemeine Regel für Data Mining: detaillierte Daten werden gegenüber aggregierten Daten bevorzugt
- Das Detailniveau im Eingabestrom ist um ein Aggregationsniveau detaillierter als das erforderliche Detailniveau in der Ausgabe

## Vorverarbeitungs-Stufen

### Duplikate oder redundante Variable

- identische Informationen in Mehrfachvariablen, z.B. „Geburtsdatum“ und „Alter“
- Probleme z.B. für neurales Netzwerk bei Kolinearität der Variablen

### Fehlende und leere Felder

- Leere Felder haben u.U. keinen entsprechenden realen Wert oder haben einen realen Wert, aber dieser wurde nicht erfasst
- Miner sollte zwischen beiden Arten von Werten differenzieren
- Data Mining Tools haben verschiedene Strategien, um diese Werte zu bearbeiten

## Vorverarbeitungs-Stufen

### Verunreinigung

Datenmüll, z.B. Komma-begrenzte Daten, die Kommata enthalten

- Menschlicher Widerstand bei der Datenerfassung, z.B. leere, unvollständige und ungenaue Datenfelder

### Beziehungen

- das Kombinieren von verschiedenen Eingabeströmen (verwende zum Beispiel Schlüssel)
- Finde die richtigen Schlüssel, eliminiere doppelte Schlüssel

### Definitionsbereich

Variable haben meist einen speziellen Wertebereich

- erkenne Ausreisser

### Defaults

Der Miner muss die Defaults von Datenerfassungsprogrammen kennen

- bedingte Defaults können scheinbar signifikante Strukturen kreieren

## Vorverarbeitungs-Stufen

### b) Datenmanipulation

#### Reverse Pivottisierung

- Modellierung wichtiger Dinge unter dem richtigen Gesichtspunkt
- Beispiel: Datenbank mit detaillierten Abrufaufzeichnungen  
Aufgabe ist, die Kunden zu analysieren  
Problem: Der Fokus der Datenbank ist nicht der Kunde

#### Reduzierung der Dimensionalität

- Eliminiere Merkmale, die für deine Aufgabe nicht wichtig sind

#### Erhöhung der Dimensionalität

- Expandiere eine Dimension, um die Information in einer besseren Weise darzustellen
- Beispiel: Postleitzahl kann in „Lat“ und „Lon“ transformiert werden

## Vorverarbeitungsstufen

### Seltenheit

- manche Variablen sind nur dünn mit Instanzwerten bevölkert
- Der Miner muss entscheiden, die Variable z.B. zu ignorieren oder zu kollabieren

### Stichproben

- nimm nur einen Teil der gesammelten Daten, aber verliere keine Information

### Monotonie

- eine monotone Variable ist (hier) eine, die sich ohne Schranke vergrößert
- Beispielvariable: Datum, Zeit, Sozialnummer
- müssen in nicht-monotone Variable transformiert (oder ignoriert) werden, da eine Vorhersage nur im Bereich der Trainingsdaten Datenmenge ausgeführt werden kann

### Ausreisser

- einzeln oder mit sehr geringer Häufigkeit auftretender Wert einer Variablen
- weit weg von der Masse der Werte der Variablen
- ist der Ausreißer ein Fehler oder eine sehr wichtige Information?

## Vorverarbeitungs-Beispiel für kategorische Daten

Wie kategorische Werte am besten dargestellt werden, hängt stark von den Anforderungen des Modelliertools ab.

### Bsp.: Aufzählung kategorischer Werte

Zeitraum	Lohnsatz (\$)
Halber Tag	100
Tag	200
Halbe Woche	500
Woche	1000
Halber Monat	2000
Monat	4000

Zeitraum	...	Lohnsatz (\$)
Halbe Woche	1	500
Halber Monat	2	2000
Halber Tag	3	100
Monat	4	4000
Tag	5	200
Woche	6	1000

- beide Tabellen zeigen den Lohnsatz für einen Zeitraum in Dollar
- man sieht keine Struktur, wenn man die Zeiträume alphabetisch sortiert (rechte Tabelle)

## Vorverarbeitungs-Stufen

### Anachronismen

- Information, die nicht tatsächlich in den Daten verfügbar ist, wenn eine Vorhersage erforderlich ist.

### Beziehungen zwischen Variablen

- genug Instanzwerte werden zur Darstellung der Variablen benötigt, um statistische Beziehungen zwischen den Variablen zu erkennen
- Ggf. müssen korrelierte Variablen gelöscht werden
- Abhängigkeiten zwischen Variablen sind auch interessant

### Kombinatorische Explosion

- Wenn man an den Interaktionen zwischen Variablen interessiert ist, muss man jede Kombination der Variablen prüfen.

Anzahl von Variablen	Anzahl Kombinationen
5	26
9	502
25	33.554.406

## Vorverarbeitungs-Beispiel für kategorische Daten

### Problem:

Schlimmstenfalls (wenn das Skalierungsniveau nominal ist) werden durch die Aufzählung kategorischer Werte Strukturen in die Daten eingeführt oder geschaffen, die nicht natürlich sind.

- **Domainenwissen kann helfen**
- **Verursache an der 'natürlichen Struktur' so wenig Schaden wie möglich**

Don't torture your data until they confess!

## Vorverarbeitung Beispiel für kategorische Daten

### Distanzmaß für kategorische Daten

Example: call detail record

KundeID	Distanz	Wochentag	Datum/Uhrzeit	komm. Minuten
1	Ort	Mo-Fr	19.11.98/9:55	20 min
1	Ort	Mo-Fr	20.11.98/10:10	18 min
2	Regional	Mo-Fr	19.11.98/21:00	120 min
2	Regional	Mo-Fr	20.11.98/17:00	2 min

**Problem:** Welche erfassten Daten ähneln einander?

- die einfache Antwort lautet: alle sind verschieden
- eine Person würde sagen: die ersten beiden erfassten Daten sind ähnlich zueinander, der Rest nicht.

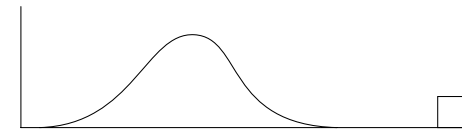
Die meisten Tools benötigen ein Ähnlichkeitsmaß, um die Daten automatisch zu verarbeiten.  
Aber prüfe immer, ob dein Maß bedeutungsvoll ist!

## Vorverarbeitungs-Beispiel für numerische Daten

### Diskretisiere numerische Variable

Nehme einen Wertebereich und bilden diesen mit einem neuen Wert ab  
Verstehe die zugrundeliegende Verteilung

Wähle den richtigen Bereich



**Warum?**

alle Variablen haben in der Praxis eine spezifische Auflösungsgrenze

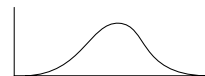
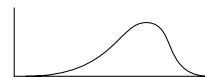
- Messgenauigkeit
- Darstellungsgenauigkeit

falls der Wert außerhalb des Bereichs liegt - werden zwei verschiedene Eingabewerte gleich

## Vorverarbeitungs-Beispiel für numerische Daten

### Normalisierung eines Variablenbereichs mit Potenzen (Tukey 1977)

p	Transformation $T(x_i)$	Name
...	...	...
10	$x_i^{10}$	dezimal
...	...	...
3	$x_i^3$	kubisch
2	$x_i^2$	quadratisch
<b>1</b>	$x_i$	<b>Rohdaten</b>
$\frac{1}{2}$	$\sqrt{x_i}$	Wurzel
0	$\log(x_i)$	logarithmisch
$-\frac{1}{2}$	$-\frac{1}{\sqrt{x_i}}$	reziproke Wurzel
1	$-\frac{1}{x_i}$	reziprok
...	...	...



## Vorverarbeitungs-Beispiel für fehlende und leere Felder

### Unterschied zwischen fehlenden und leeren Feldern

Leere Werte haben keinen entsprechenden realen Wert  
Fehlende Werte haben einen realen Wert, aber dieser wurde nicht erfasst

### Tools können Schwierigkeiten haben, solche Werte zu bearbeiten

- ignoriere fehlende und leere Felder  
- Verwende ein Maß, um einen „passenden“ Ersatz zu bestimmen  
Aber: Automatisierte Ersatztechniken sind kritisch

- Kennt der Miner die Probleme der Technik?
- Kennt der Miner die verwendete Ersatzmethode?
- Was sind ihre Begrenzungen?

### Aufgabe für Miner

Ersatz muss so **neutral** wie möglich sein  
Verwende eine vom Miner verstandene und gesteuerte Methode

### **Ersatz: Probleme und Aspekte**

- Einige Modellierungstechniken beschäftigen sich mit fehlenden Werten
- Ersetzungen mit Defaults können Verzerrung einführen
- kenne und steuere die Eigenschaften jeder Ersatzmethode
- Wichtige Informationen sind manchmal in der Verteilung der fehlenden Werte enthalten