Classifying Medical Literature Using k-Nearest-Neighbours Algorithm

Andreas Lüschow and Christian Wartena

University of Applied Sciences and Arts, Hanover/Germany

17th European Networked Knowledge Organization Systems (NKOS) workshop, 21st September 2017

Contents

- Status Quo
  - Automatic Classification in Libraries
- Our Approach
  - Data Analysis
  - Data Preprocessing
  - Data Mining
- Results
- Discussion

Status Quo

- Increasing number of digital resources
- Many different classification systems

- 38,000 classes (DDC)
- 860,000 classes (RVK)

- Mapping between classification systems ?

Automatic Classification in Libraries

- ≠ Automatic assignment of keywords
- Often based on keywords, titles or full texts
- Use of electronic resources

- Larson (1992): 46.6 % up to 74.4 % (Classification: LCC)
- Wang (2009): 90 % (with user interactions, Classification: DDC)
- Pong et al. (2007): *kNN* better than *Naive Bayes* (Classification: LCC)

Our Approach

- Classification system: *National Library of Medicine (NLM)*
- Document representation: Assignments to other classification systems

- Goal: Use already established classifications to predict the NLM notation for a specific dataset
- Why? – Generate additional metadata for retrieval purposes

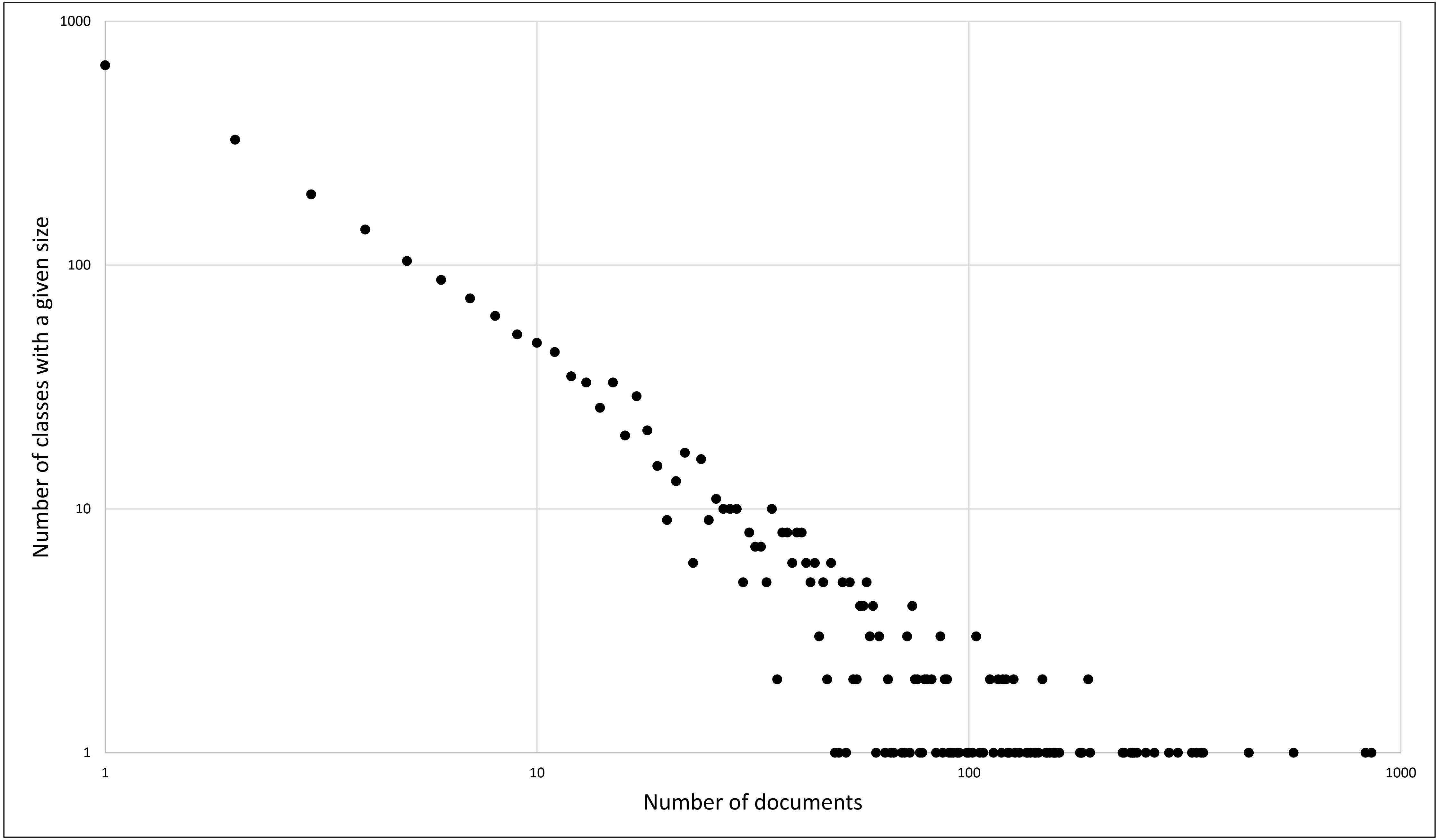
DDC	NLM	RVK	LCC	Basisklassifikation	LokaleNotation
1797			R726	'44.02'; '89.21'	W 50
306461		LC 56000	RA418	'44.06'; '44.01'	WZ
838912	WZ 330		PT2625		WZ 330
	WB 50.1		R130	'44.98'	WB 50.1
	WZ 51			'17.87'; '18.42'; '18.45'; '44.01'	WZ 51
			RC503		WA 31
				'44.02'	W 50

Data Analysis

- 45,350 datasets from the *Hanover Medical School (MHH)*
- Medical classes QS—QZ and W—WZ: 34,705 datasets
  - 2,368 different classes
  - 1,174 classes (49.6 %) with three or less assigned documents
  - 24 largest classes: 7,774 documents (22.4 %)

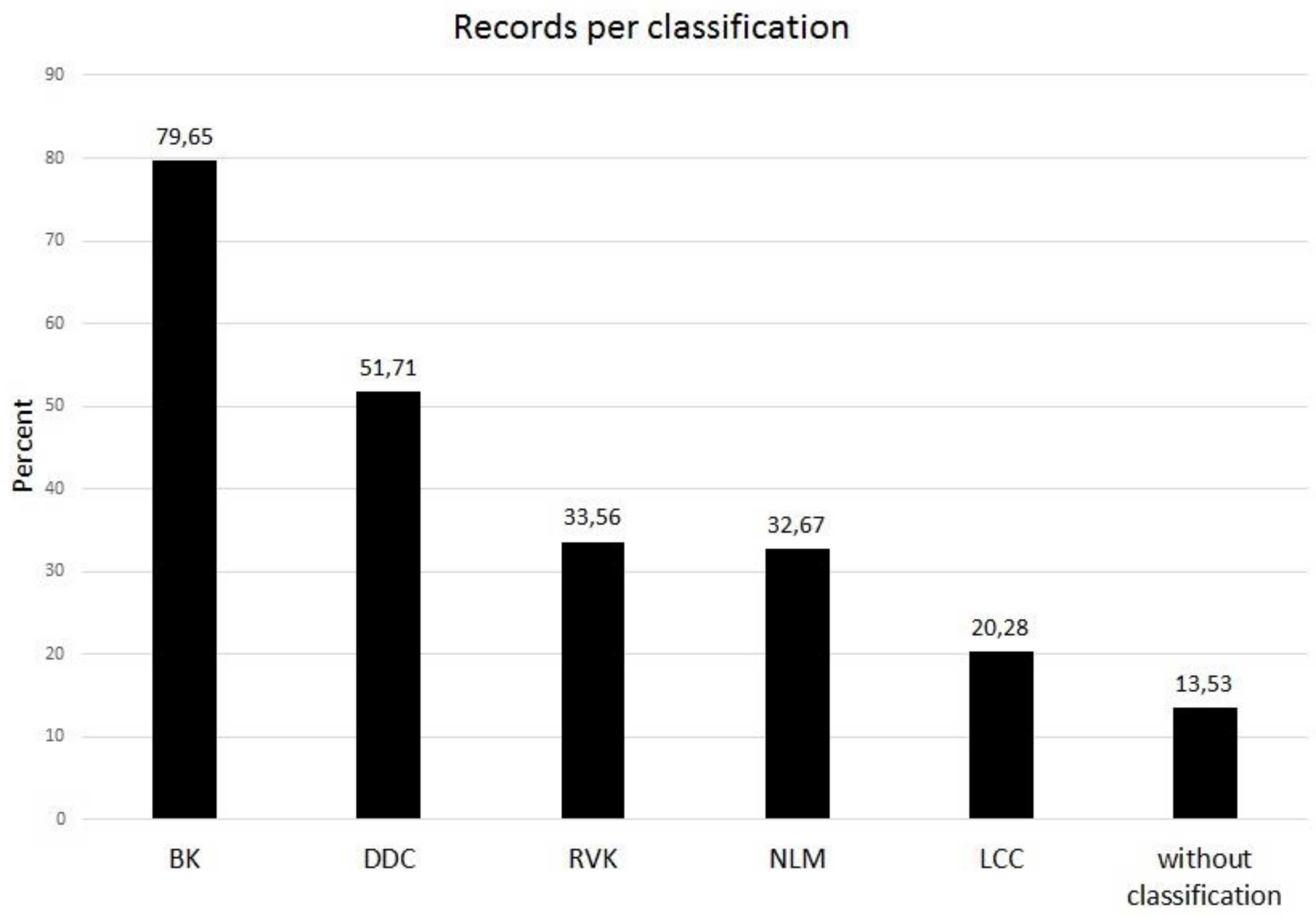
→ Skewed distribution!

→ Problematic for automatic classification


1000
100
10
1
1
10
100
1000
Number of classes with a given size
Number of documents

Data Preprocessing

- Remove datasets with no assignment to a classification system (see next slide)
- More than one assignment in a classification system → leave only the first mentioned notation in the dataset
- Exception: Convert Basisklassifikation to a vector
- Three hierarchical levels added (LN1-4, LN1-3, LNmain)



Records per classification
Percent
90
80
70
60
50
40
30
20
10
0
79,65
51,71
33,56
32,67
20,28
13,53
BK
DDC
RVK
NLM
LCC
without classification

After first preprocessing:

- 29,946 datasets, still sparse → Remove all main classes (e.g. *WB* or *QS*) and classes with less than 10 documents
- Results in: 19,348 datasets with 514 classes

	Before	During	After
Classes with 1 document	656	458	
Classes with 2 documents	323	254	
Classes with 3 documents	195	171	
Classes with max. 3 documents	1,174 (49.6 %)	883 (45.6 %)	
Classes (total)	2,368	1,935	514
Documents in classes with max. 3 documents	1,887 (5.4 %)	1,479 (4.9 %)	
Documents in the 1 % largest classes	7,774 (22.4 %)	6,290 (21.0 %)	1,646 (8.5 %)
Documents (total)	34,705	29,946	19,348

Datasets after Preprocessing

DDC	NLM	RVK	LCC	001.24	001.30	001.31	002.00	002.01	…	LNfull	LN1-4	LN1-3	LNmain
?	WB105	YT01703	?	0	0	0	0	0	…	WB105	WB10	WB1	WB
?	?	?	?	0	1	0	0	0	…	WZ100	WZ10	WZ1	WZ
610	?	?	?	0	0	0	0	0	…	WI700	WI70	WI7	WI
?	WU011	?	?	0	0	0	0	0	…	WU011	WU01	WU0	WU
61892	WL385	?	RJ001	0	0	0	0	0	…	WL385	WL38	WL3	WL

Data Mining

- WEKA
- Instance-Based Algorithm: *k-Nearest-Neighbours (kNN)*

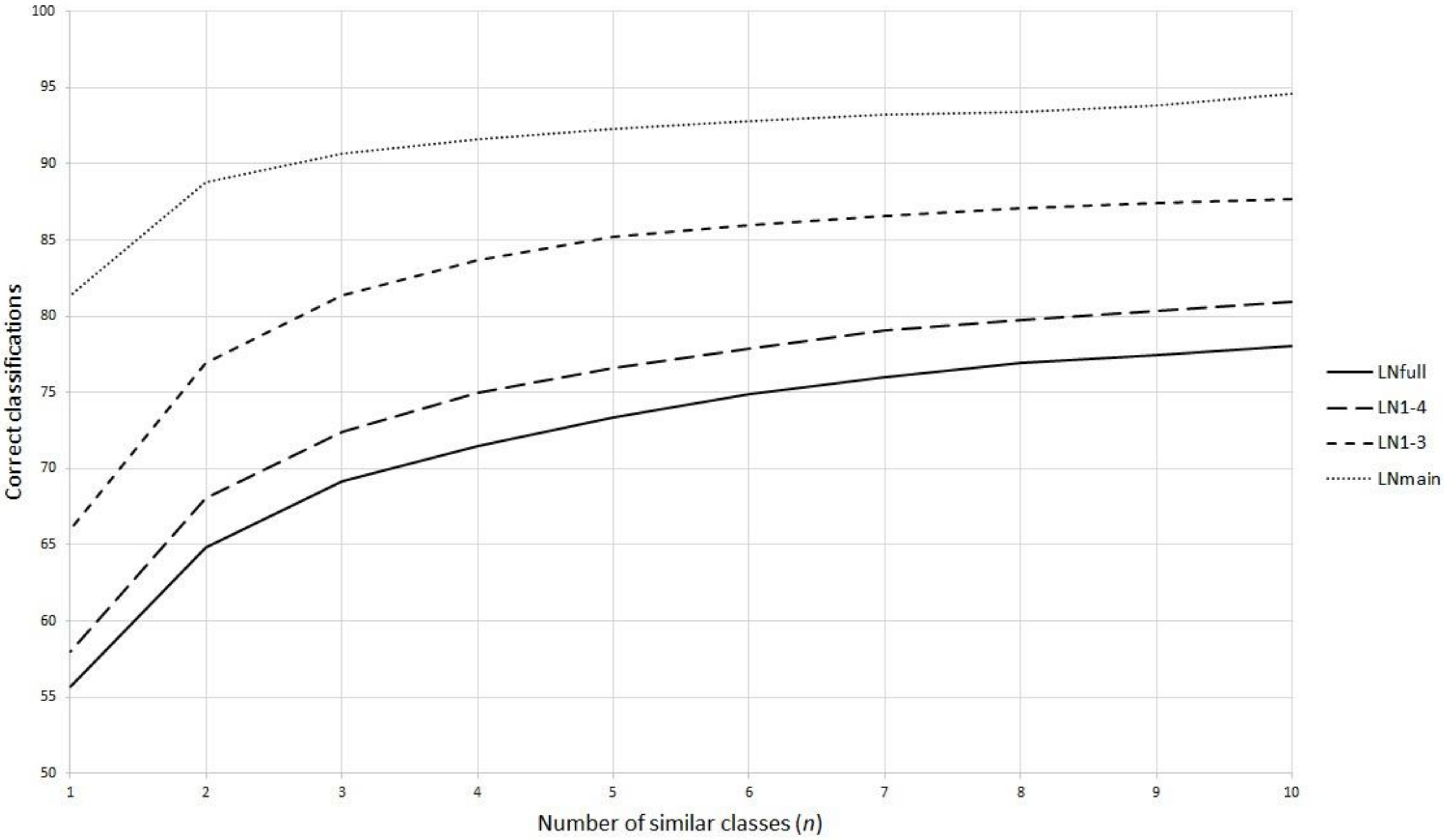
- Evaluation with ten-fold cross validation
- WEKA gives most likely class for each record → For further evaluation, we also looked at the first 10 most likely classes

Results

n	Recall LNfull	LN1-4	LN1-3	LNmain
1	55.7	58.0	66.0	81.4
2	64.9	68.1	76.9	88.8
3	69.1	72.4	81.4	90.7
4	71.5	75.0	83.6	91.6
5	73.3	76.6	85.3	92.3
8	77.0	79.8	87.0	93.4
10	78.0	81.0	87.7	94.6

Baselines

Classification	Target class LNfull	LN1-4	LN1-3	LNmain
DDC	10.6	12.2	19.8	26.6
NLM	34.5	34.9	39.7	41.4
RVK	12.1	12.9	19.9	22.5
LCC	7.3	7.9	15.2	17.8
BK	39.2	42.3	53.5	75.4
all, except NLM	44.0	47.0	57.9	77.7



100
95
90
85
80
75
70
65
60
55
50
Correct classifications
1
2
3
4
5
6
7
8
9
10
Number of similar classes (n)
LNfull
LN1-4
LN1-3
LNmain

Discussion

- Most frequent class is *W 50* (2.8 %)
- 34.5 % of the datasets are represented by only one classification system → hard for machine learning to detect differences
- Correct notation is often found in the most likely classes as determined by the algorithm

→ Semi-automatic classification could lead to good results in classification practice

Possible Optimizations

- Including more than one notation (where available)
- Definition of „similarity“
  - Our research: two different notations are completely diverse (= similarity is 0)
  - But in fact: *WN 190* is probably very similar to *WN 195* but rather different to *WC 534*!
- Using other algorithms
- Weighting attributes differently
  - The NLM attribute is more important than the others

The End

Thank you for your attention!