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

11.05.04

Knowledge Discovery

Übungsblatt 1

Sommersemester 2004

Vorbemerkungen:

Vorlesungsfolien und Übungsblätter können Sie im Internet unter folgender Adresse einsehen:
http://www.kde.cs.uni-kassel.de/lehre/ss2004/kdd

Weitere Fragen bitte an Andreas Hotho, hotho@cs.uni-kassel.de

Aufgabe 1: Allgemeines

a) Was ist KDD und was ist insbesondere das Ziel davon?

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

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

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

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

Aufgabe 2: Überwachte vs. Unüberwachte Verfahren

a) Was sind die wesentlichen Unterschiede zwischen einem überwachten und einem unüberwachten Verfahren?

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

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

Aufgabe 3: CRISP-DM Methodologie

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

b) Was sind die wichtigsten Schritte in der Datenpräparierung?

c) Was für Probleme ergeben sich typischerweise dabei?

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

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. Bestimmen Sie dafür vorersteinmal was die Kennzahl und was die Dimensionen dieser sind.

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

c) Erweitern Sie das obige Sternschema so, daß möglichst einfach der Monatsumsatz eines bestimmten Produktes in einer bestimmten Filiale berechnet werden kann.

d) Wie würden Sie die Daten visualisieren, damit der Logistikexperte der Firma IDLA möglichst effizient die Logistikplanung für den nächsten Zeitraum vornehmen kann.

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. Als Consultant und Klassifikationsexperte werden Sie von dem Internet-Provider angeheuert, um bei der Klärung folgender Fragen zu helfen:

a) Wie gut wird die Klassifikation vermutlich sein?

b) Wie viele Attribute sollten für die Klassifikation verwendet werden?

c) Welche Attribute werden in Bezug auf die Klassifikation gut funktionieren?

Verwenden Sie ihre Kenntnisse aus der Vorlesung um eine Antwort auf diese Fragen zu geben!