Universität Kassel
Fachbereich Mathematik/Informatik
Fachgebiet Wissensverarbeitung
Hertie-Stiftungslehrstuhl
Wilhelmshöher Allee 73
34121 Kassel

Email: hotho@cs.uni-kassel.de
Tel.: ++49 561 804-6252

Dr. Andreas Hotho
Prof. Gerd Stumme

24.05.04

Knowledge Discovery

Lösungsblatt 1

Sommersemester 2004

Aufgabe 1: Allgemeines

a) 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)

b) Was ist der Unterschied zwischen Data Mining und Knowledge Discovery?

Zwei Interpretationen:

1) Data Mining = gesamter KDD Prozess
2) Data Mining = Teil des KDD Prozesses (Mustergewinnung, Modellierung, Anwendung von Algorithmen => 4. Schritt in CRISP-DM)

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

1) Data Mining in Kaufhäusern (Chipkarte)
2) Segementierung von Kunden in der Telekommunikationsbranche
3) Betrugserkennung (z.B. Handy)

d) Welche Geschäftsziele werden typischerweise durch KDD verfolgt? Diskutieren sie diese anhand der von Ihnen in c) genannten Anwendungsbereiche.

1) Kundenbindung: - besondere Angebote für bestimmte Gruppen von Kunden
   - schnellere Reaktion auf Änderungen im Kaufverhalten
2) gezieltere Werbemaßnahmen
3) Optimierung (z.B. Lagerkosten)

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

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

1) Segmentierung: Einteilung der Daten in Gruppen mit gemeinsamen Eigenschaften
2) Klassifikation: Zuweisung von Objekten in vordefinierte Klassen (diskret)
3) Vorhersage: im Prinzip wie Klassifikation, aber mit kontinuerlichen (numerischen) Zielvariable
4) Abhängigkeitsanalyse: Entdecken von Beziehungen zwischen Objekten, z.B. „Kunden, die Nachos kaufen, kaufen auch Salsa Dip“
5) Abweichungsanalyse: Festellen von abweichenden Werten in Daten

Aufgabe 2: Überwachte vs. Unüberwachte Verfahren

a) 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.

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

Aus a) folgt eben, 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.

c) 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)

Aufgabe 3: CRISP-DM Methodologie

a) Nennen Sie die sechs Phasen der CRISP-DM Methodologie.

Business Understanding (was will man erreichen)

Data Understanding (mit was für Daten habe ich es zu tun)

Data Preparation (Vorverarbeitung für Modellierung)

Modelling (Anwendung der Algorithmen)

Evaluation (Interpretation, Ziel erreicht?)

Deployment (Umsetzung der Ergebnisse im Unternehmen)

b) Was sind die wichtigsten Schritte in der Datenpräparierung?

1) Datenselektion
2) Datenreinigung (Fehler, Defaults, fehlende Werte)
3) Datenkonstruktion: - Normalisisierung
   - Transformation
   - Ableitung von Attributen (Summe, Mittelwert, etc.)
4) Integration (verschiedene Quellen)
5) Formatierung

c) Was für Probleme ergeben sich typischerweise dabei?

1) Schätzung fehlender Werte

2) Erkennung falscher Werte
3) Formatkonflikte bei Integration (Börsenbeispiel)

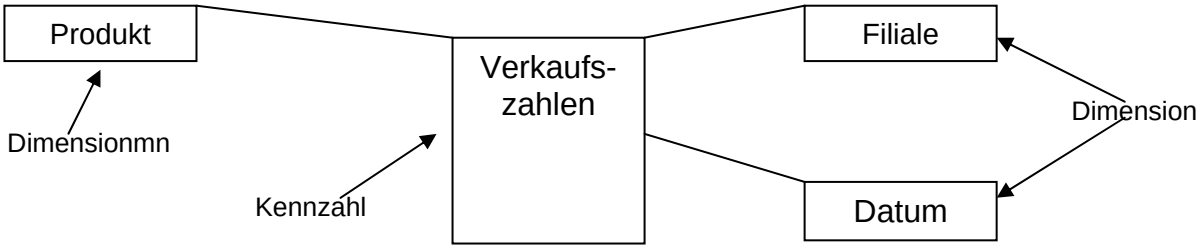
d) Wie hängt diese Phase konzeptuell mit den anderen Phasen zusammen?

1) Datenselektion hängt von den Zielen ab (Business Understanding)
2) Datenpräparierung, Modellierung, Evaluierung hängt stark von ausgewählten Daten ab
3) Modellierung, Evaluierung hat wiederum Einfluss auf Datenselektion (iterativer Prozess)

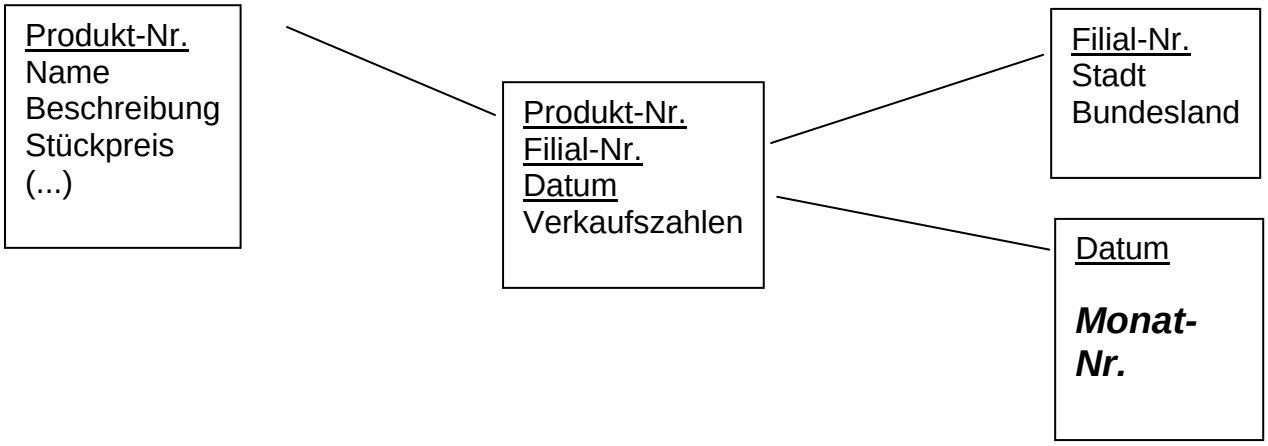
Aufgabe 4: Stern-Schema

Die Supermarktkette IDLA möchte ihre Lagerkosten optimieren. Dazu hat sie Daten darüber gesammelt, welche Produkte in welchen Filialen und in welcher Menge über einen Zeitraum von zwei Jahren verkauft worden sind. Die Einheit der Zeitmessung sind Tage. Zur Analyse dieser Daten möchte IDLA ein OLAP System einsetzen.

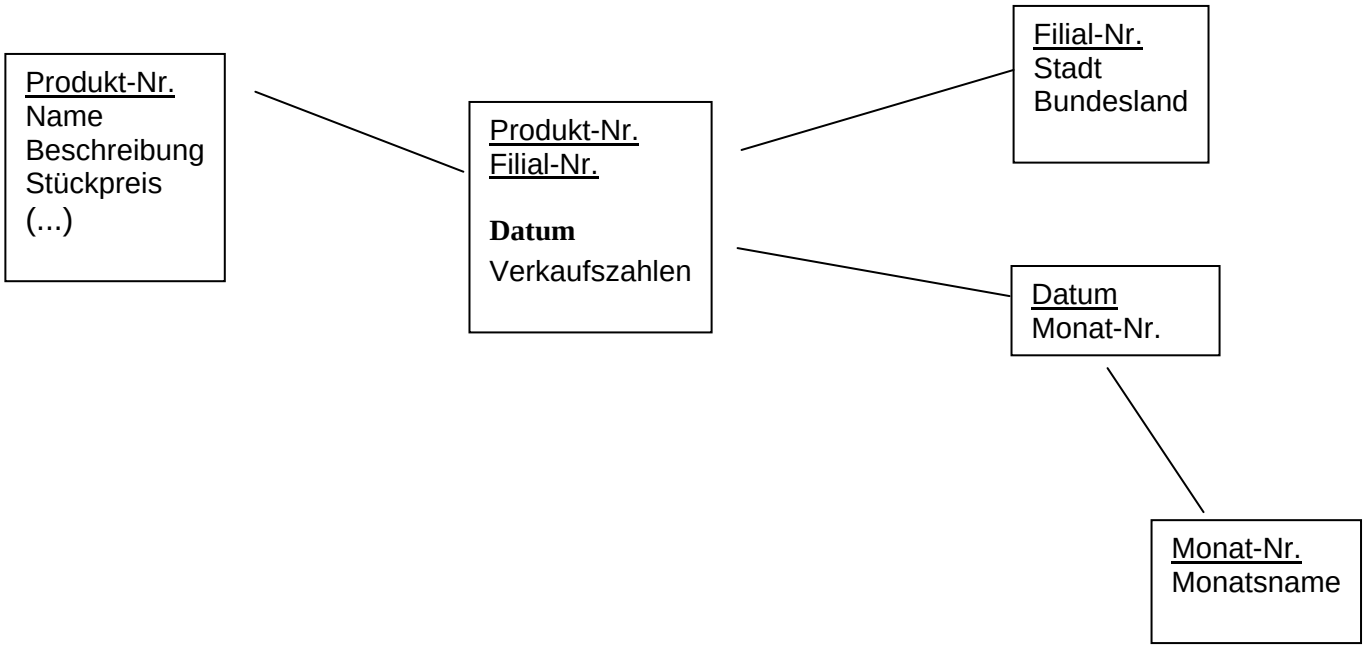
a) Entwerfen sie ein Stern-Schema für die Analyse dieser Daten. Geben sie zuerst die Kennzahl und die Dimensionen an!



b) Skizzieren Sie sowohl die Kennzahlentabelle als auch die Dimensionstabellen.



c) Erweitern Sie das obige Sternschema zu einem Schneeflockenschema unter Berücksichtigung, dass die Tagesdaten zu Monatsdaten aggregiert werden sollen.



Aufgabe 5: Datencharakteristiken

Der Internet-Provider *ich-bin-drin.de* möchte einen neuen Spam-Filter einsetzen, der auf einem Klassifikator aufbaut, der Mails in die Kategorien SPAM und NON_SPAM einteilt. Dazu hat ich-bin-drin.de einen großen Datensatz zur Verfügung, in dem Mails zusammen mit bestimmten Attributen (Empfänger, Sender, Subject usw.) gespeichert sind sowie auch als SPAM oder NICHT_SPAM markiert sind. Die Verteilung ist wie folgt: 60% der Mails sind als NICHT_SPAM, 40% als SPAM markiert. Als Klassifikationsxperte werden Sie von dem Internet-Provider angeheuert, um bei der Klärung folgender Fragen zu dienen:

a) Könnte man im Vorhinein abschätzen, ob die Klassifikation erfolgreich sein wird?

Ja, z.B. unter Verwendung von Wilks-Lambda (ohne Formel)

Wilks-Lambda: measures **class differences**:

- value near 1: no good distinction between classes
- value near 0: good distinction between classes => **good classifier** can be learned

Oder durch Verwendung der Attributentropie:

Attributentropie: gegeben ein Attribut B mit n möglichen verschiedenen Werten $b_1 ... b_n$ sowie die Warscheinlichkeit p_i für das Attribut b_i ist die Attributentropie gegeben durch:

$$H(B) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Wenn alle Werte die gleiche Wahrscheinlichkeiten haben, dann ist H(B) =1. Man sagt auch „B ist maximal unsicher“.
Wenn man nun die Attributentropie des Klassifikationsattributes berechnet bekommt man eine Aussage über die Komplexität der Klassifikationsaufgabe, d.h. je höher die Entropie des Klassifikationsattributes desto schwieriger ist die Klassifikationsaufgabe.
(Hier mal für die Studenten ausrechnen, wie hoch die Attributentropie für die Spam-Klassifikation ist.)

b) Könnte man theoretisch voraussagen, wie viel Attribute man für eine solche Klassifikation benötigen werden?

Ja, und zwar kann man diese Zahl durch die **Equivalent Number** of Attributes abschätzen:

Equivalent Number of Attributes: schätzt die Anzahl der Attribute, die benötigt werden, um den Wert de Zielattributes c zu bestimmen:

$$EN.attr = \frac{H(c)}{\bar{I}_{gain}(c, A)}$$

$\bar{I}_{gain}$ ist nun das Mittel über das I_{gain} aller Attribute:

$\bar{I}_{gain}(c, A) = \frac{1}{s}\sum_{i=1}^{s} I_{gain}(c, A_i)$, wobei

$$I_{gain}(c, A_i) = H(c) - H(c \mid A_i)$$

d.h. von H(c) (die Unsicherheit dass c einen bestimmen Wert hat), wird die Unsicherheit abgezogen, dass c einen bestimmten Wert hat unter der Voraussetzung dass der Wert von A_i gegeben ist. Wenn der Wert von c und A_i z.B. immer gleich ist, so ist die Unsicherheit von c gegeben A_i , d.h. $H(c|A_i)=0$ und damit der Information Gain von A in Bezug auf c, d.h. $I_{gain}(c,A)=H(c)$. Der Informationsgewinn ist also genau die Unsicherheit von c. Wenn alle Attribute nun c vollständig bestimmen, dann ist EN.attr=1, d.h. man braucht genau ein Attribut um den Wert von c zu bestimmen.
Hmm... die Erklärung war jetzt ein bisschen wirr, aber eine bessere fällt mir gerade nicht ein...
En.attr gibt nun an, wie viel Information (im Mittel) alle Attribute über die Klassenzugehörigkeit geben.
Wenn die Anzahl der relevanten Attributen im Datensatz nun größer als EN.attr ist, dann hat man eine gute Chance, einen guten Klassifikator zu lernen.

c) Wie könnte man die Güte eines bestimmten Merkmals in Bezug auf die Klassifikationsaufgabe beurteilen?

Die **Joint Entropy** berechnet die Entropy der Kombination bestimmter Attribute. Wenn man für jedes Attribut die Joint Entropy in Bezug auf das Zielattribut berechnet, erhält man das relative Gewicht dieses Attributes für die Klassifikationsaufgabe. Hier gilt wieder: je größer die Entropy, desto unwichtiger das Attribut für die Klassifikationsaufgabe.