Data Mining, Human Language Technology, and Pattern Recognition

Thomas Deselaers, Arne Mauser
lastname@cs.rwth-aachen.de

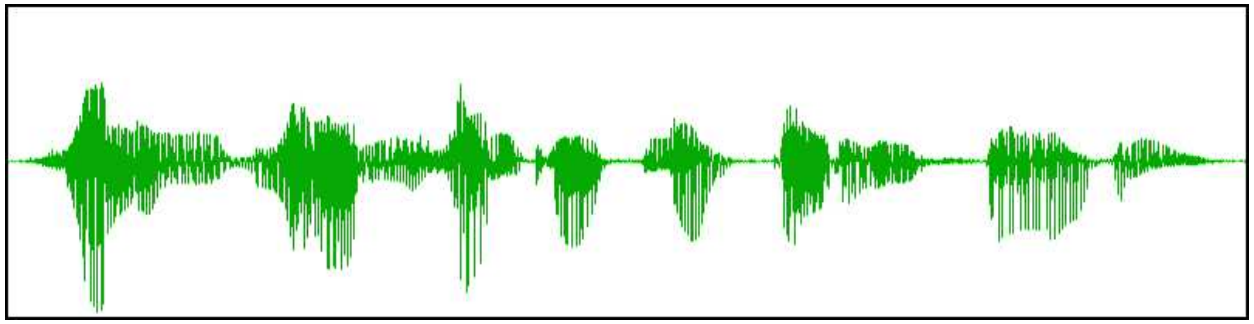
TDWI Roundtable Rheinland – 31.03.2008

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Outline

- Human Language Technology and Pattern Recognition Group
- Data Mining Cup 2004 - 2007 and beyond
- Preprocessing
- Classifiers
- Model Selection
- Summary and Outlook

Speech Recognition and Pattern Recognition

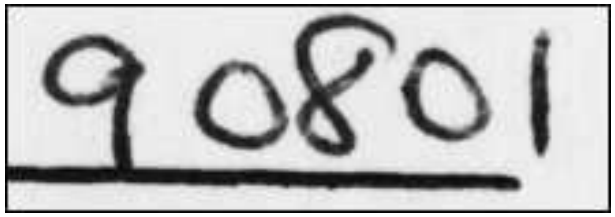


↓

Speech Recognition System

↓

`Sollen wir am Sonntag nach Berlin fahren?`



↓

OCR System

↓

`9 0 8 0 1`

Typical Tasks:

- **Speech Recognition (signal ⇒ text)**
- **Machine Translation (z.B. chinese text ⇒ english text)**
- **Language Understanding (text ⇒ category)**
- **Optical Character Recognition (hand writing ⇒ text)**
- **Image/Object Recognition (image ⇒ class)**

signals and unstructure data (audio, video, hand writing, text, ...)

↓

discrete symbols

Speech Recognition and Pattern Recognition

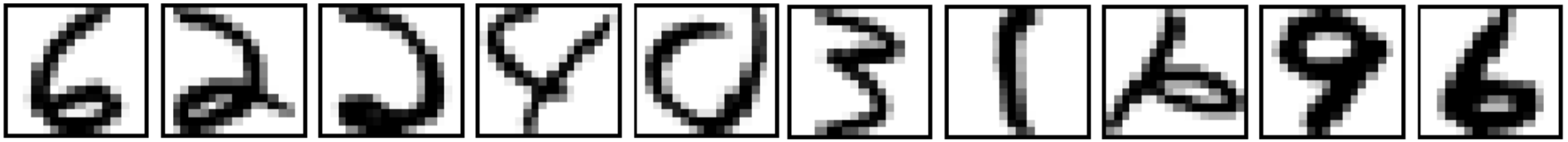
Tasks and Methods:

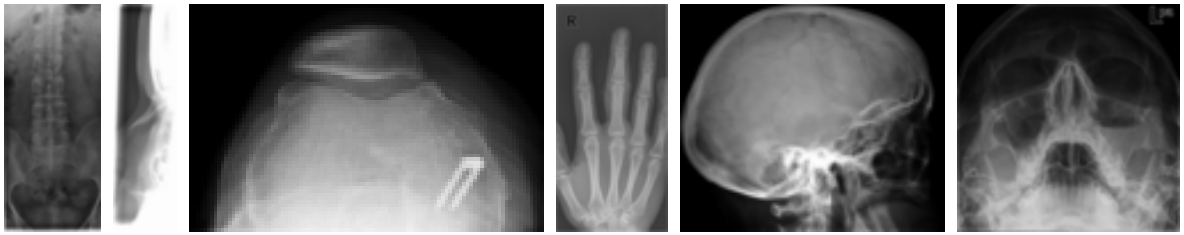
- **How to learn from data automatically?**
  $\Rightarrow$statistical modelling
- **How to analyse large amounts of data fully automatically?**
  ($>$1000 hours audio/$>$100M running words)
  $\Rightarrow$efficient algorithms
- **How to model series of discrete symbols?**
  $\Rightarrow$formal languages

Relevant disciplines:

- **probabilistic models:**
  pattern recognition, statistics,
  neural networks, data mining
- **information theory**
- **efficient algorithms**
- **digital signal processing and numerics**
- **formal languages and grammars**
- **artificial intelligence**

Image Recognition

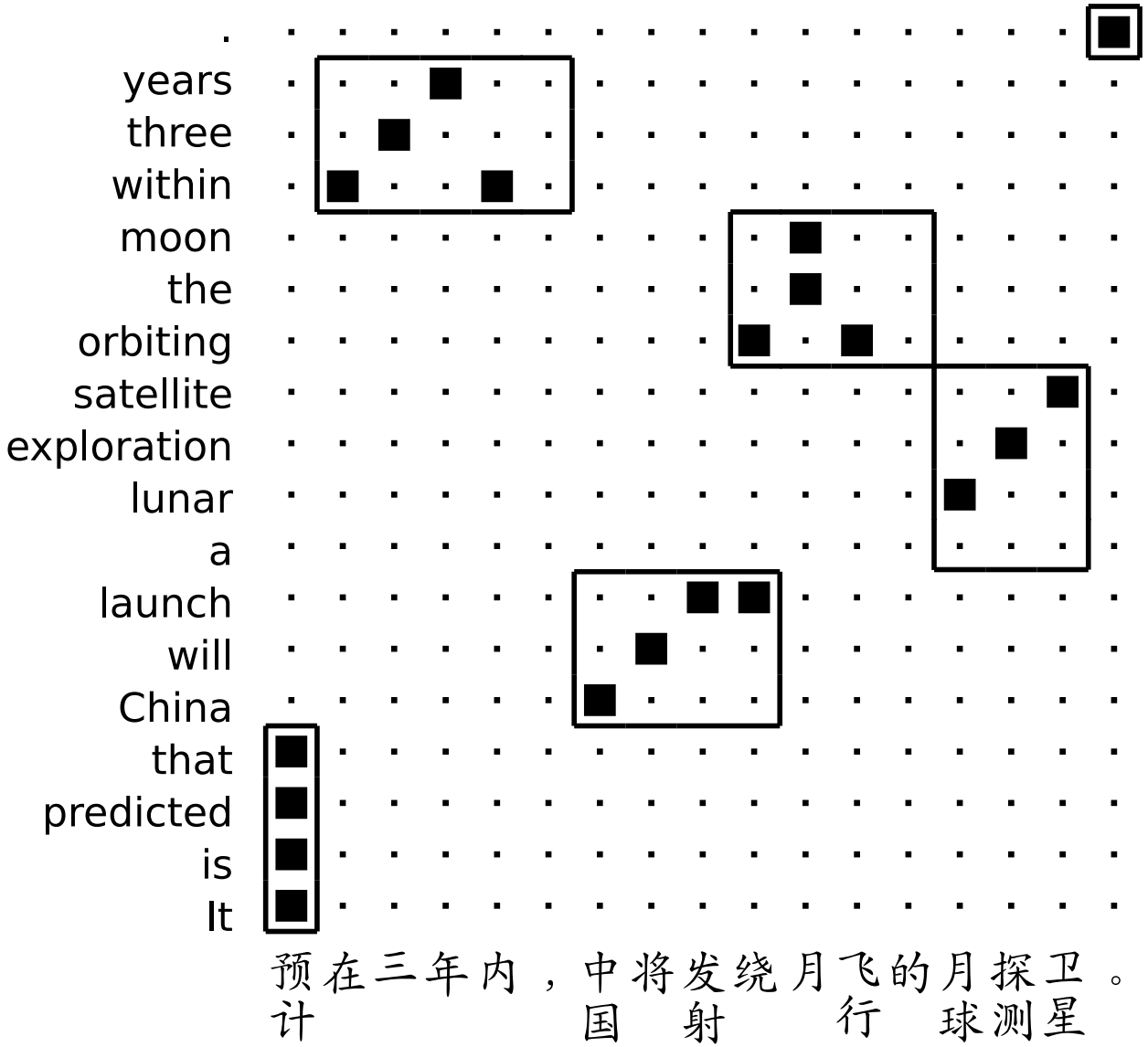




Content-based Image Retrieval

http://www-i6.informatik.rwth-aachen.de/~deselaers/fire.html

Statistical Machine Translation



Data Mining

- **Data Mining Cup is a competition for students**
- **We are experts in natural language processing, speech recognition, and computer vision**
- **Application of *our* methods to *Data Mining*?**
  - **Data Mining Cup 2004:**
    3 students of i6 participated and obtained ranks 1,3, and 5
  - **Data Mining Cup 2005:**
    9 particpants from our lab,
    1st rank & all other students amoung the first 20 (from 147 submissions)
  - **GfKl Data Mining Competition in 2005:**
    5 submission among the top 10 (from 40 submissions)
  - **Data Mining Cup 2006:**
    11 participants from our lab, most among the top third submissions
  - **Data Mining Cup 2007:**
    9 participants from our lab, ranks 1,2,4,5,7,8 from 230 submissions
    (all our submissions among top 20)

Data Mining Cup 2004

- prediction of returning behavior of mail order customers
- given a set of approx. 20,000 classified training records
- and approx. 20,000 unclassified test records
- classify into one of 3 classes: low returners, high returners, indefinite
- given features e.g.:
  - ordering and returning behavior in different time spans
    (e.g. number of orders, number of returns in time span A)
  - statistical information about customer
    (e.g. age)
  - information on geographical environment of customer
    (e.g. purchasing power per citizen in ZIP-code area)

Data Mining Cup 2005

- prediction whether a person ordering in an online-shop will pay or not
- 30,000 trainnig records
- 20,000 unclassified test records
- classify into one of 3 classes: will pay/won't pay/indefinite
- given feature e.g.:
  - paying by creditcard, Nachname, ...
  - last order was payed
  - article numbers of ordered articles
  - ...

Data Mining Cup 2006

- predict whether iPods will be sold above or below average price in their category on eBay
- 8,000 training auctions
- 8,000 unclassified test auctions
- classify into 2 classes: above/below average price
- given features e.g.:
  - category
  - start price
  - textual description
  - feedback scores of seller
  - ...

Data Mining Cup 2007

- predict which type of rebate coupons are appropriate for which customer
- 50,000 training customer records
- 50,000 unclassifierd test customer records
- classify into 3 classes: rebate coupon type A, B, or none
- given features e.g.:
  - rebate coupons that were used by the particular customer

Approach

4 stages:

- data preprocessing
- feature selection
- classification
- combination of classifiers

Additional steps:

- visualization
- clustering
- regression vs. classification
- data reduction (e.g. PCA, LDA)
- ...

Preprocessing

Problem:

- different variables of data have different ranges
- missing values
- outliers
- *insane distributions*
- noisy values

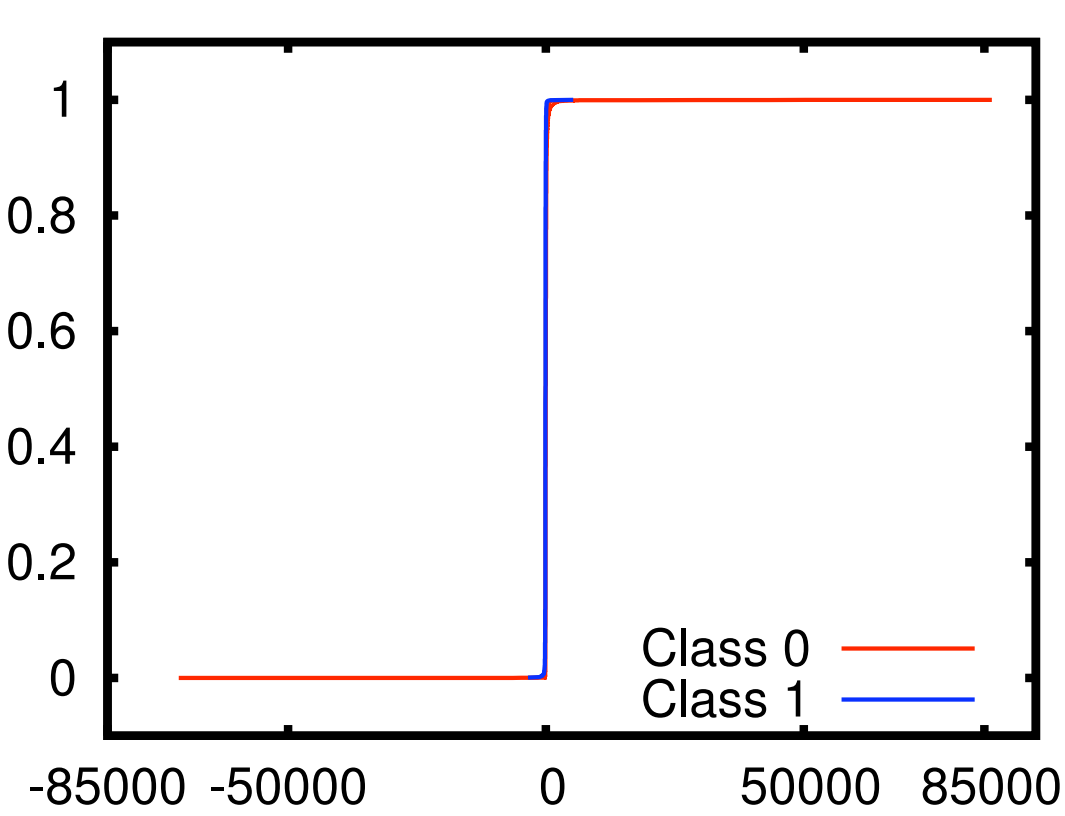
$\Rightarrow$**strong impact on some classifiers**

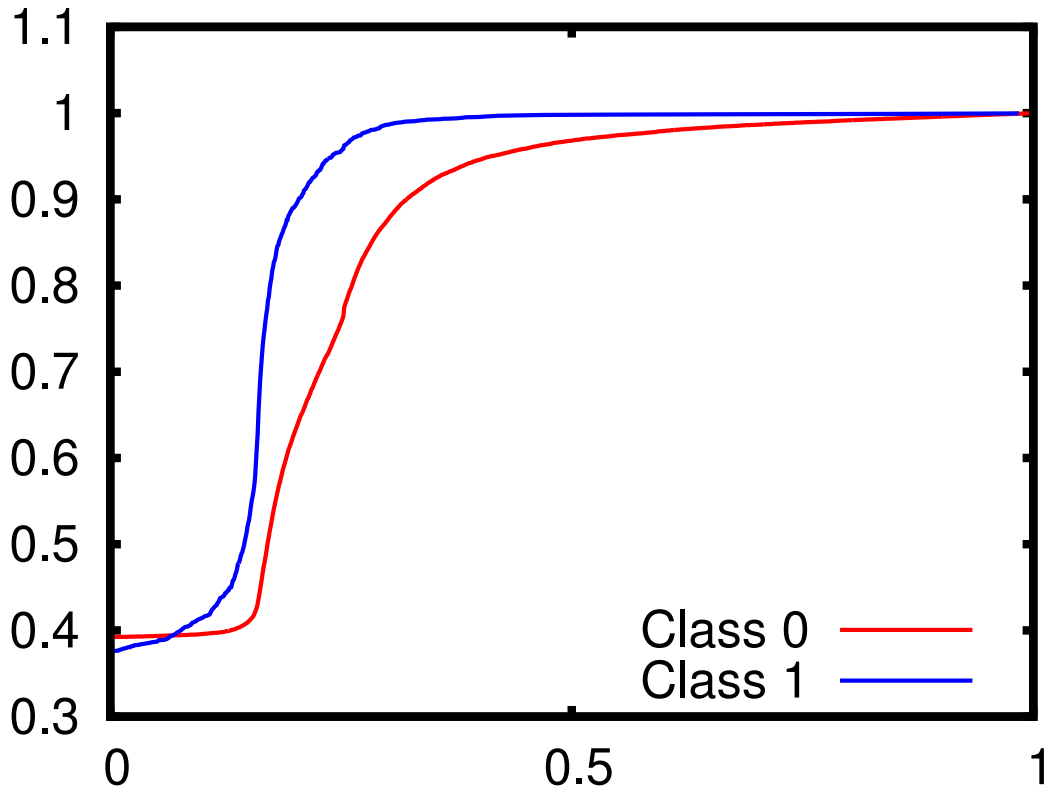
Solution: preprocessing of data: e.g.

- linear scaling to common ranges
- replace values by their quantiles
- histogramization
- add binary features
- classifier stacking
- ...

Order preserving transformation

- contains a bin for each unique feature value
- feature values are replaced with the [0..1]-normalized index of their bins





result: normalization of distances between neighboring feature values

Binary features

Idea: Generalizing or emphasizing particularities

- missing values
- special values (e.g. zero)
- represent categorial values that are expressed numerically

Feature Selection

Problem:

- Even after preprocessing some features may be good for classification, others not.

Idea: find out which features are suited to the task at hand.

- forward selection
- examination of feature correlation
- weighted combination of features
- ...

Classification Methods

various classification methods available:

- **neural nets**
- **nearest neighbor/kernel densities**
- **support vector machines**
- **Gaussian distributions**
- **naive Bayes**
- **decision trees**
- **boosting**
- **logistic regression**
- **...**

Combination of Classifiers

Problem:

- different classifiers make different errors and have different advantages.

Idea: combining classifiers may lead to improved performance

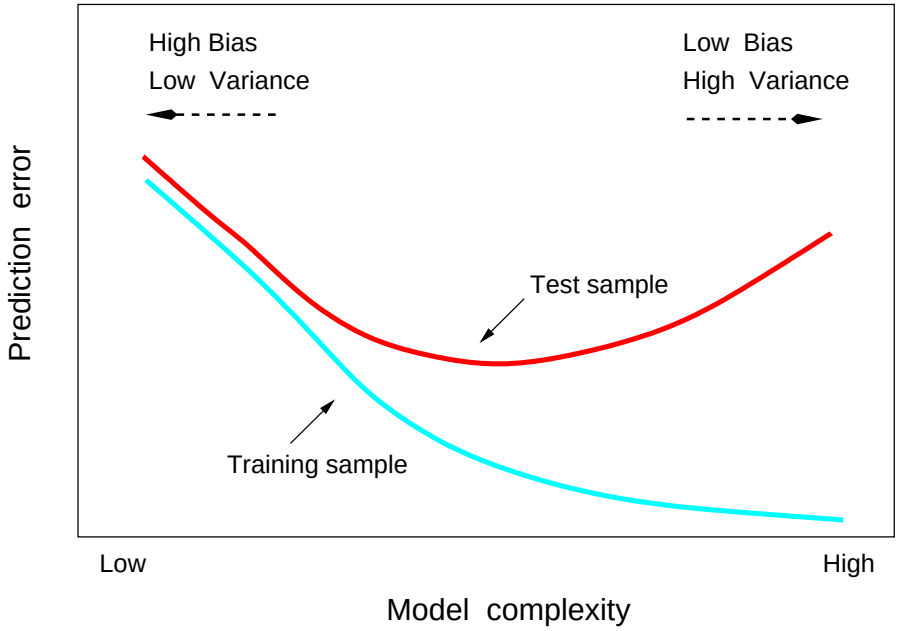
- sum rule
- product rule
- maximum rule
- stacking
- boosting
- ...

Selecting the “right” methods

Problem: How to select preprocessing, classification, and ... methods?

not an easy task:

- training on training data, measuring errors on training data $\Rightarrow$ no errors for nearest neighbor classifier, but not always the best classifier.
- not possible to count errors on test data, because correct classification not known
- other problem: over-fitting



Problem

free parameters:

- **classifier**
- **preprocessing**
- **feature combination**
- **classifier combination**
- **parameters to classifier**

problem: model assessment and selection

- **how to select the best for each of the above parameters?**
- **how good generalizes a considered model to unseen test data?**

Splitting the Data

given situation:

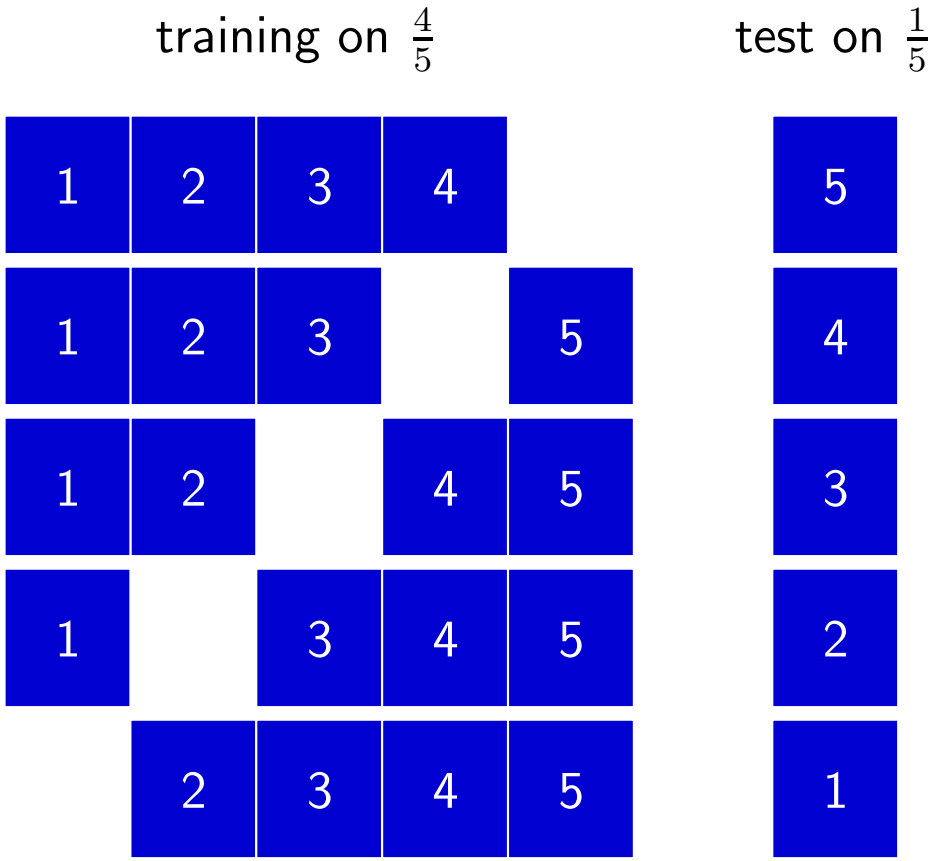
take some data from the training data away:

perform experiments with 5-fold cross validation on remaining training data:

Cross-Validation

assessing classifier performance:

- **cross-validation on reduced training data:**



aim:

- **determine some “good” setups**

Evaluation on Validation data

- given the set of “good” setups
- evaluate these on the so-far unseen training data:

Result:

- performance measure (here: recall) on the validation data
- average these performance measures with the according measures from cross-validation experiments to select the “best” method

Classification of the Testdata

from our current setup

go back to the initial setup

and classify the test data using all training data.

Resources

- **weka –** `http://www.cs.waikato.ac.nz/~ml/weka/`
- **libsvm –** `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
- **Netlab –** `http://www.ncrg.aston.ac.uk/netlab/`
- **R –** `http://www.r-project.org`
- **libsvmtl –** `http://lmb.informatik.uni-freiburg.de/lmbsoft/libsvmtl/`
- **Maximum Entropy Toolkit – (available from i6)**
- **...**

Thank you for your attention

Thomas Deselaers

deselaers@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de/dmclab

Publications

C. Buck, T. Gass, A. Hanning, J. Hosang, S. Jonas, J.T. Peter, P. Steingrube, H. Ziegeldorf
Data Mining Cup 2007: Vorhersage des Einlöseverhaltens.
Informatik Spektrum, Springer, 2008