WPF Data Mining praktisch – Vorbereitung DMC

Lehrveranstaltung von Profs. Drs. Wolfgang Konen & Thomas Bartz-Beielstein, WS09/10
E-mail: wolfgang.konen@fh-koeln.de, bartz „at“ gm.fh-koeln.de
Campus: Raum 3.230, Tel. -6275, -6385.

Literaturverzeichnis **Themen im Detail** **Materialien**

Informationen zum Kurs

Organisatorische Hinweise

Die Veranstaltung besteht aus Dozentenvorträgen, Übungen, die alle bearbeiten, sowie Projektarbeiten der Studierenden (jeder versucht zu lösen, Zusammenbringen der Lösungen im Team) und der Vorstellung der Ergebnisse.

1. Termin: 13.10.2009, 12.15 Uhr , Raum 3.113

Anmeldung zum WPF: bitte online über **http://advbs06.gm.fh-koeln.de:8080/WPF-Anmeldung**, falls das nicht geht, per E-Mail an die Dozenten, bei über 20 Teilnehmern entscheidet das Datum der Anmeldung.

Der Kurs ist eine ideale Vorbereitung für spätere Projekt-, Bachelor- oder Masterarbeiten oder SHK-/WHK-Tätigkeiten in den BMBF-geförderten Forschungsprojekten **SOMA** *und* **FIWA**, *die unter unserer Leitung von 2009-2012 am Campus Gummersbach durchgeführt werden!*

Zielsetzung

We are drowning in data, but starving for knowledge!
John Naisbett

Dieser mittlerweile berühmtgewordene Ausspruch von John Naisbett unterstreicht die Wichtigkeit der Themengebiete *Data Warehousing*, *Knowledge Discovery* und *Data Mining (DM)*. Gerade der letzte Bereich (DM) hat das Ziel, Strukturen oder Muster in (großen) Datenbeständen semiautomatisch zu finden und diese in Vorhersage- oder Klassifikationsmodellen für vielfältige Anwendungen nutzbar zu machen. Beispiele: Internet, Kundenabwanderung, Bedarfsprognose, Bioinformatik, Text Mining u.v.a.m.

Der Kurs *Data Mining praktisch – Vorbereitung DMC* will eine **praxisorientierte Einführung zum Data Mining** liefern für Studenten aller Informatik-Studiengänge.

Was ist DMC? – Der **Data Mining Cup (DMC)** ist ein jährlich von den Veranstaltern TU Chemnitz und prudsys AG angebotener Wettbewerb für Studenten (national und international, 240 Teilnehmer in 2009), in dem es darum geht, eine aus der Anwendung stammende Data Mining Challenge möglichst gut zu lösen.

Wir wollen mit dem Kurs *Data Mining praktisch – Vorbereitung DMC* ein ganz konkretes Ziel verfolgen: Sie als Studierende der FH Köln fit zu machen für eine mögliche Teilnahme am DMC'2010. Dabei werden Sie natürlich auch jede Menge über praxisorientiertes Data Mining und Optimierung lernen, was auch später im Beruf nützlich sein kann.

(Der DMC'2010 wird voraussichtlich von April-Juni 2010 laufen, also nicht mehr in diesem WS. Das ist aber auch gut so, damit Sie ausreichenden Vorlauf zur Vorbereitung haben. Für Interessenten kann aber die Durchführung im Sommersemster'2010 im Rahmen eines Projektes Sinn machen; sprechen Sie uns hierzu an)

Im WPF wird über weite Strecken ein „Learning by Doing"-Ansatz gewählt: Nach einer kurzen Einführung in grundlegende Begriffe des DM, den **CRISP-DM-Prozess** (Entwicklungsprozess im Data Mining) und einem Tutorial für die Statistik- und Modellierungssprache **R**, begeben wir uns direkt in die praktische Datenanalyse und den Aufbau von ersten **DM-Modellen**. Aus den typischen Schwierigkeiten heraus, die sich am Anfang stellen, werden wir unser Wissen schrittweise verfeinern: wichtige Aspekte der **Datenvorverarbeitung** und der **Evaluierung** von Modellen werden behandelt. Wir lernen (einige) „state-of-the-art" Modelle und **Meta-Modelle** des Data Mining kennen. Jedes Modell hat (einige oder viele) Parameter, deren optimale Einstellung durch geeignetes **Experiment Design** ein wichtiges Thema ist.

Mit diesem Rüstzeug sind wir dann fit für die **Haupt-Projekte**: Alle Teilnehmer entwickeln Lösungen für die DMC-Challenges vergangener Jahre, stellen ihre Ergebenisse vor, wir diskutieren sie im Team und erreichen evtl. durch Kombinationen verbesserte Resultate. Hierbei stoßen wir sicher auf diverse Probleme, aber genau durch deren Behandlung gewinnt man Kompetenz in Datenanalyse und Data Mining sowie Erfahrung bei der Einschätzung von Daten allgemein.

Zielgruppe

Der Kurs wendet sich an alle, die etwas über die Methoden des Data Mining und der Computational Intelligence erfahren wollen. Besonders relevant für Studierende der Studiengänge der Wirtschaftsinformatik (WI), aber auch für Allgemeine Informatik (AI), Technische Informatik (TI) und Medieninformatik (MI).

Voraussetzungen

Gute Mathematikkenntnisse. Für ein tiefergehendes Verständnis der mathematischen Zusammenhänge sind grundlegende Kenntnisse in der Statistik hilfreich.

Optional: Grundkenntnisse in **R**, hier wird jedoch auch im Kurs eine kurze Einführung geboten. (Für jemanden, der schon einmal in **MATLAB** gearbeitet hat, wird der Einstieg auch leichter fallen, da beide Umgebungen gewisse Parallelen aufweisen)

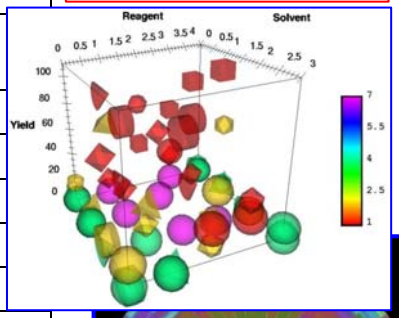
Ausbildungsziele

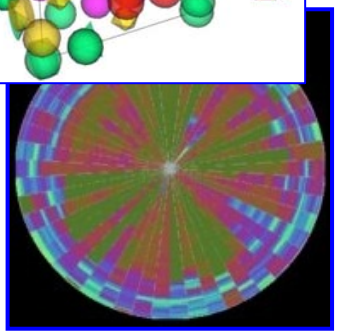
1. Verständnis der grundlegenden Konzepte Data Mining (s. Lerninhalte)
2. Überblick zu Toolboxen und Programmen des Data Mining
3. Kompetenz im Umgang mit **R** als *einem* Standardwerkzeug zur Datenanalyse
4. Die Studierenden sind in der Lage, komplexe Aufgabenstellungen des Data Mining zu bearbeiten sowie eigene Modelle hierzu zielorientiert zu entwickeln

Lerninhalte

(V: Vorlesung, Ü: Übung, S: Seminar, P: Projektarbeit, Änderungen möglich, für aktuelle / genauere Angaben s. **Themen im Detail**)

V	Einführung DM, CRISP-DM, Datenvorverarbeitung
V / Ü	Datenvorverarbeitung
V / Ü	**R** Tutorial und **R** Übungen
V	Modellierung, Entscheidungsbäume, Meta-Lerner
Ü / S	Random Forests – Ideen, Begriffe, Fragen
V	Evaluierung: Trainings-/Test-Set, Overfitting, Kostenmatrix, Fehlermaße
S/P	Projekt DMC 2007
S/P	Projekt DMC 2006
S/P	evtl. weitere DMC-Projekte
V / Ü	Optimierung und Experiment Design
V / Ü	Fehlende Werte – Replacement-Strategien
S/P	Vergleichende Umsetzung mit **RapidMiner**
Ergänzende Themen	
V / Ü	Naive Bayes – Algorithmus u. Implementierung in **R**





Die Veranstaltung umfasst

1. Vorträge und Demonstrationen
2. Übungen am Computer
3. Projektarbeiten und Referate
4. Diskussionen

Umfang / Teilnehmerbegrenzung

4 SWS / max. 20 Studenten

Medien

Overhead-Projektor. Wandtafel. Demonstrationsrechner. Rechnerarbeitsplätze für die Teilnehmer. Programme: **R** und **Tinn-R**, evtl. auch **MATLAB**.

Kommentiertes Literaturverzeichnis

Themen im Detail

Materialien