HAIS 2010

Rough Set-Based Analysis of Characteristic Features for ANN Classifier

Urszula Stańczyk

Institute of Informatics

Silesian University of Technology

Gliwice, POLAND

Agenda

- Stylometric Analysis
- Input Data
- ANN Classifier
- DRSA-Based Analysis for Features
- Experiments
- Conclusions

Stylometric Analysis

- Stylometric tasks
  - Author characterisation
  - Author comparison
  - Author recognition
- Textual descriptors
  - Lexical
  - Syntactic
  - Structural
  - Content specific
- Techniques employed
  - Statistics
  - Artificial intelligence

Input data

- Texts by Thomas Hardy and Henry James
  - Learning set: 180 samples

    30 parts from 3 novels per writer
  - Testing set: 80 samples

    8 parts from 5 novels per writer
- 25 Textual markers
  - 17 lexical

    (but, and, not, in, with, on, at, of, this, as, that, what, from, by, for, to, if)
  - 8 syntactic

    (fullstop, comma, question mark, exclamation mark, semicolon, colon, bracket, hyphen)

ANN Classifier

- Feed-forward Multilayer Perceptron
- Number of inputs: maximum 25, outputs: 2
- Sigmoid activation function

$$y(n) = \frac{1}{1 + e^{-\beta n}}$$

- Backpropagation algorithm

$$e(W) = \frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{I} (d_i^m - y_i^m(W))^2$$

- Multistarting approach

Experiment 1

- Classification results for the whole set of 25 features

	Classification accuracy
Worst case	**77,50%**
Average case	**81,25%**
Best case	**86,25%**

- Question: can some attributes be disregarded while maintaining the power of the classifier?
- Answer: none from stylometry

DRSA-Based Analysis for Features

- Rough Set Approach – rule-based attitude to data mining
- CRSA versus DRSA – dominance relation instead of indiscernibility
- Procedure
  - Decision table constructed
  - Concept of relative reducts employed
  - Rules for decision algorithms generated
- Selection of reducts and rules - frequency analysis for features

Reduct-Based Frequency Analysis

- 6664 relative reducts found

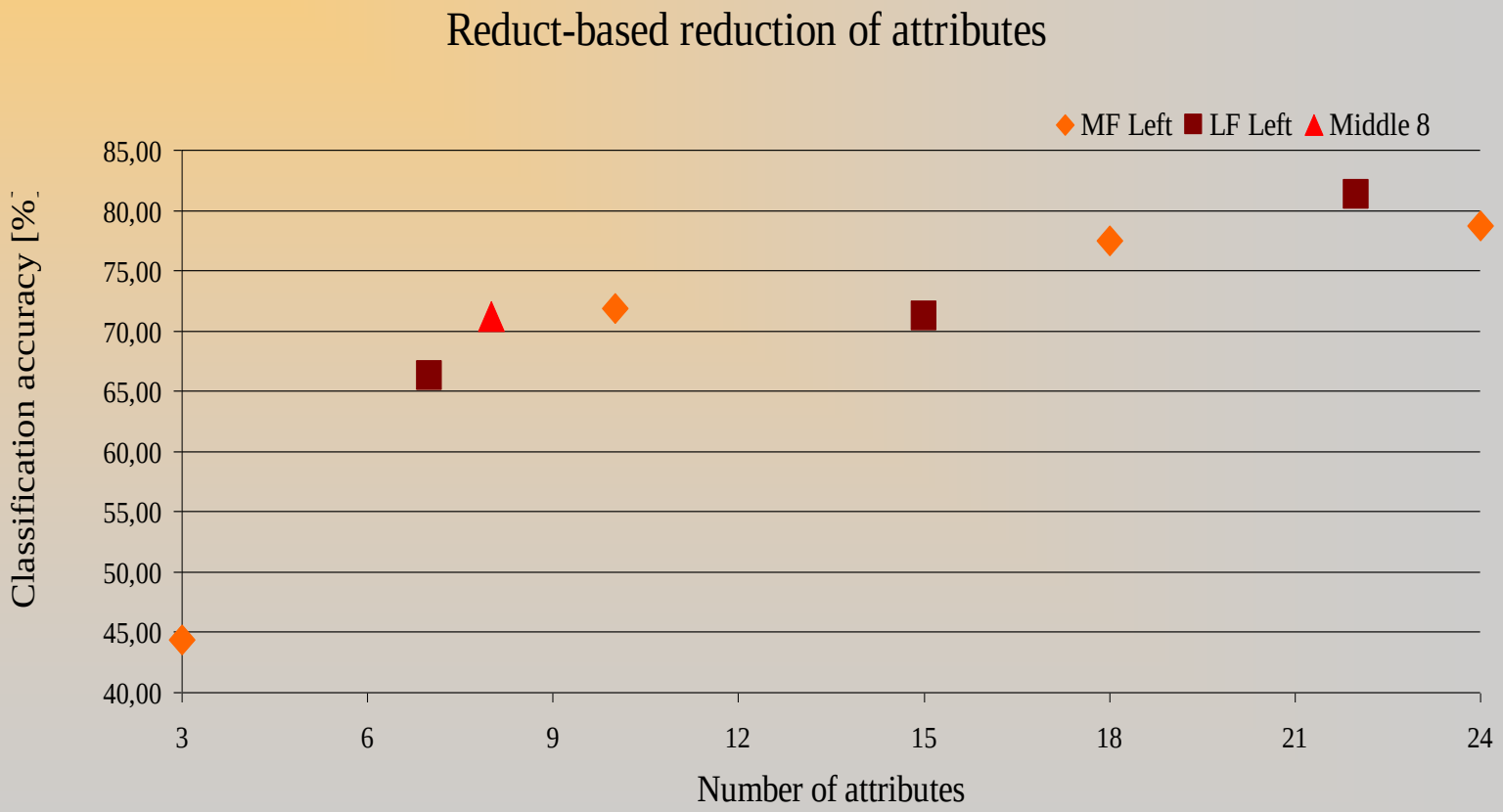
Attribute	Occ. ind.	Attribute	Occ. ind.	Attribute	Occ. ind.
of	3478	at	2585	question	1712
fullstop	3190	to	2497	for	1609
on	3083	colon	2384	if	1584
comma	2943	exclamation	2368	what	1415
not	2778	and	2324	bracket	1395
semicolon	2740	from	2273	that	1343
in	2726	with	2161	but	893
by	2648	as	2108		
this	2585	hyphen	2035		

Rule-Based Frequency Analysis

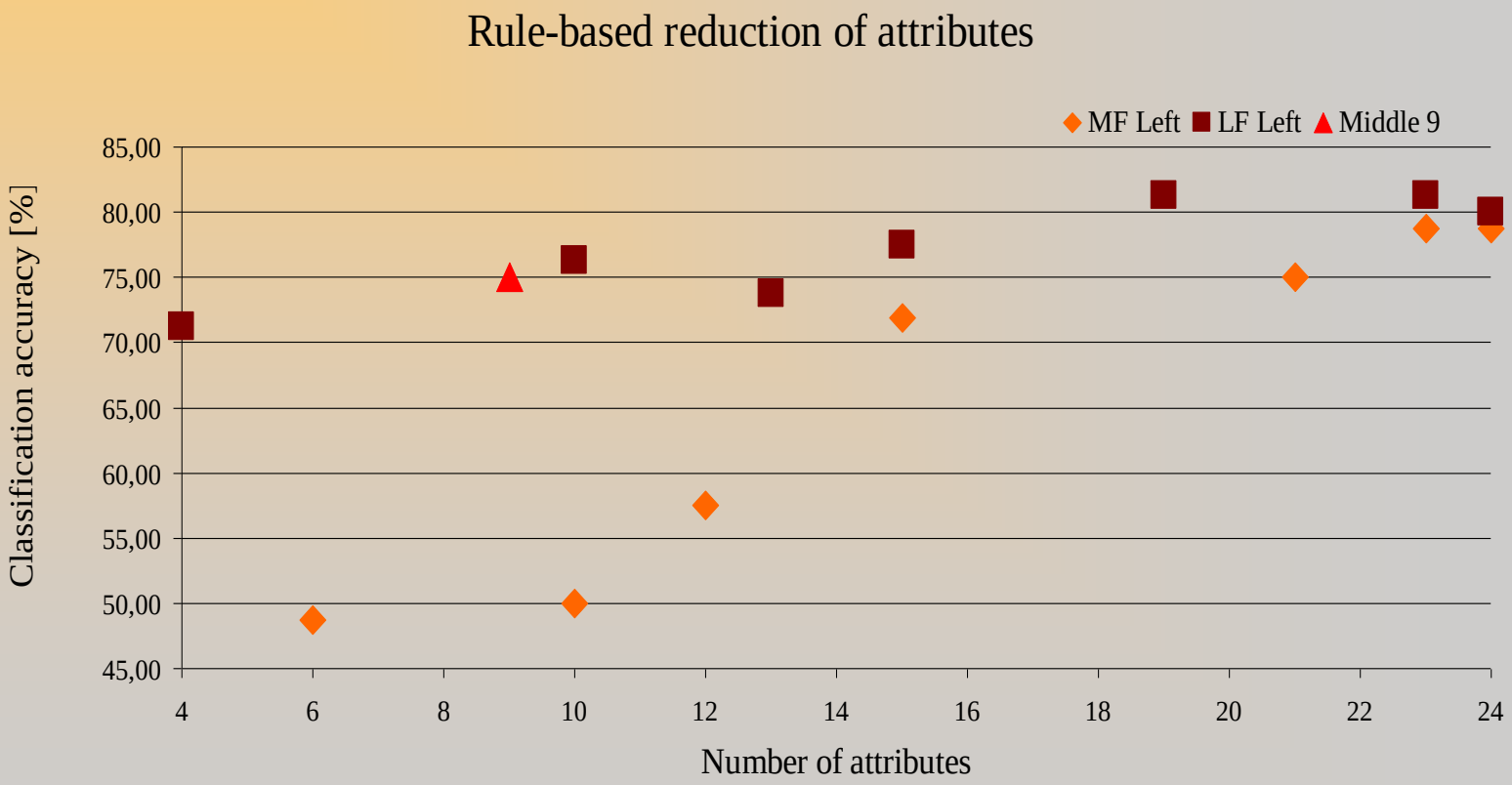
- 46191 rules found

Attribute	Occ. ind.	Attribute	Occ. ind.	Attribute	Occ. ind.
of	13310	in	10240	from	7614
on	12921	semicolon	9797	question	7468
to	11838	at	9082	for	7449
this	11426	with	8646	what	6172
comma	11176	as	8471	that	6166
fullstop	11004	by	8450	and	4172
exclamation	10639	hyphen	7996	but	3927
colon	10326	bracket	7950		
not	10305	if	7691		

Experiments 2 Reduct-Based



Experiments 2 Rule-Based



Conclusions

- Results with at least the same classification ratio for 25% disregarded input features
- Better results from rule- than reduct-based analysis
- Better results when keeping least frequently used features
- Instead of domain knowledge frequency analysis used in selection of features

Thank you for your attention