Dr. Andreas Hotho, Robert Jäschke
Fachgebiet Wissensverarbeitung

30.10.2008

1. Übung „Knowledge Discovery“

Wintersemester 2008/2009

Vorbemerkungen

Vorlesungsfolien und Übungsblätter können Sie im Internet unter der Adresse `http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd` oder mittels folgender RSS-Feeds einsehen:

- Übungen: `http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd/uebung/rss`
- Folien: `http://www.kde.cs.uni-kassel.de/lehre/ws2008-09/kdd/folien/rss`

Bei Fragen wenden Sie sich bitte an Robert Jäschke (jaeschke@cs.uni-kassel.de).

1 Allgemeines

1. Was ist KDD und was ist insbesondere das Ziel davon?

   Definition: *„Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data“*
   Ziel ist die Entdeckung von Wissen in Form von Mustern aus einer Menge von Fakten (Daten)

2. Was ist der Unterschied zwischen Data Mining und Knowledge Discovery?

   Zwei Interpretationen:

   a) Data Mining = gesamter KDD Prozess

   b) Data Mining = Teil des KDD Prozesses (Mustergewinnung, Modellierung, Anwendung von Algorithmen $\Longrightarrow$ 4. Schritt in CRISP-DM)

3. Geben Sie Beispiele für Bereiche an, in denen KDD angewendet wird.

   a) Data Mining in Kaufhäusern (Chipkarte)

   b) Segmentierung von Kunden in der Telekommunikationsbranche

   c) Betrugserkennung (z.B. Handy)

4. Welche Geschäftsziele werden typischerweise durch KDD unterstützt? Diskutieren sie diese anhand der von Ihnen in 3. genannten Anwendungsbereiche.

   a) Kundenbindung:
      - besondere Angebote für bestimmte Gruppen von Kunden
      - schnellere Reaktion auf Änderungen im Kaufverhalten

   b) gezieltere Werbemaßnahmen

   c) Optimierung (z. B. Lagerkosten)

   d) Planung, Prognose (z. B. Telekommunikationsinfrastruktur)

5. Geben Sie vier typische Verfahren/Methoden an, die im Rahmen von KDD Anwendung finden und beschreiben Sie diese kurz.

   Segmentierung: Einteilung der Daten in Gruppen mit gemeinsamen Eigenschaften

   Klassifikation: Zuweisung von Objekten in vordefinierte Klassen (diskret)

   Vorhersage: im Prinzip wie Klassifikation, aber mit kontinuierlichen (numerischen) Zielvariablen

   Abhängigkeitsanalyse: Entdecken von Beziehungen zwischen Objekten, z.B. „Kunden, die Nachos kaufen, kaufen auch Salsa Dip“

   Abweichungsanalyse: Feststellen von abweichenden Werten in Daten

2 Überwachte vs. Unüberwachte Verfahren

1. Was sind die wesentlichen Unterschiede zwischen einem überwachten und einem unüberwachten Verfahren?

   Bei überwachten Verfahren sind die Klassen, in die Daten eingeteilt werden sollen vorgegeben, beim Unüberwachten hingegen nicht. Das überwachte Verfahren lernt dementsprechend anhand einer bestimmten Anzahl von positiven oder negativen Beispielen.

2. Was für Konsequenzen hat die Verwendung eines überwachten bzw. unüberwachten Verfahrens für die zur Verfügung zu stellenden Daten?

   Aus 1. folgt, dass man für ein überwachtes Verfahren immer einen Trainingsdatensatz braucht, bei dem die Objekte der korrekten Klasse zugeordnet sind. Zur Verifikation des Verfahrens braucht man dann auch einen Testdatensatz.

3. Nennen sie jeweils zwei Anwendungen für ein überwachtes und ein unüberwachtes Verfahren.

- Überwachte Verfahren:
  - Klassifikation z. B. BETRUG und NICHT_BETRUG
  - Vorhersage, z. B. wie viel werden die Kunden diese Weihnachten ausgeben?
- Unüberwachte Verfahren:
  - Segmentierung von Kunden (Telekom)
  - Entdeckung von Assoziationsregeln (Kaufverhalten)

3 CRISP-DM Methodologie

1. Nennen Sie die sechs Phasen der CRISP-DM Methodologie.

   a) Business Understanding (was will man erreichen)
   b) Data Understanding (mit was für Daten habe ich es zu tun)
   c) Data Preparation (Vorverarbeitung für Modellierung)
   d) Modelling (Anwendung der Algorithmen)
   e) Evaluation (Interpretation, Ziel erreicht?)
   f) Deployment (Umsetzung der Ergebnisse im Unternehmen)

2. Was sind die wichtigsten Schritte in der Vorverarbeitung?

   a) Datenselektion
   b) Datenreinigung (Fehler, Defaults, fehlende Werte)
   c) Datenkonstruktion
      - Normalisisierung
      - Transformation
      - Ableitung von Attributen (Summe, Mittelwert, etc.)
   d) Integration (verschiedene Quellen)
   e) Formatierung

3. Was für Probleme ergeben sich dabei typischerweise?

   a) Schätzung fehlender Werte
   b) Erkennung falscher Werte
   c) Formatkonflikte bei Integration

4. Wie hängt diese Phase konzeptuell mit den anderen Phasen zusammen?

   a) Datenselektion hängt von den Zielen ab (Business Understanding)

   b) Datenpräparierung, Modellierung, Evaluierung hängt stark von ausgewählten Daten ab

   c) Modellierung, Evaluierung hat wiederum Einfluss auf Datenselektion (iterativer Prozess)

4 Vorverarbeitung

Ein Versandhändler möchte seinen Kundenbestand analysieren, um den aktivsten Kunden besondere Angebote zu machen. Dazu hat er Ihnen folgende Stichprobe seiner Daten bereitgestellt:

Kunden					
Id	Name	E-Mail-Adresse	Strasse	Ort	PLZ
1	Carla D. Eiffel		Forsthausweg 2	Duisburg	47057
2	F. Ganter	ganter@gxm.de	Geschwister-Scholl-Platz 1	München	80539
3	Jan Klein	jan_klein@gmail.com	Kaiserswerther Str. 16	Berlin	14195
4	Anton BlÃcher	bluecher@gmx.de	Rosengarten 10	Halle/Saale	6132
6	Irving, Hans	hans.irving@web.de	Christian-Albrechts-Platz 4	Kiel	24118
7	Ludwig Mann	lm@lumann.com	Kaiserswerther Strasse 16	Berlin	14195

Kaufdaten Online-Shop					
Id	Kunden-Id	Datum	Artikel-Id	Preis	Anzahl
1	1	1.1.1970	1	12,99	2
2	1	1.1.1970	5	5,49	1
3	2	12.6.2006	3	15,00	1
4	5	20.6.2007	2	2,00	4
5	3	21.6.2006	5	5,99	1
6	1	1.1.1970	1	12,99	255

Kaufdaten telefonische Bestellung				
Kunden-Nr	Datum	Artikel	Preis	Menge
3	3.6.06	2	2	2
3	10.6.06	1	12,99	1
4	4.6.06	2	2,00	1
1	3.6.06	1	12,99	5
7	9.6.06	5	5,99	1

1. Wenden Sie (soweit möglich) die in Aufgabe 3, Teil 2 genannten Schritte der Datenvorverarbeitung auf den folgenden Datensatz an. Machen Sie sich insbesondere klar, welches konkrete Vorgehen zu welchem Schritt gehört.

Kunden					
Id	Name	E-Mail-Adresse	Strasse	Ort	PLZ
1	Carla D. Eiffel		Forsthausweg 2	Duisburg	47057
2	F. Ganter	ganter@gxm.de	Geschwister-Scholl-Platz 1	München	80539
3	Jan Klein	jan_klein@gmail.com	Kaiserswerther Strasse 16	Berlin	14195
4	Anton Blücher	bluecher@gmx.de	Rosengarten 10	Halle/Saale	06132
~~6~~	~~Hans Irving~~	~~hans.irving@web.de~~	~~Christian-Albrechts-Platz 4~~	~~Kiel~~	~~24118~~
7	Ludwig Mann	lm@lumann.com	Kaiserswerther Strasse 16	Berlin	14195

Durchgeführte Schritte: Selektion (Streichen von Kunde Nr. 6, da er in den Kaufdaten nicht auftaucht), Datenreinigung (Umlaute, führende Nullen in PLZ), Normalisierung (Str. wird zu Strasse).

Kaufdaten Online-Shop					
Id	Kunden-Id	Datum	Artikel-Id	Preis	Anzahl
1	1	11.6.2006	1	12,99	2
2	1	11.6.2006	5	5,99	1
3	2	12.6.2006	3	15,00	1
~~4~~	~~5~~	~~20.6.2007~~	~~2~~	~~2,00~~	~~4~~
5	3	21.6.2006	5	5,99	1
6	1	23.6.2006	1	12,99	1

Durchgeführte Schritte: Datenreinigung (fehlende Datums-Werte eingefügt aus Logfile, Preis für Artikel 5 korrigiert, Anzahl für Transaktion Nr. 6 korrigiert), Datenselektion (Entfernen von Transaktion Nr. 4, da Kunde Nr. 5 in Kundendaten fehlt).

Kaufdaten telefonische Bestellung					
Id	Kunden-Id	Datum	Artikel-Id	Preis	Anzahl
7	3	3.6.2006	2	2,00	2
8	3	10.6.2006	1	12,99	1
9	4	4.6.2006	2	2,00	1
10	1	3.6.2006	1	12,99	5
11	7	9.6.2006	5	5,99	1

Durchgeführte Schritte: Integration (Spalten-Namen), Formatierung (Datum), Normalisierung (Preis), fehlende Werte ergänzt (Transaktions-Ids),

Wie hängen die vorzunehmenden Schritte vom Ziel der Datenanalyse ab?

Das Ziel der Analyse ist wesentlicher Faktor bei der Entscheidung, welche Schritte der Vorverarbeitung anzuwenden sind. Beispielsweise ist das Bereinigen der Namen und E-Mail-Adressen der Kunden nur notwendig, wenn diese zur Analyse notwendig sind. Für viele Ziele dürften zudem die Spalten *Id* und *PLZ* der Kundentabelle ausreichend sein. Des weiteren kann es je nach Ziel sinnvoll sein, die Datums-, Preis- oder Anzahl-Daten weiter zu aggregieren (beispielsweise nach Monat, Preisgruppe- oder Mengengruppe).

2. Diskutieren Sie in der Gruppe die auftretenden Probleme und wie Sie diese lösen könnten.

- Fehlende Werte können u. U. aus einem Logfile beschafft werden.
- Falsche Werte (z. B. 255 als Kaufanzahl) können durch Mittelwerte ersetzt oder der entsprechende Datensatz komplett entfernt werden.
- Strengere, spezifischere Eingabemasken und Überprüfungen bei der Eingabe der Kunden-/Kaudaten, saubere Modellierung des Datenbankschemas sowie Wertüberprüfungen, Verwendung von Transaktionen und Logging können helfen, Fehler zu vermeiden bzw. im Nachhinein zu korrigieren.

5 Datenbanken

1. Definieren Sie informell ein Datenbanksystem und beschreiben Sie den prinzipiellen Aufbau.

   Ein *Datenbanksystem* (DBS) ist ein Software System zur dauerhaften Speicherung und zum effizienten Suchen in großen Datenmengen.
   Ein DBS besteht aus einer *Datenbank* (DB), die die eigentliche Sammlung der Daten einer gegebenen Anwendung darstellt und aus einem *Datenbank-Management-System* (DBMS). Beim DBMs handelt es sich um ein Computer-Programm zum Management von Datenbanken. Gleichzeitig erlaubt es beliebigen Anwendungen in einem spezifizierten Format den Zugriff auf die Daten einer DB.

2. Beschreiben Sie den prinzipiellen Unterschied für den Zugriff auf Daten einer Datenbank beim Data Mining gegenüber einer klassischen Datenbankanwendung.

   Beim klassischen Zugriff auf Daten einer Datenbank werden typischerweise viele Einfüge-, Update- und auch Löschoperationen von vielen Anwendungen in konkurrierender Weise auf die Datenbank ausgeführt. Beim Data Mining werden eher einmalig große Mengen, auch von Bestandsdaten die über die Zeit aggregiert sind, angefragt. Dabei werden die Daten typischerweise nur gelesen.

3. Geben Sie die Eigenschaften eines B-Baumes wieder und begründen Sie, warum ein B-Baum balanciert sein muss.

   - Jeder Knoten enthält höchstens 2m Schlüssel.
   - Jeder Knoten außer der Wurzel enthält mindestens m Schlüssel, die Wurzel mindestens einen Schlüssel.
   - Ein Knoten mit k Schlüsseln hat genau $k + 1$ Söhne.
   - Alle Blätter befinden sich auf demselben Level.

   Nur ein balancierter B-Baum bietet die Möglichkeit in $O(\log n)$ auf die Seite einer Datenbank mit den entsprechenden Daten zuzugreifen. Ist der Baum nicht mehr balanciert, so erhöht sich die Zugriffszeit (im schlechtesten Fall bis auf $O(n)$).

6 Einführung Statistik

1. Gegeben sei folgende Tabelle:

Kredithöhe in Euro	300 schlechte Kunden (in Prozent)	700 gute Kunden (in Prozent)
$0 < \cdots \leq 500$	1.00	2.14
$500 < \cdots \leq 1000$	11.33	9.14
$1000 < \cdots \leq 1500$	17.00	19.86
$1500 < \cdots \leq 2500$	19.67	24.57
$2500 < \cdots \leq 5000$	25.00	28.57
$5000 < \cdots \leq 7500$	11.33	9.71
$7500 < \cdots \leq 10000$	6.67	3.71
$10000 < \cdots \leq 15000$	7.00	2.00
$15000 < \cdots \leq 20000$	1.00	0.29
Frühere Kredite		
gut	82.33	94.85
schlecht	17.66	5.15

Die Tabelle gibt die Merkmale von 1000 Kunden einer Bank wieder, die ihren Kredit mit/ohne Probleme zurückgezahlt haben. Die Bank möchte von Ihnen ein Modell zur Vorhersage, ob neue Kunden Probleme bei der Rückzahlung ihres Kredites machen werden oder nicht. Stellen Sie dazu die Informationen der Tabelle in geeigneter Form dar, um sich einen ersten Eindruck von den Daten zu machen. Interpretieren Sie das Ergebnis und vergleichen Sie vor allen Dingen die Verteilungen der Merkmale. Berechnen Sie das arithmetische Mittel und den Median. Interpretieren Sie die Ergebnisse und setzen Sie die Ergebnisse in Bezug zu den Verteilungen der dargestellten Merkmale.

$$\overline{x}_{\text{art}} = \sum_{i=1}^{k} f_i m_i$$

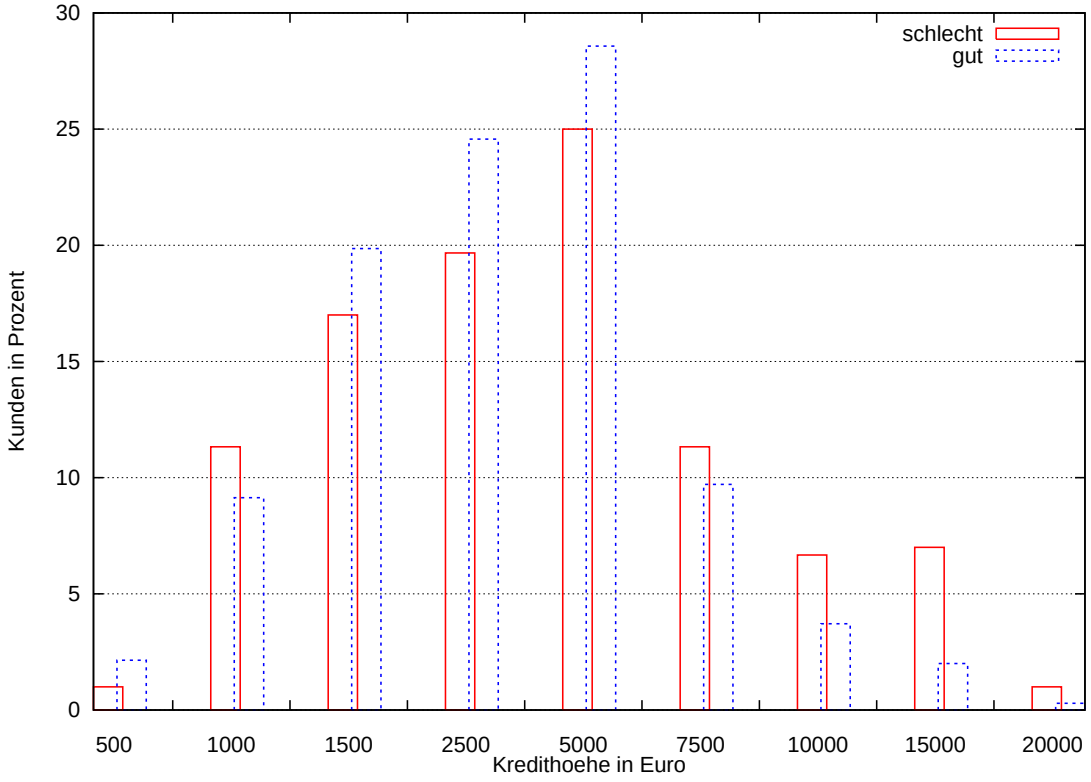
$$\overline{x}_{\text{med}} = c_{i-1} + \frac{d_i(0.5 - F(c_{i-1}))}{f_i}$$

$$x_{\text{art,schlecht}} = 3972 \text{ Euro}$$

$$x_{\text{art,gut}} = 3117 \text{ Euro}$$

$$x_{\text{med,schlecht}} = 2603 \text{ Euro}$$

$$x_{\text{med,gut}} = 2267 \text{ Euro}$$



2. Ein Experiment bestehe aus dem Werfen eines Würfels und einer Münze. Geben Sie einen geeigneten Ergebnisraum Ω an. Zeigt die Münze Wappen, so wird die doppelte Augenzahl des Würfels notiert, bei Zahl nur die einfache. Wie groß ist die Wahrscheinlichkeit, dass eine gerade Zahl notiert wird?

$$\Omega = \{(1, Z), (2, Z), (3, Z), (4, Z), (5, Z), (6, Z), (2, W), (4, W), (6, W), (8, W), (10, W), (12, W)\}$$

$$P = \frac{\text{Anzahl der Ergebnisse mit geraden Zahlen im Ergebnisraum}}{\text{Anzahl aller Ergebnisse im Ergebnisraum}} = \frac{9}{12} = \frac{3}{4}$$