WPF Data Mining praktisch – Vorbereitung DMC

Kommentiertes Literaturverzeichnis – Stand 11/2009

Data Mining & CI

- **[Kramer09]** Oliver Kramer: *Computational Intelligence*, Springer, 2009. **Sehr gute Kurzdarstellung** verschiedener Themengebiete in der CI, als Einstieg geeignet.
- **[Witten&Frank01]** Ian H. Witten, Eibe Frank: ***Data Mining***, Hanser (2001). S. 238-245. Das Buch zum **WEKA-System**. Allgemeine Einführung in Data Mining, Schwerpunkt Deci-Trees + Klassifikation. Wenige Formeln, dafür textuell gute Beschreibung. **Sehr gut verständliche Erklärung** der Grundzüge von Bagging & Boosting.
- **[Witten&Frank05, 2. Auflage]**: mehr zu WEKA, einige Neuerungen (>> 2 Ex. bestellt für Lib GM, noch zu lesen!)
- [Brierley07] P. Brierley: *Committee of Experts*, Tech.Report Tiberius Data Mining, **http://www.tiberius.biz/pakdd07.html**
- **[Mierswa+06]** Mierswa, Ingo and Wurst, Michael and Klinkenberg, Ralf and Scholz, Martin and Euler, Timm: ***YALE**: Rapid Prototyping for Complex Data Mining Tasks*, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006. – Basisreferenz zu **RapidMiner (früher YALE)**
- **[Hastie, Tibshirani, Friedeman 2001]** *The Elements of Statistical Learning*. Springer. Insbesondere Kap. 10 (1st ed.) bzw. Kap. 15 (2nd ed.) über Random Forests ist hier interessant. **NEU 2. Auflage 2008**: Das auch insgesamt sehr lesenswerte Buch ist unter **http://www-stat.stanford.edu/~tibs/ElemStatLearn/** als Free-PDF-Download erhältlich, mit verschiedenen Ergänzungen wie Benchmark-Datensätzen u.a.m.
- [Morik&Joachims01] Katharina Morik, Thorsten Joachims: ***Data Mining*** in: G. Görz et al.: *Handbuch Künstliche Intelligenz*, Oldenbourg Verlag, 2001. Guter Überblick DM, darin sehr gute Kurzeinführung in Verstärkungslernen und in Subgruppenentdeckung.
- [Faeskorn+07] Heide Faeskorn-Woyke, Birgit Bertelsmeier, Petra Riemer, Elena Bauer: **Datenbanksysteme** - Theorie und Praxis mit SQL2003, Oracle und MySQL, Pearson Studium, 2007. Companion-Website: **www.pearson-studium.de**, Schnellsuche Buchnr. 7266.
- [Hornick+07] Mark F. Hornick, Erik Marcadé and Sunil Venkayala. ***Java Data Mining: Strategy, Standard and Practice***, Morgan Kaufmann, **Elsevier**; 2007. Beschreibt mehr eine SW-Architektur, nur kurz die Data Mining Methoden. Leider nicht sehr tiefgehend: Z.B. gibt es drei Kapitelüberschriften "Attribute Importance", an jeder wird weitschweifig erklärt, was man darunter verstehen soll, aber an keiner Stelle wird auch nur ein Ansatz gegeben, wie man es praktisch macht.
- [Torge03] Luis Torge: *Data Mining with R: learning by case studies*.
  Das Buch ist ganz gut zu lesen als Einführung in R und Data Mining. Es ist als Free PDF Download unter **http://www.liaad.up.pt/~ltorgo/DataMiningWithR/** verfügbar. Es scheint allerdings ein noch nicht ganz fertiger Entwurf zu sein (aber immerhin 140 Seiten!)

Random Forests

- **[Breiman01]** Leo Breiman, ***Random forests***, Machine Learning 45 (1) 5-32, 2001. – *Die* Haupt-Veröffentlichung zum Thema, zum Zitieren in wiss. Arbeiten, allerdings

recht mathematisch und für's pragmatische Arbeiten etwas weniger geeignet. **http://oz.berkeley.edu/users/breiman/randomforest2001.pdf**.

- [Breiman02] Leo Breiman, ***Manual On Setting Up, Using, And Understanding Random Forests V3.1,*** **http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf**. [Breiman02b] Leo Breiman, ***Manual On Setting Up, Using, And Understanding Random Forests V4.0,*** **http://oz.berkeley.edu/users/breiman/Using_random_forests_V4.0.pdf**. Manuals zum Fortran-Code, aber auch lesenswert wg. Beschreibung wie RF arbeitet und der Data-Mining-Beispiele. Das V3.1-Manual enthält ein paar mehr Erklärung zur Variable-Importance, die in V4.0 leider fehlen. Dafür enthält V4.0 neue Features zu Missing Values.
- [BreimanCutler] **http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm**: Die "description"-Seite auf dem "classification/clustering"-Teil der Random-Forest-Page. Bringt eigentlich **die beste Einführung** zum Thema für den Praktiker. Zum Teil Überlapp mit dem obigen Manual, z.T. aber auch eine griffigere und umfassendere Darstellung der Konzepte.
- **[Breiman03]** Leo Breiman, **RF / tools – A Class of Two-eyed Algorithms**, Talk at SIAM Workshop, May 2003, **http://www.cs.cmu.edu/~lafferty/ml-stat/breiman.pdf**
- **[UC-Berkeley05]** Nachruf auf Leo Breiman, der am 05.07.05 starb. **http://www.berkeley.edu/news/media/releases/2005/07/07_breiman.shtml**

Software und Datensätze

Ein paar (kostenfreie) Software-Angebote zu Data Mining. Es gibt eine fast unübersehbare Fülle von Data Mining Tools, für einen umfassenden Überblick siehe **http://www.kdnuggets.com** (eine auch ansonsten recht interessante Website für Neues aus dem Bereich Knowledge Discovery & Data Mining)

- [WEKA] **http://www.cs.waikato.ac.nz/~ml/weka/**. Eine der ersten Data Mining Suites in Java. [Witten&Frank01 & 05] haben die Popularität von WEKA begünstigt.
- [RapidMiner] **http://rapid-i.com/**. An der Uni Dortmund entstandene Data Mining Suite, hieß früher YALE. Hat alles von WEKA integriert.
- [R] **http://www.r-project.org/**: The R-Project for Statistical Computing
- [Tinn-R] **http://www.sciviews.org/Tinn-R/**: komfortabler Entwicklungs-Editor für R
- [**Tiberius] http://www.tiberius.biz/** – sollte free sein für Academia, noch nicht getestet
- **http://www.cs.waikato.ac.nz/~ml/weka/index_datasets.html**: Hier finden sich alle Datasets, die in [**Witten&Frank05**] benutzt werden (und viele mehr). Meist im ARFF-Format, läßt sich aber leicht konvertieren.
- **http://www.data-mining-cup.de/** Hier finden sich die Datensätze der DMC Challenges aus den Jahren 2000-2007

R-Tutorials

- **http://lectures.molgen.mpg.de/Genexpression_WS0506/material/einfuehrungR.pdf**: Kurzeinführung (5 Seiten, also sehr kurz) im Rahmen eines Kurses „Statistik für Bioinformatiker" an der FU Berlin.
- [Klar06]: Bernhard Klar, Datenanalyse + Graphik mit R, **http://www.mathematik.uni-karlsruhe.de/stoch/lehre/statprak2006s/media/biopraktikum.pdf**: Ausführliches Tutorial (140 Seiten!) im Rahmen eines Kurses „Statistik für Biologen" an der Uni

Karlsruhe. Für eine (sehr gute!) Einführung in R reichen aber erstmal auch die ersten 36 Seiten (Kap. 1-3).

- **<R-Verzeichnis>/doc/manual/R-intro.html** oder **R-intro.pdf** : "An Introduction to R"
- **<R-Verzeichnis>/doc/manual/refman.pdf**: “R Reference Manual”
- [**Torge03**]

Info zum WPF DMC

Themen im Detail