Classification of Multilingual Mathematical Papers in DML-CZ

Petr Sojka, Radim Řehůřek

NLPlab, Masaryk University, Brno
Czech Republic

December 15, 2007

Outline

1 Goals, motivation

2 Similarity search

3 MSC classification

Goals, motivation

- DML-CZ project
- Finding similar articles
  - based on metadata (citations, fixed taxonomy – MSC)
  - based on fulltext similarity
- Automated MSC classification
  - what is MSC?
  - why automated?
- Plus incorporating tools into DML workflow
  - The less research, the better

Similarity search

- Requirement to offer similar articles to the user
  - based on same MSC, fulltext
- Issues with MSC
  - one article may belong to more categories: use of secondary MSCs
  - certain level of arbitrariness wrt. person classifying
  - plus MSC is evolving: versions incompatibility, topic drift
  - test on compactness (we used only primary MSC codes)
- Fulltext issues
  - with OCR, errors already at the character level; deep, fine analysis problematic
  - $\rightarrow$ simple (even stupid, but robust) IR techniques
- We used Vector Space Model with TFIDF and LSI

TFIDF vs LSI

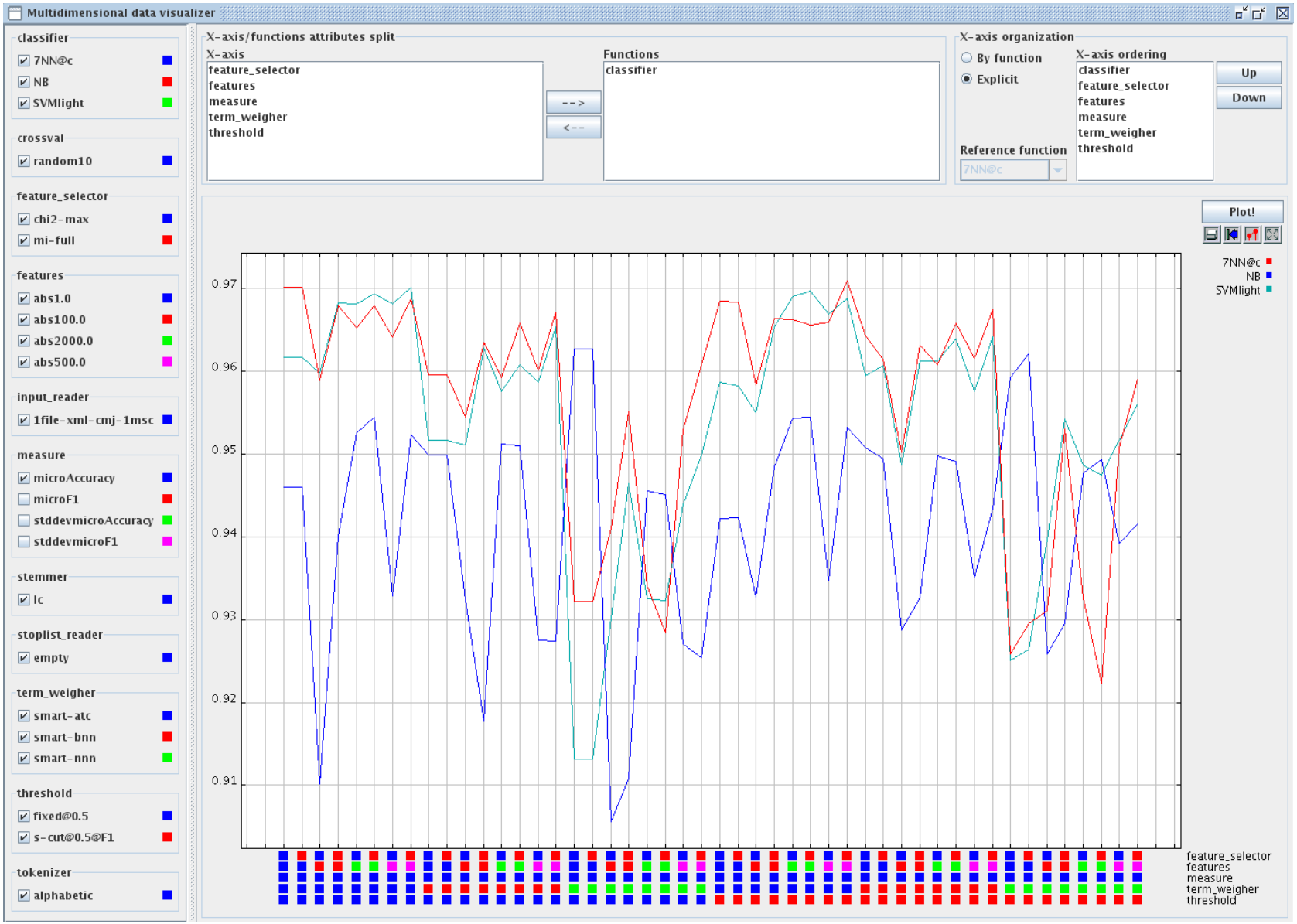
- LSI purely statistical method of topic extraction
  - topic = linear combination of terms
- Results of both look ok to layman's eye
  - but a real user would evaluate this better!
- Advantage of LSI
  - smaller, more compact matrices, simpler to work with
- Disadvantages of LSI
  - concepts not interpretable
    - hard to assign "labels" in natural language to concepts
  - resource demanding for extra large matrices
    - extra large=doesn't fit in RAM, order of 10^8 non-zero elements in TFIDF matrix

MSC classification

- Assignment of one or more MSC codes to a math article
- Questions
  - what is input?
    - possibilities abstract—fulltext—citations
    - what combination gives best performance/cost ratio?
  - internal parameters of the system
    - use of GVP know-how and code
  - what is output?
    - top MSC (first two MSC digits)
    - for example category 18-xx definition:
      Category theory; homological algebra (For commutative rings see 13Dxx, for associative rings 16Exx, for groups 20Jxx, for topological groups and related structures 57Txx ; see also 55Nxx and 55Uxx for algebraic topology)

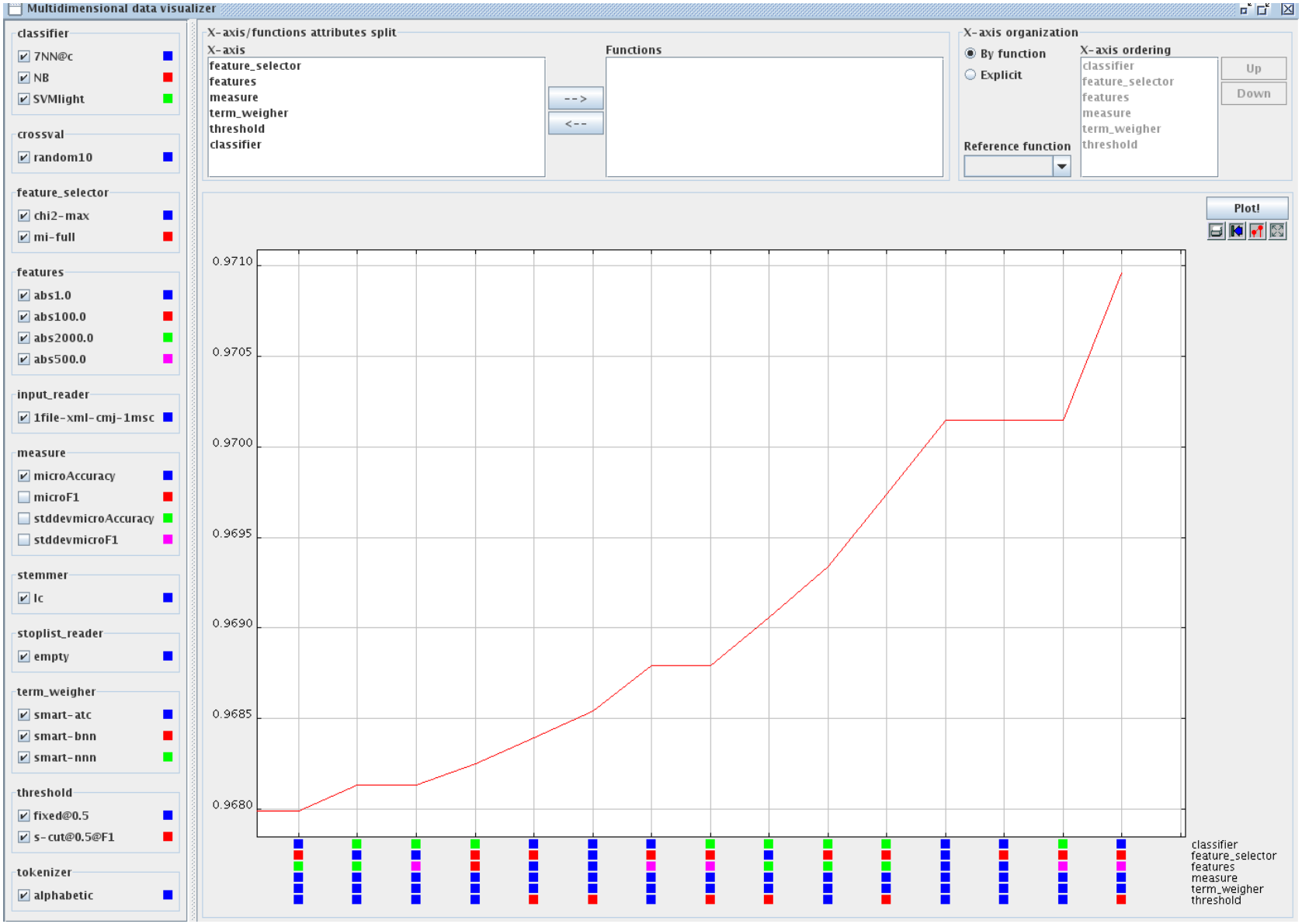
Experiments

- From CMJ fulltext.txt files
- From Archivum source files
- Use of GVP results to constrain internal parameter space
  - feature selection, term weighting, classifiers, ...
- Also use of GVP code to do the actual work
  - Java code, from tokenizing to evaluation to visualization


Multidimensional data visualizer
classifier
7NN@c
NB
SVMlight
crossval
random10
feature_selector
chi2-max
mi-full
features
abs1.0
abs100.0
abs2000.0
abs500.0
input_reader
1file-xml-cmj-1msc
measure
microAccuracy
microF1
stddevmicroAccuracy
stddevmicroF1
stemmer
lc
stoplist_reader
empty
term_weigher
smart-atc
smart-bnn
smart-nnn
threshold
fixed@0.5
s-cut@0.5@F1
tokenizer
alphabetic
X-axis/functions attributes split
X-axis
feature_selector
features
measure
term_weigher
threshold
-->
<--
Functions
classifier
X-axis organization
By function
Explicit
Reference function
7NN@c
X-axis ordering
classifier
feature_selector
features
measure
term_weigher
threshold
Up
Down
Plot!
0.97
0.96
0.95
0.94
0.93
0.92
0.91
7NN@c
NB
SVMlight
feature_selector
features
measure
term_weigher
threshold

Classifier comparison

- 10-fold crossvalidation
- *k*NN best, but does not scale well
- SVM also good
- Naive Bayes worst out of the three


Multidimensional data visualizer
classifier
7NN@c
NB
SVMlight
crossval
random10
feature_selector
chi2-max
mi-full
features
abs1.0
abs100.0
abs2000.0
abs500.0
input_reader
1file-xml-cmj-1msc
measure
microAccuracy
microF1
stddevmicroAccuracy
stddevmicroF1
stemmer
lc
stoplist_reader
empty
term_weigher
smart-atc
smart-bnn
smart-nnn
threshold
fixed@0.5
s-cut@0.5@F1
tokenizer
alphabetic
X-axis/functions attributes split
X-axis
feature_selector
features
measure
term_weigher
threshold
classifier
-->
<--
Functions
X-axis organization
By function
Explicit
Reference function
X-axis ordering
classifier
feature_selector
features
measure
term_weigher
threshold
Up
Down
Plot!
0.9710
0.9705
0.9700
0.9695
0.9690
0.9685
0.9680
classifier
feature_selector
features
measure
term_weigher
threshold

Top configurations

- *atc* best term weighter
- Mutual Information better than χ^2
- micro accuracy around 97%
  - baseline around $\frac{1}{8} = 12.5\%$
- More results in the paper

What next

- So far good results, but
  - theoretical questions
    - how will the system cope with the whole MSC taxonomy?
    - MSC 2000: 65 classes (theoretically)
    - but only 8 classes with more than 60 examples used so far
    - these experiments deferred until more data
  - practical issues
    - more data
    - incorporating automated system into DML workflow

Thank you for your attention