Hypertext Classification

Diploma thesis

Hervé Utard

Supervisor Professor Johannes Fürnkranz

Technische Universität Darmstadt

Presentation

- **Hypertext Classification**
- Related Work
- Our Model
- Implementation
- Results
- Conclusion

Presentation

- **Hypertext Classification**
  - **Accessing the information**
  - Text Classification
  - Hypertext Classification

Web $\leftrightarrow$ Libraries

Quality of a library

- completeness
- accessibility of the information

The Web is

- more and more complete
- hardly accessible

Web $\leftrightarrow$ Libraries

Accessibility in a library

- Classification by themes
- Alphabetical sort
- Ask the librarian

Accessibility on the Web

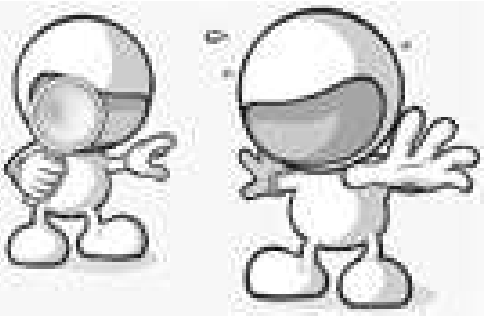
- No general Web Directory
- No global URLs list
- Search engines and Web Directories

Search engines

Boolean query $\longrightarrow$ list of web pages

Brief history

- 1990 Emtag (FTP filenames)
- 1994 Web crawler (first web search engine)
- 1998 Google (PageRank)
- 2003 Start of the Nutch project (Open source)



Web directory

- Links to other web sites
- Categorizes those links
- Historically collected by hand
  - Manual categorization is slow and costly
  - Categorization is subjective
- Automated Wed pages categorization
  - Understand both the document and the category

Presentation

- **Hypertext Classification**
  - Accessing the information
  - **Text Classification**
  - Hypertext Classification

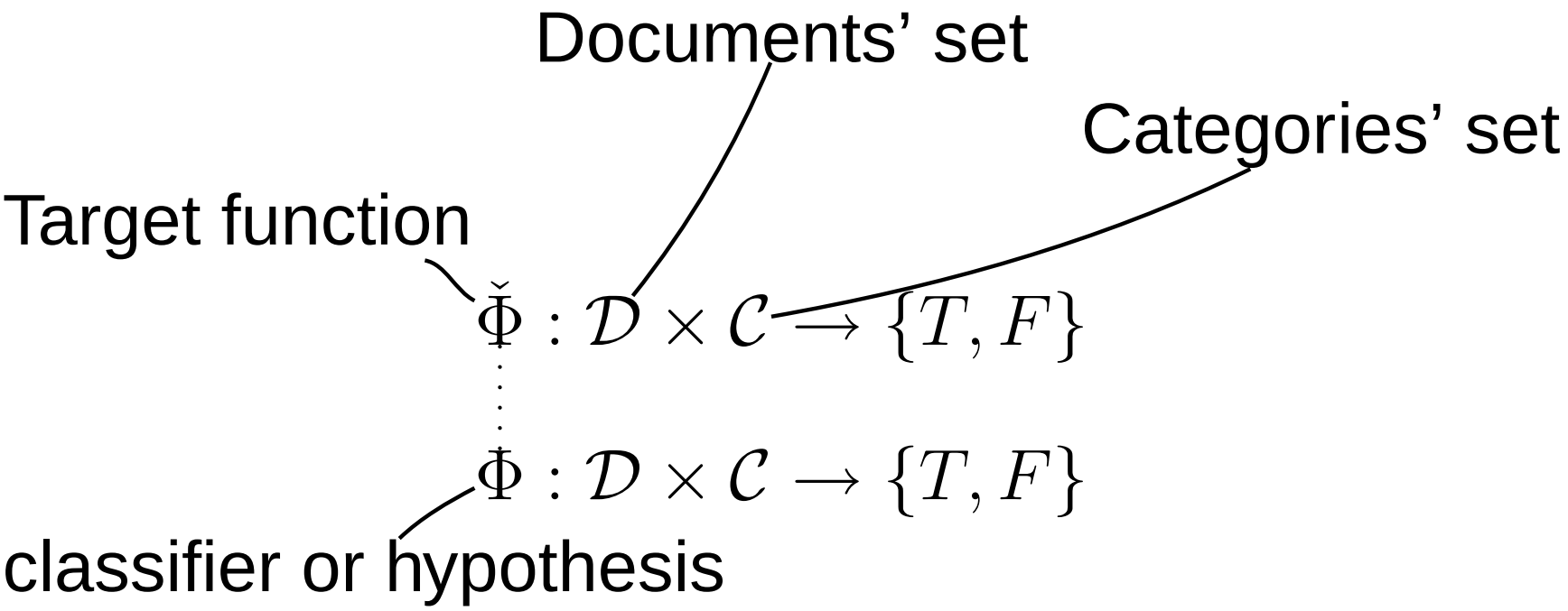
Categorization

- Task of predicting if a given document is related to a given category
- subfield of the information systems discipline
- born in the early ’60s
- first approach: ask a human expert to define manually a set of rules encoding his knowledge
- late ’80s, Machine Learning paradigm (extracting inductive knowledge from pre-classified documents)

Text Categorization

- indexing of digital libraries
- filing of newspaper articles
- Spam Filtering
- word sense disambiguation for polysemous words (*just*, *stand*)

Automated categorization



Approximate $\check{\Phi}$ by means of a function Φ such that $\check{\Phi}$ and Φ coincide as much as possible.

Automated categorization

Search engines

- polysemy
- response time

Web directories

- number of references
- update frequency
- cost
- same accuracy as manually designed models

Automated categorization

- Probabilistic classifiers
- Decision Rules
- Decision Trees
- Neural Networks
- Support Vector Machines

Presentation

- **Hypertext Classification**
  - Accessing the information
  - Text Classification
  - **Hypertext Classification**

Hypertext classification

Text categorization on the web

- big heterogeneousness
  - many authors
  - many languages
  - variety of topics
- irrelevant content
  - pictures
  - *Page under construction*

Hypertext classification

New information sources

- intern HTML structure
  - keywords
  - headings
  - lists
- Graph structure of the web
  - Predecessors or in-neighbors
  - Successors or out-neighbors
  - co-cited neighbors

Presentation

- Hypertext Classification
- **Related Work**
- Our Model
- Implementation
- Results
- Conclusion

Presentation

- **Related Work**
  - **Categorization using hyperlinks**
  - Link Mining
  - Categorization without the Web Page

Categorization using hyperlinks

Soumen Chakrabarti, 1998

- append the text of the neighbors $\rightarrow$ increase of the error rate
- relaxation labeling classifier using the class prediction of the neighbors $\rightarrow$ error rate reduced by 70%

Presentation

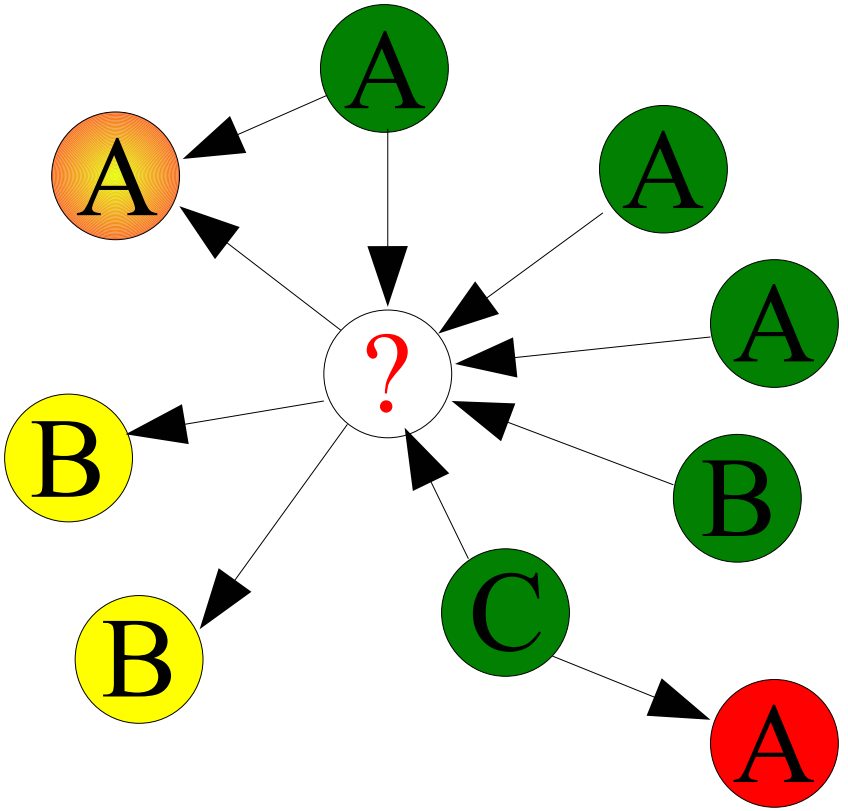
- **Related Work**
  - Categorization using hyperlinks
  - **Link Mining**
  - Categorization without the Web Page

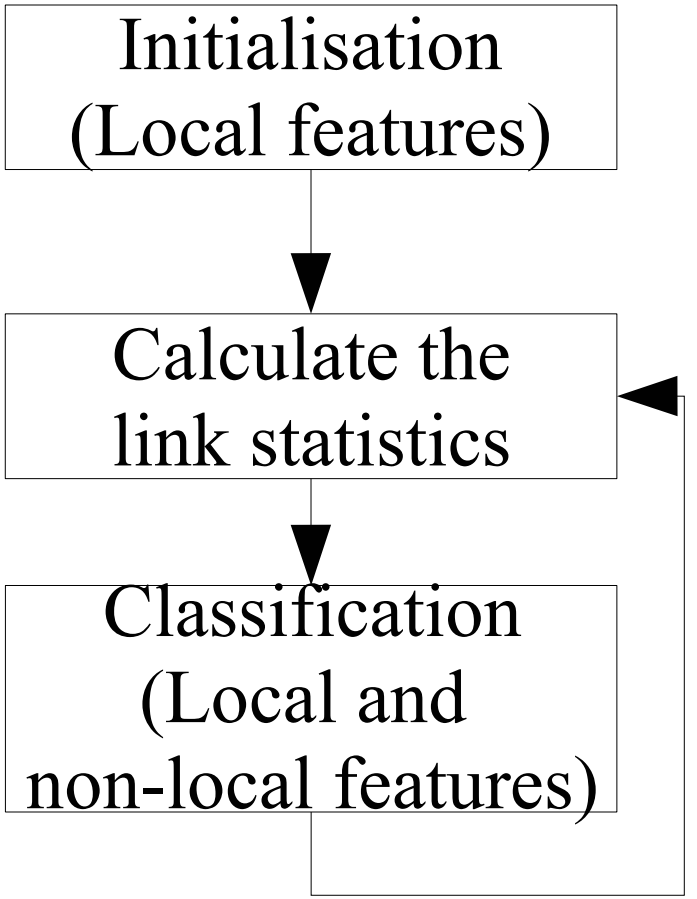
Link Mining

Lise Getoor and King Lu, 2003

- Feature mining
  - local features: words
  - non-local features: statistics about the category distribution of the neighbors
- Support Vector Machine

Link Mining


A
A
A
A
?
B
B
B
C
A


Initialisation
(Local features)
Calculate the
link statistics
Classification
(Local and
non-local features)

Link Mining

- Flat model: the local features and the non-local ones were concatenated into a common vector
- 2-step model: a local and a non-local prediction are computed independently and combined

The 2-step model outperforms the flat model

Presentation

- **Related Work**
  - Categorization using hyperlinks
  - Link Mining
  - **Categorization without the Web Page**

Categorization without the Web Page

Min-Yen Kan, 2004

- Web crawlers collect more URLS than classifiers can process
- Feature mining:
  - split the URL (scheme:://host/path-elements/document.extension)
  - expand the abbreviations
- results
  - $\frac{3}{4}$ as effective as text-based classifiers
  - outperforms title or anchor words

Presentation

- Hypertext Classification
- Related Works
- **Our Model**
- Implementation
- Results
- Conclusion

Presentation

- **Our model**
  - **Overview**
  - Various predecessors
  - Binarization of the multiclass problem
  - Various feature patterns

Overview

- Getoor and Chakrabarti showed that using the class predication of the neighbors increases the performances
- We believe that more than the categories of the neighbors, we should identify the category of each link

Overview

Mine local and link-specific non-local features

- Local features: the text content of the document
- Link-specific non-local features
  - anchor description
  - words neighboring the anchor
  - headings structurally preceding the link
  - heading of the list of link
  - paragraph surrounding the link

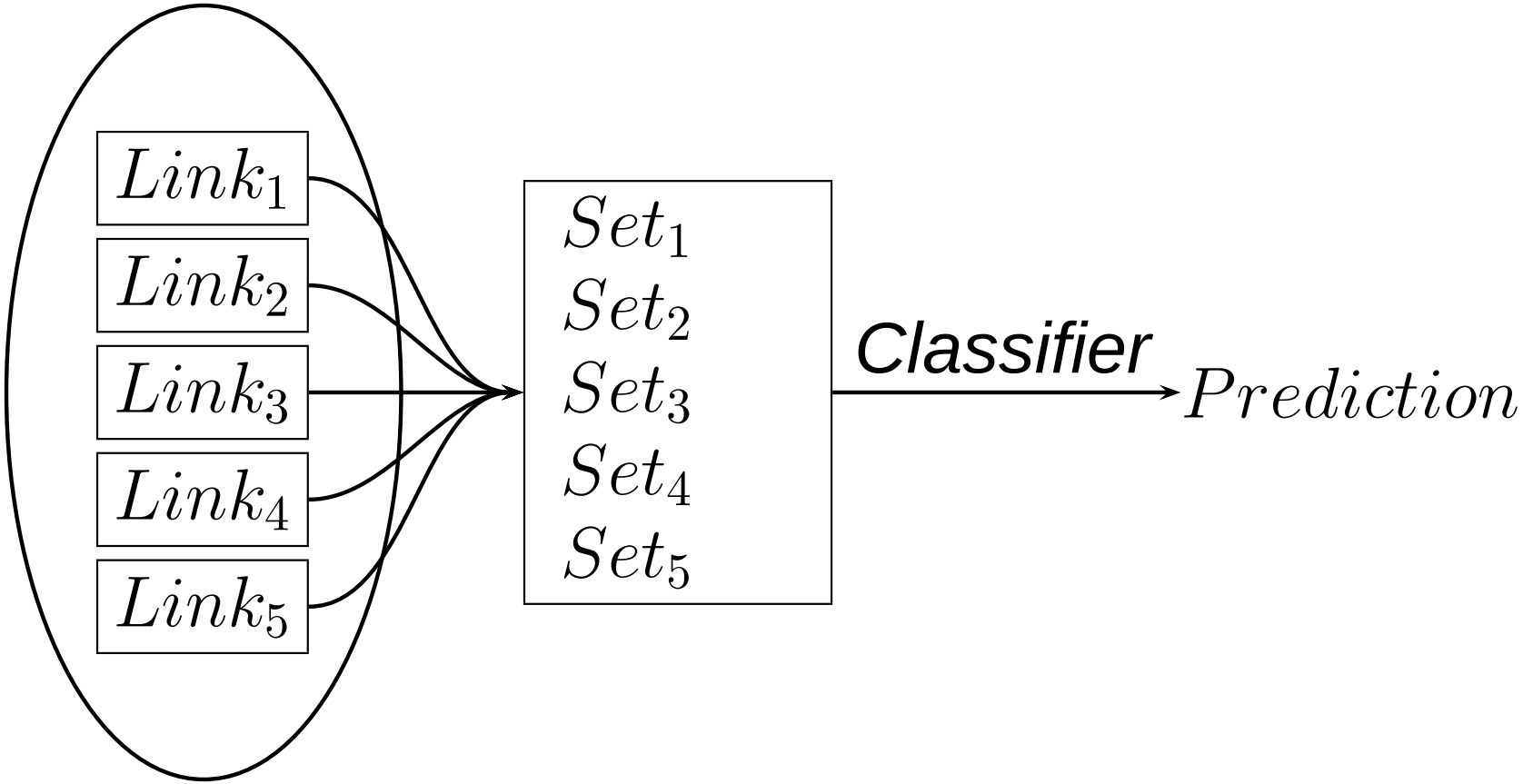
Presentation

- **Our model**
  - Overview
  - **Various predecessors**
  - Binarization of the multiclass problem
  - Various feature patterns

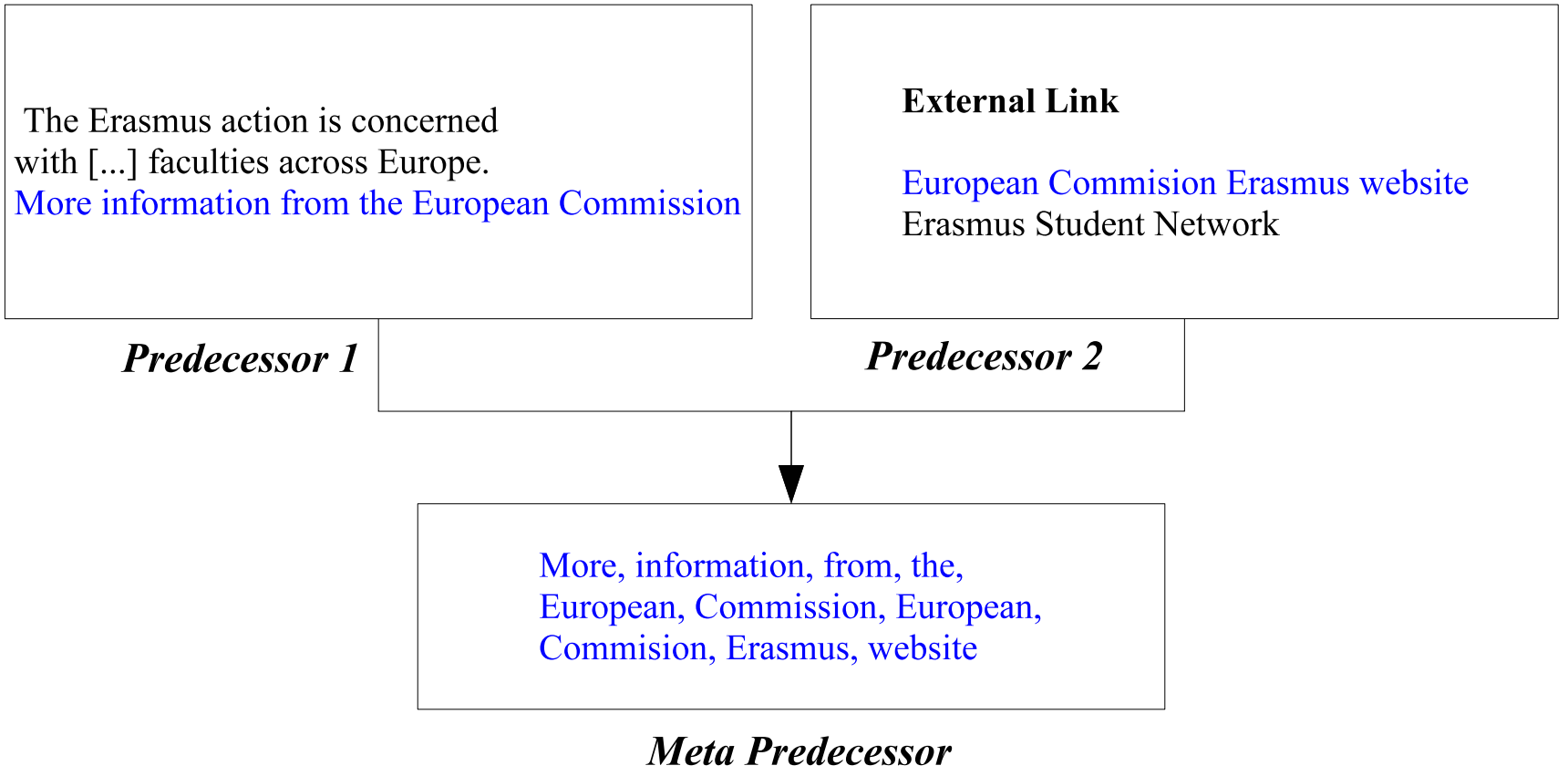
Learning from various predecessors

- Traditional classification problems: one features set per example
- Hyperlink-based Classification: one ensemble of features set per example

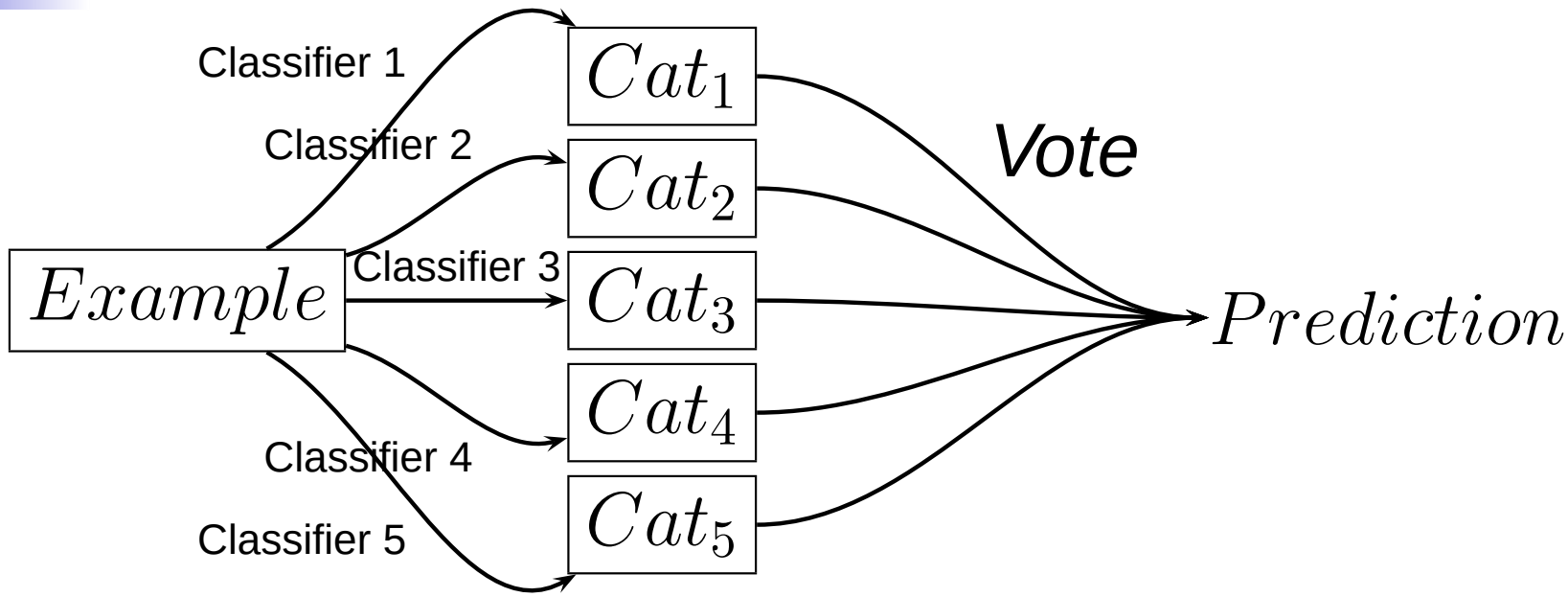
Meta Predecessor



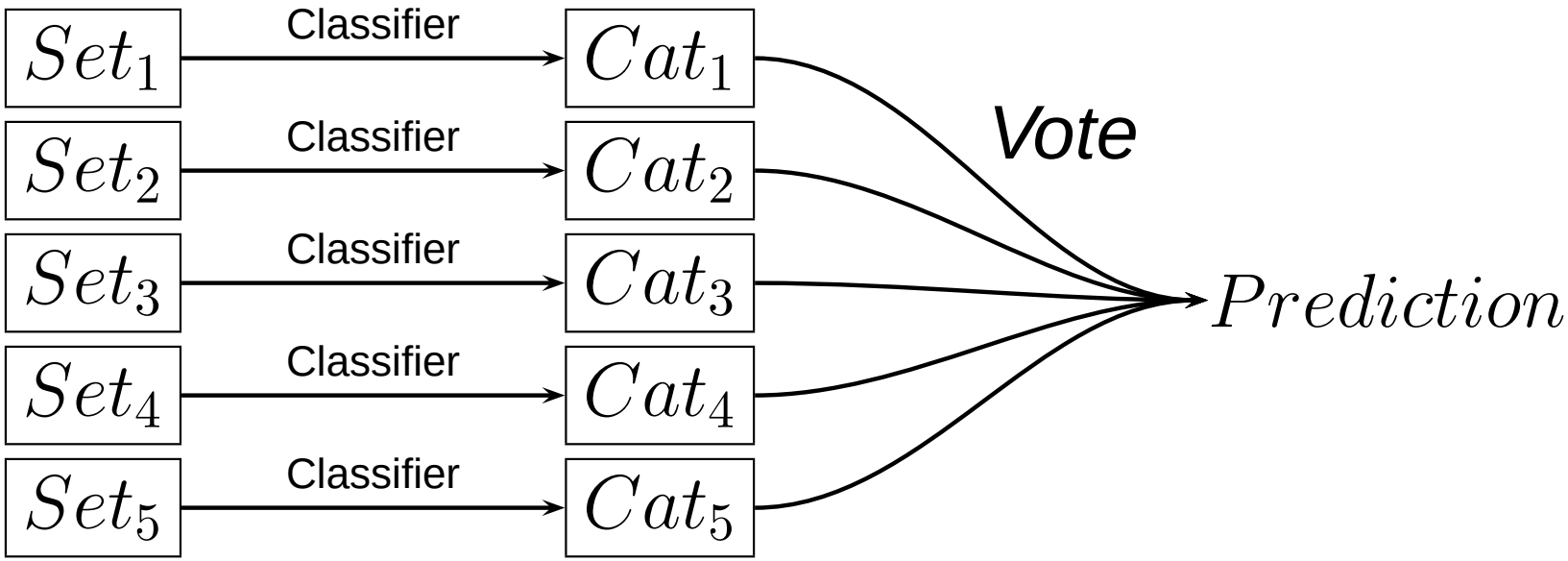
Meta Predecessor



Stacking


Classifier 1
Classifier 2
Classifier 3
Classifier 4
Classifier 5
$Example$
Cat_1
Cat_2
Cat_3
Cat_4
Cat_5
Vote
$Prediction$

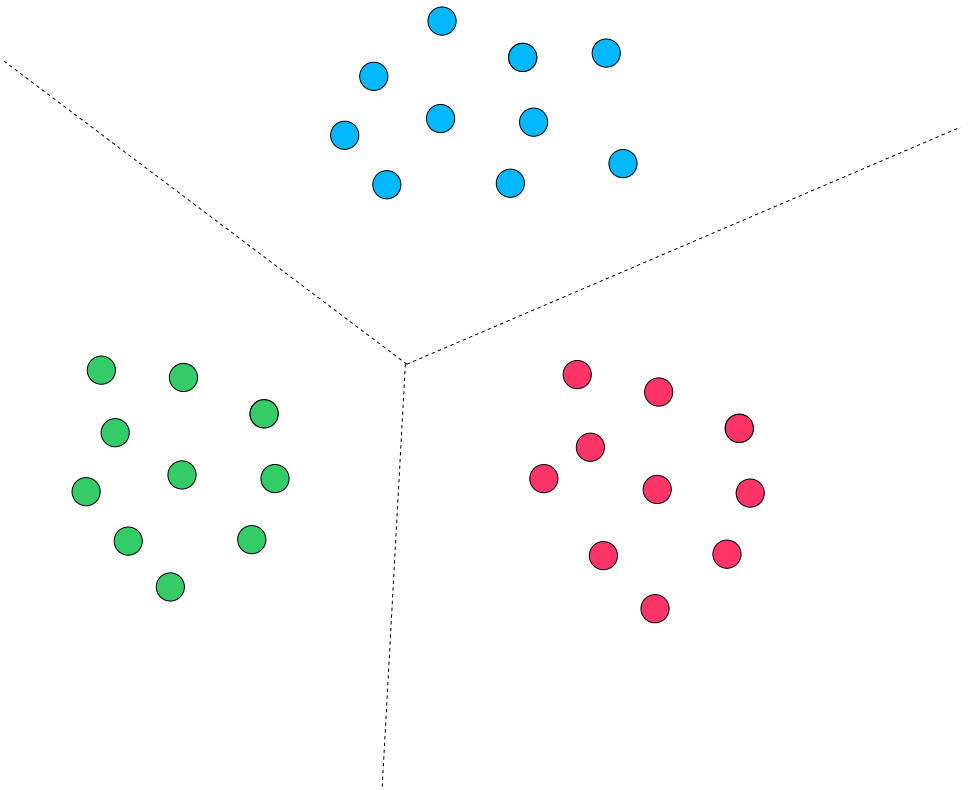
Hyperlink Ensembles



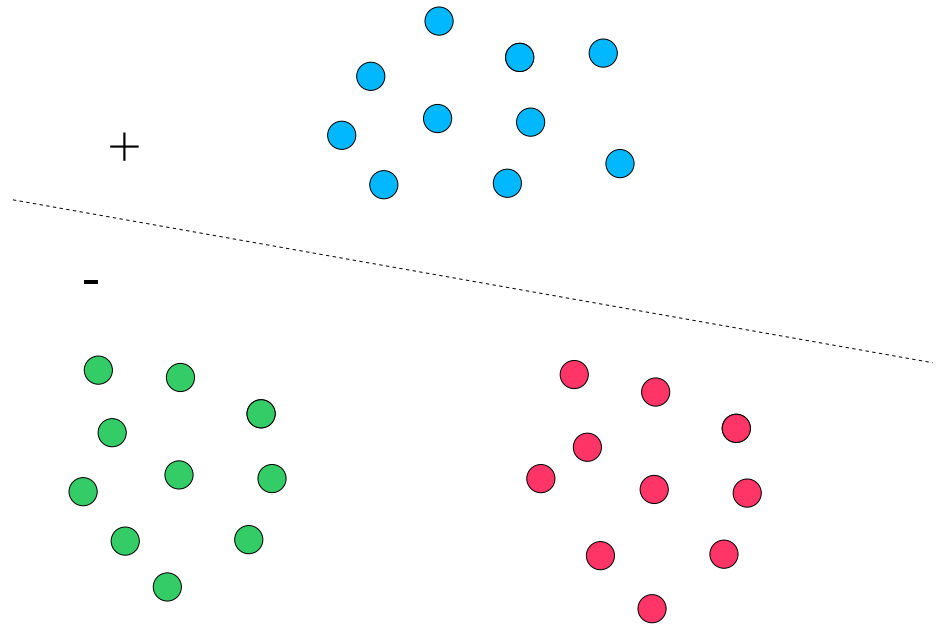
Presentation

- **Our model**
  - Overview
  - Various predecessors
  - **Binarization of the multiclass problem**
  - Various feature patterns

Multiclass binarization



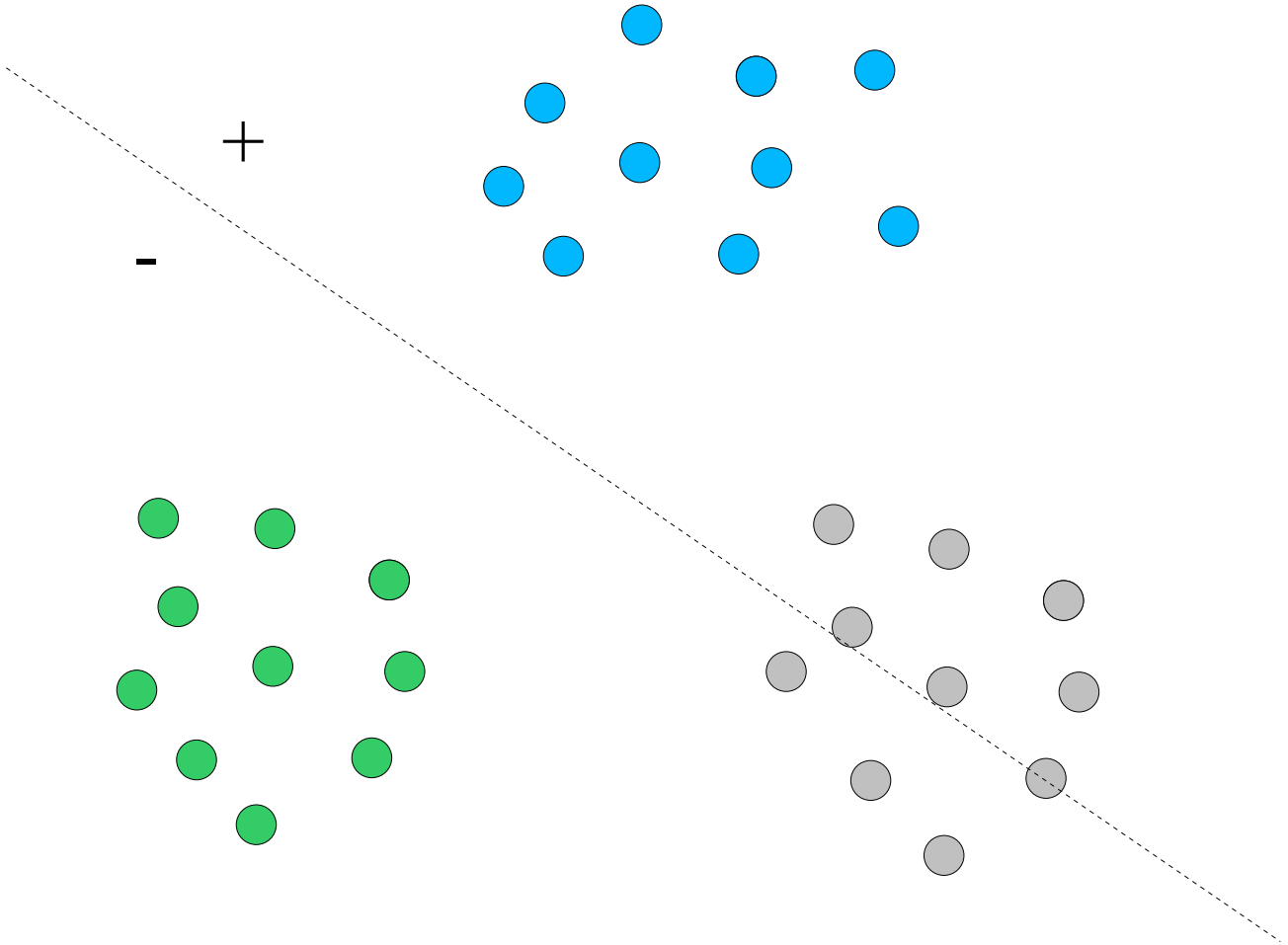
One against all



One against all

	Answer	*English*	*German*	*French*
Is it English ?	No	0	1	1
Is it German ?	No	1	0	1
Is it French ?	Yes	0	0	1
Sum		**1**	**1**	**3**

Round Robin



Round Robin

	Answer	*English*	*German*	*French*
Is it English or German ?	English	1	-1	0
Is it English or French ?	French	-1	0	1
Is it German or French ?	French	0	-1	1
Sum		0	-2	**2**

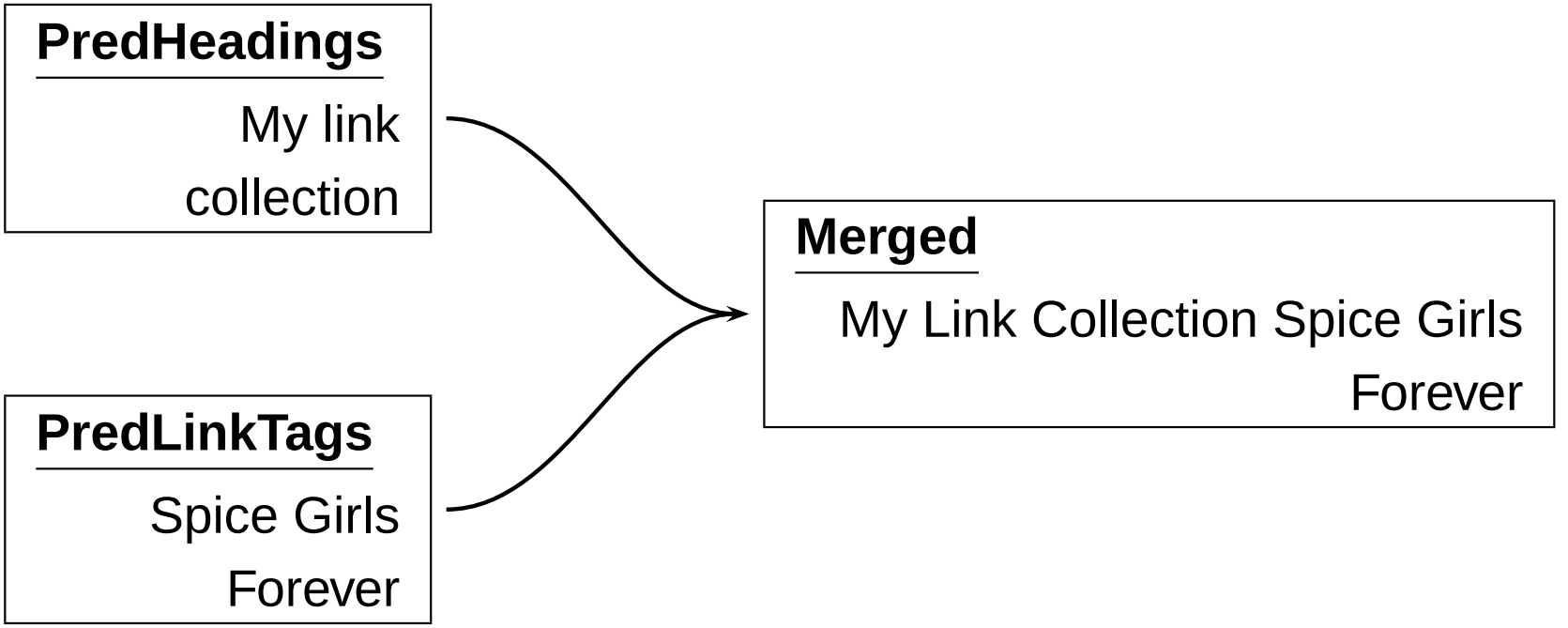
Presentation

- **Our model**
  - Overview
  - Features mined
  - Various predecessors
  - Binarization of the multiclass problem
  - **Various feature patterns**

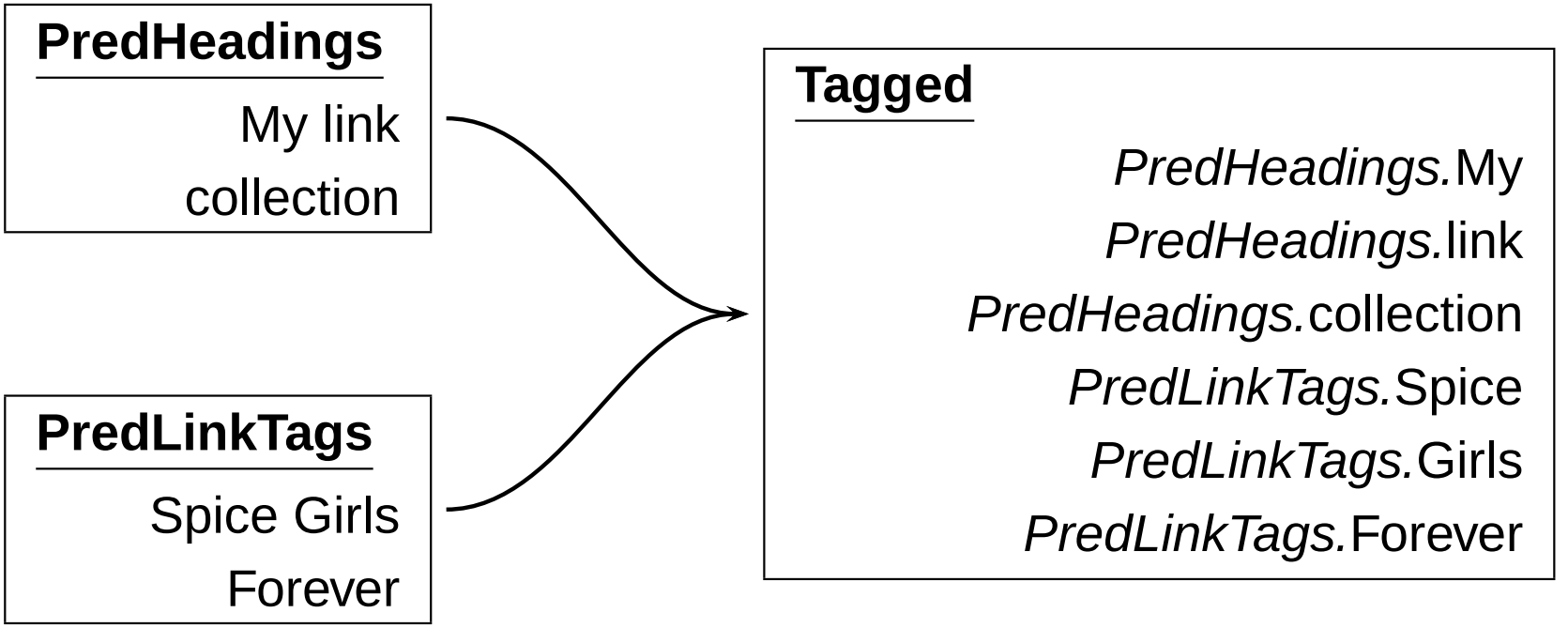
Feature patterns

- **PredLinkTags** The anchor description
- **PredLinkHeadings** The headings *structurally* preceding the link
- **PredLinkParagraph** The paragraph surrounding the link
- **PredListHeadings** The heading of the list of links
- **PredNWordsAroundAnchor** n words preceding or following the anchor
- **OwnText** content of the target page

Merging


PredHeadings
My link collection
PredLinkTags
Spice Girls Forever
Merged
My Link Collection Spice Girls Forever

Tagging


PredHeadings
My link
collection
PredLinkTags
Spice Girls
Forever
Tagged
PredHeadings.My
PredHeadings.link
PredHeadings.collection
PredLinkTags.Spice
PredLinkTags.Girls
PredLinkTags.Forever

Presentation

- Hypertext Classification
- Related Work
- Our Model
- **Implementation**
- Results
- Conclusion

Presentation

- **Implementation**
  - **The Benchmark Collections**
  - Support Vector Machines
  - Preprocessing
  - Mining the features
  - Cross validation

The benchmark collections

- Allesklar
  - strongly connected
  - specifically mined for this study
- WebKB
  - weakly connected
  - already tested by other researchers

The Allesklar dataset

- German generic web directory
- http://www.allesklar.de
- About 3 million of German web sites referenced
- 16 main categories (between 30 000 and 1 000 000 sites per main category)

The Allesklar Dataset

We chose 5 main categories

- Arbeit und Beruf (Work and Jobs)
- Bildung und Wissenschaft (Education and Science)
- Freizeit und Lifestyle (Hobbies and Lifestyle)
- Gesellschaft und Politik (Society and Politics)
- Immobilien und Wohnen (Accommodation)

The Allesklar dataset

Crawling

- Breadth-first traversal of each category
- Altavista predecessors request (ex:`link:europa.eu.int`)
- Proxy
- URL $\longrightarrow$ filename
- `_Classification`
- Graph structure: `_Predecessors`

Categories distribution

Category	Examples
Arbeit&Beruf	578
Bildung&Wissenschaft	809
Freizeit&Lifestyle	752
Gesellschaft&Politik	833
Immobilien&Wohnen	793

Classification

aaa-botzke.de	, Immobilien-Wohnen	, aaa-botzke.de
aaonline.dkf.deˆbbˆp109.htm	, Arbeit-Beruf	, aaonline.dkf.de/bb/p109.htm
abb-angermuende.de	, Immobilien-Wohnen	, abb-angermuende.de
action5.toplink.de	, Gesellschaft-Politik	, action5.toplink.de
agenturohnegrenzen.de	, Freizeit-Lifestyle	, agenturohnegrenzen.de
aib-backnang.de	, Arbeit-Beruf	, aib-backnang.de
akzente-zuelpich.de	, Immobilien-Wohnen	, akzente-zuelpich.de
allschutz.de	, Immobilien-Wohnen	, allschutz.de
anahato.bei.t-online.de	, Freizeit-Lifestyle	, anahato.bei.t-online.de
anderswelt.comˆkreiszeit	, Freizeit-Lifestyle	, anderswelt.com/kreiszeit

Predecessors

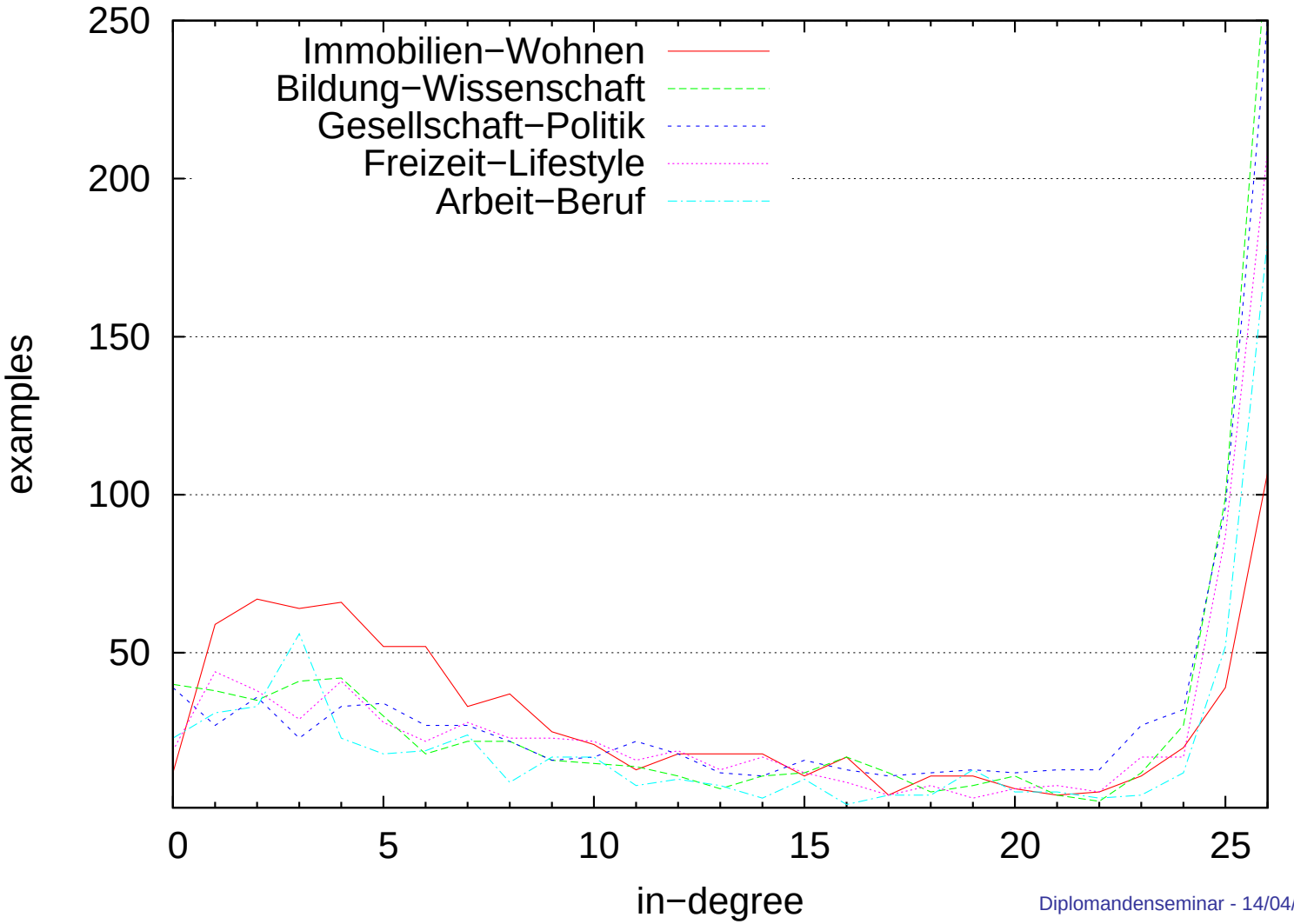
from aaonline.dkf.deˆbbˆp109.htm : www.ralf-bales.deˆgesamt.htm ; www.open-skies.orgˆh

from berufenet.arbeitsamt.de : www.studienwahl.deˆfmg.htm ; www.was-werden.de ; . . .

from home.degnet.deˆkoller_stefanˆlyricsˆly_start.htm : lyrics.berger-rangers.de ; elcapitan

from home.t-online.deˆhomeˆschmidt.re : www.lyrik.chˆlyrikˆlinks.htm ; www.lyrik.de ; www.

In-degree on Allesklar



The WebKB dataset

- Web pages collected from computer science departments
  - Cornell
  - Washington
  - Wisconsin
  - Texas
  - misc
- used as test set in numerous papers

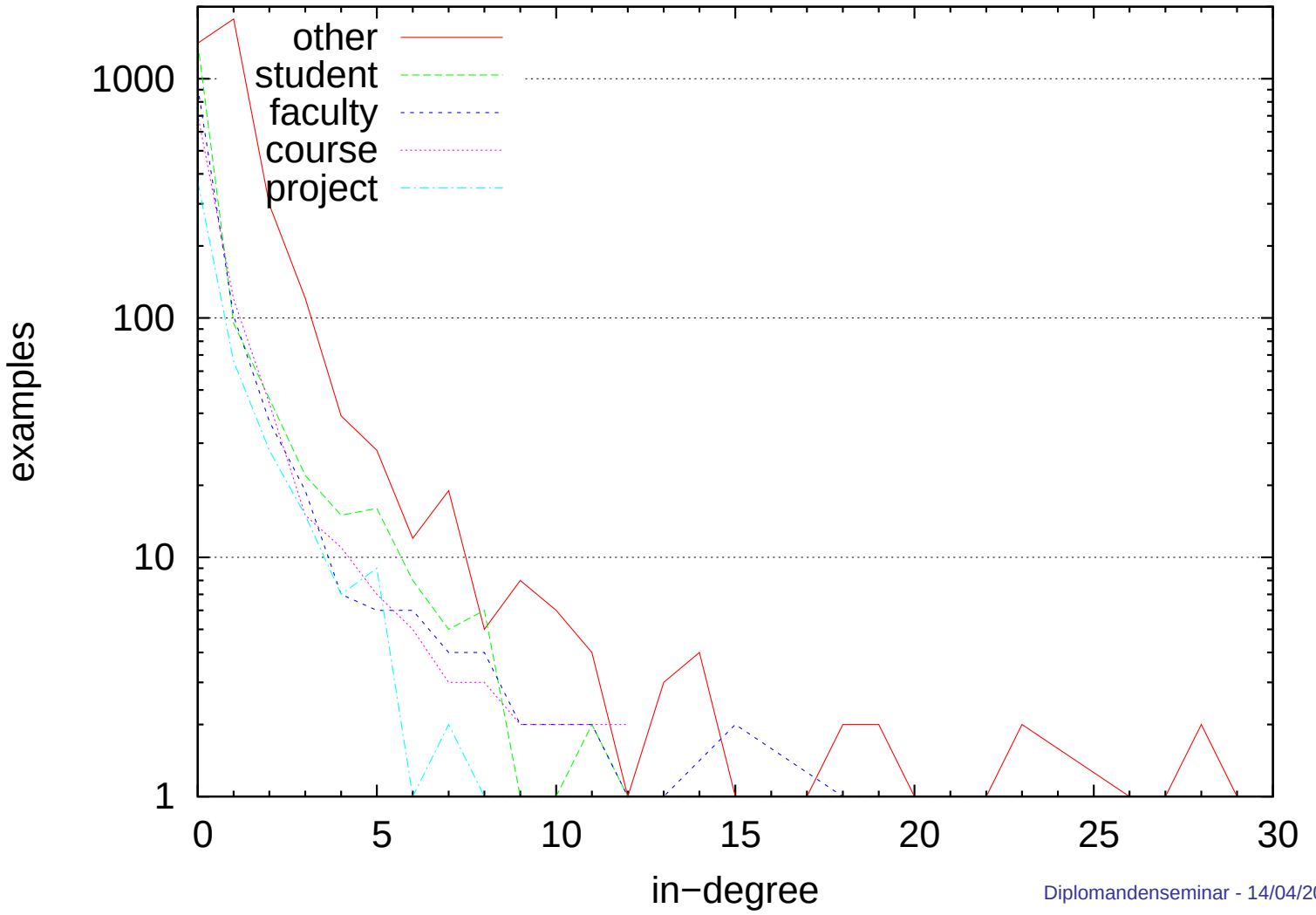
The WebKB dataset

- 7 categories
  - student
  - faculty
  - course
  - project
  - department
  - staff
  - other ($\approx 75\%$ of the examples)

The WebKB Dataset

category	Examples
other	3756
student	1639
faculty	1121
course	926
project	506
department	181
staff	135

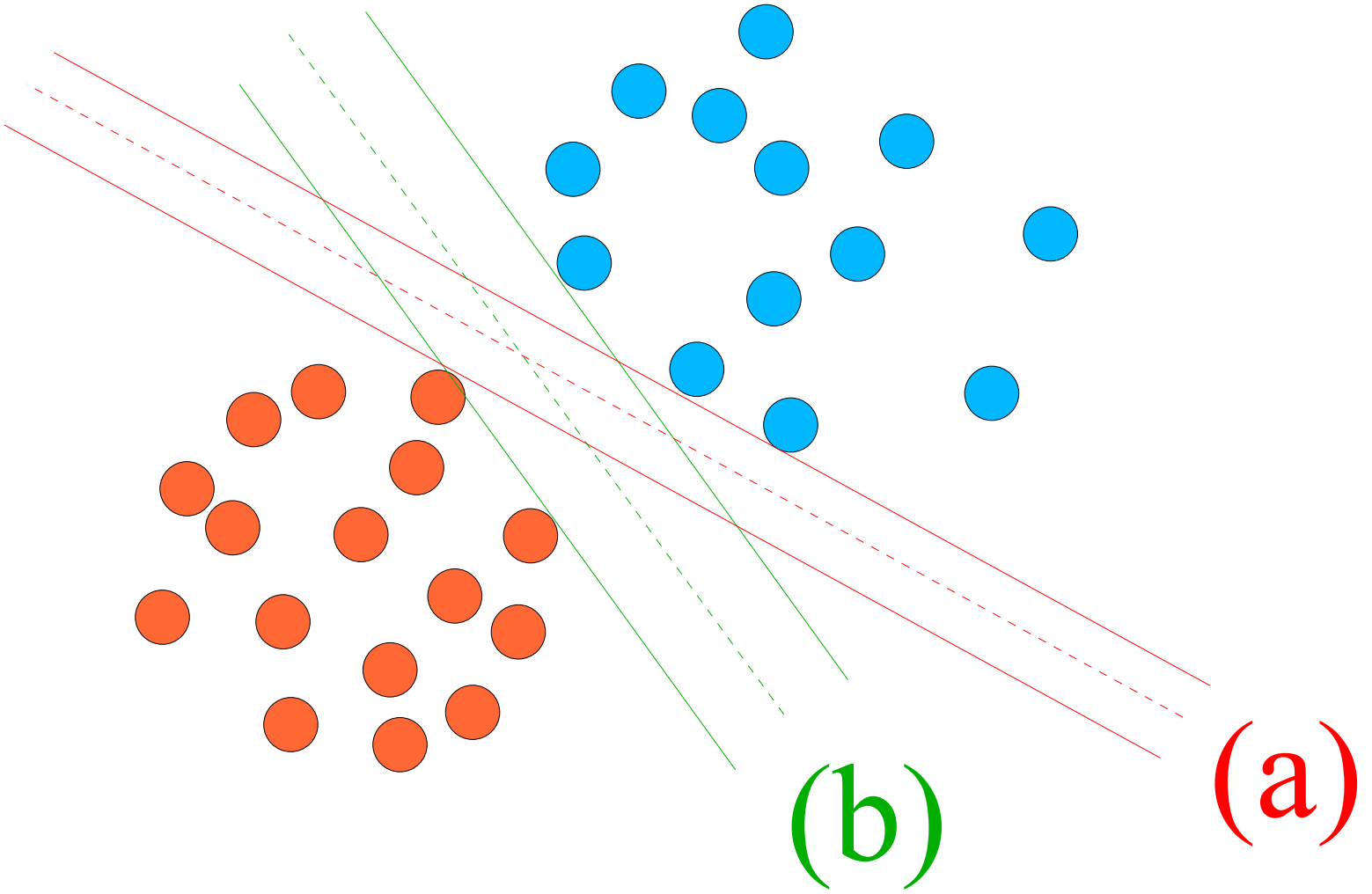
In-degrees on WebKB



Presentation

- **Implementation**
  - The Benchmark Collections
  - **Support Vector Machines**
  - Preprocessing
  - Mining the features
  - Cross validation

Support Vector Machines



Support Vector Machines

Set of examples: $\vec{x_i} \in \mathcal{R}^n$ with $i = 1, 2, ..., N$.

$\forall i, \vec{x_i} \in y_i \in \{-1, 1\}$

$\exists\, \vec{w} \in \mathcal{R}^n, b \in \mathcal{R}$

(1) $$y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1, i = 1, 2, ..., N$$

Support Vector Machines

$(\vec{w}, b)$: hyperplane of equation $\vec{w} \cdot \vec{x_i} + b = 0$ named separating hyperplane.

We rescale the pair $(\vec{w}, b)$ in $(\vec{w_0}, b')$ so that the distance of the closest point, say x_j, to the hyperplane equals $\frac{1}{\|\vec{w_0}\|}$

Support Vector Machines

The signed distance d_i of a point $\vec{x_i}$ is given by

(2) $$d_i = \frac{\vec{w_0} \cdot \vec{x_i} + b'}{\| \vec{w_0} \|}$$

And thus, with 1 and 2,

(3) $$\forall x_i \in S, y_i d_i \geq \frac{1}{\| \vec{w_0} \|}$$

Support Vector Machines

Maximize $\frac{1}{\|\vec{w_0}\|}$

Minimize $\| \vec{w_0} \|$

Minimize $\frac{1}{2} \vec{w_0} \cdot \vec{w_0}$.

Support Vector Machines

- Not linearly separable datasets
- Give a weight to each example
  - 1629 Pierre de Fermat
  - 1797 Lagrange
  - 1951 Kuhn and Tucker extended the Lagrangian theory in 1951

Support Vector Machines

The problem of minimizing $\frac{1}{2}\ \vec{w_0} \cdot \vec{w_0}$ subject to the correct classification constraint $\forall i \in [1, N]\,, y_i(\vec{w_0} \cdot \vec{x_i} + b) \geq 1$ becomes with the relative weight α_i granted to each example x_i the problem of finding the saddle point of the function L.

(4) $$L = \frac{1}{2}\ \vec{w} \cdot \vec{w} - \sum_{i=1}^{N} \alpha_i (y_i(\vec{w} \cdot \vec{x_i} + b) - 1)$$

Support Vector Machines

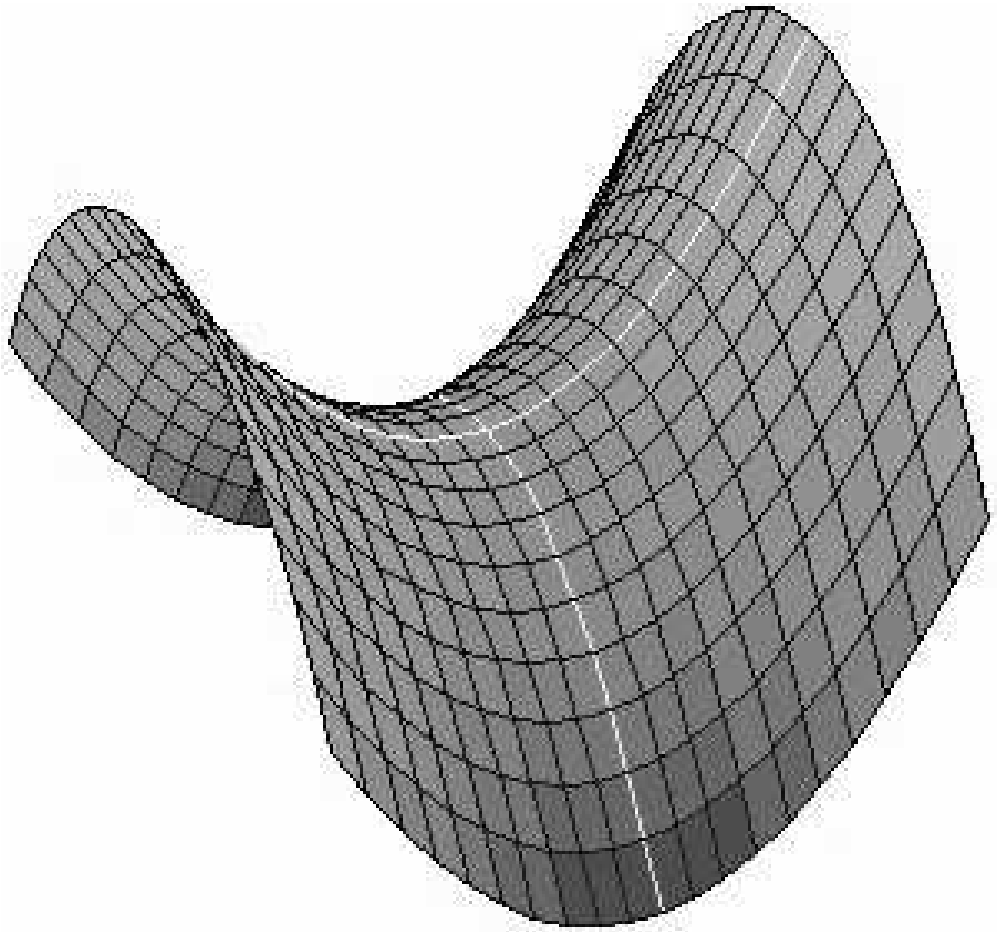


Figure 1: Saddle point

Support Vector Machines

At the saddle point,

(5)
$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} y_i \alpha_i = 0$$

(6)
$$\frac{\partial L}{\partial \vec{w}} = \vec{w} - \sum_{i=1}^{N} y_i \alpha_i \vec{x}_i = \vec{0}$$

with

(7)
$$\frac{\partial L}{\partial \vec{w}} = \left(\frac{\partial L}{w_1}, \frac{\partial L}{w_2}, ..., \frac{\partial L}{w_N}\right)$$

Support Vector Machines

The hyperplane coordinates $(\vec{w}, b)$

$$\begin{cases} \vec{w} & = & \sum_{i=1}^{N} y_i \alpha_i \ \vec{x_i} \\ b & = & ArgMax(\sum_{i=1}^{N} \alpha_i y_i (\vec{w} \cdot \vec{x_i} \ -1)) \end{cases}$$

Support Vector Machines

Classifying $\vec{d}$

$$\vec{w} \cdot \vec{d} + b \begin{cases} \geq +1 & \vec{d} \text{ is positive} \\ \in [0;1[& \vec{d} \text{ is probably positive} \\ \in]-1;0[& \vec{d} \text{ is probably negative} \\ \leq -1 & \vec{d} \text{ is negative} \end{cases}$$

Decision function $D(\vec{d}) = sign(\vec{w} \cdot \vec{d} + b)$

Comparison

- Support vector machines and boosting-based classifier committees
- Neural networks and on-line linear classifiers
- Rocchio classifiers and naive Bayes classifiers

Presentation

- **Implementation**
  - The Benchmark Collections
  - Support Vector Machines
  - **Preprocessing**
  - Mining the features
  - Cross validation

Preprocessing

- Remove the HTML tags
- Lower case the text
- Remove the diacritic signs
- Replace the remaining non alphanumeric characters by a _
- Replace the numbers by a single D
- Stop words list
- Document frequency based dimensionality reduction

Presentation

- **Implementation**
  - The Benchmark Collections
  - Support Vector Machines
  - Preprocessing
  - **Mining the features**
  - Cross validation

Mining the features

- HTML $\xrightarrow{Tidy}$ XHTML
- XHTML $\rightarrow$ DOM tree
- XPath patterns

XPath patterns

PredLinkTags	`//a[\@href='Target_SURL']`
PredLinkParagraph	`//a[\@href='Target_SURL']/ancestor:`

PredLinkHeadings	`//a[\@href='Target_SURL']/preceding`
	`| //a[\@href='Target_SURL']/preceding`
	`| //a[\@href='Target_SURL']/preceding`

PredListHeadings	`//a[\@href='Target_SURL']/ancestor`
	`::ul/preceding::h1[last()]`
	`| //a[\@href='Target_SURL']/ancestor`
	`::ul/preceding::h2[last()]`
	`| //a[\@href='Target_SURL']/ancestor`
	`::ul/preceding::h3[last()]`

Presentation

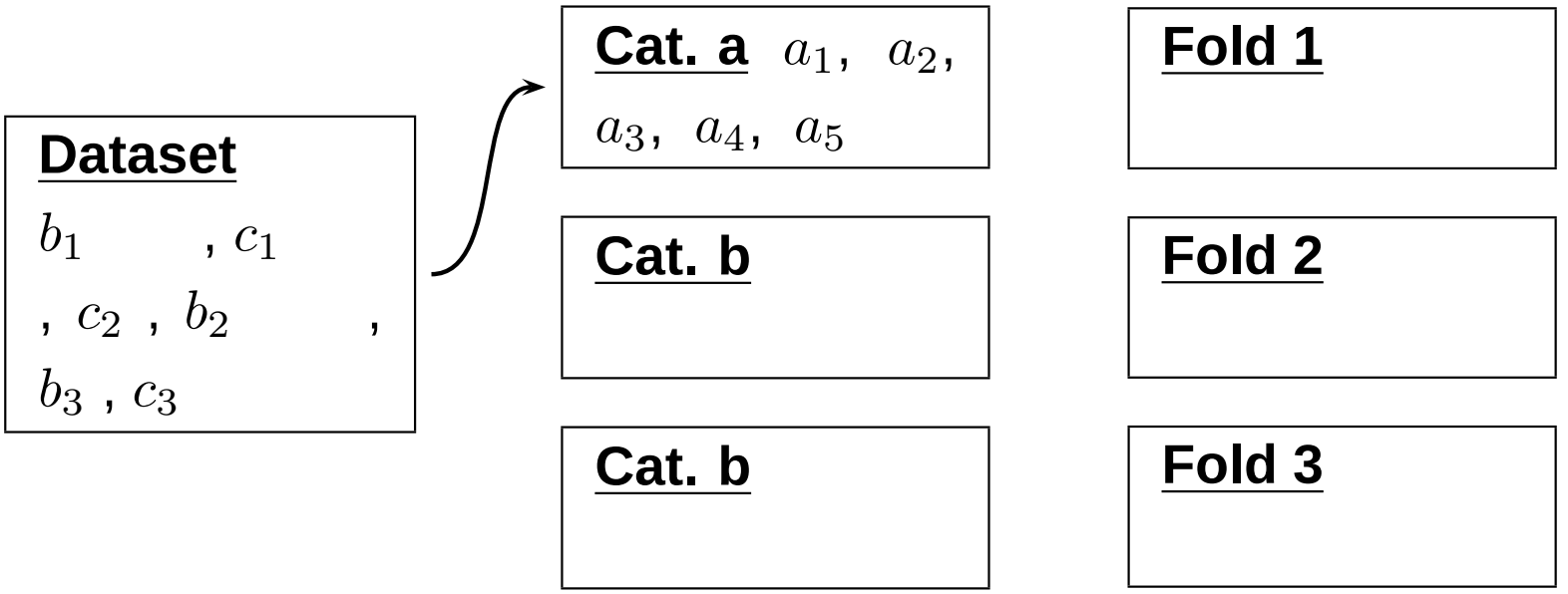
- **Implementation**
  - The Benchmark Collections
  - Support Vector Machines
  - Preprocessing
  - Mining the features
  - **Cross validation**

Stratified cross validation

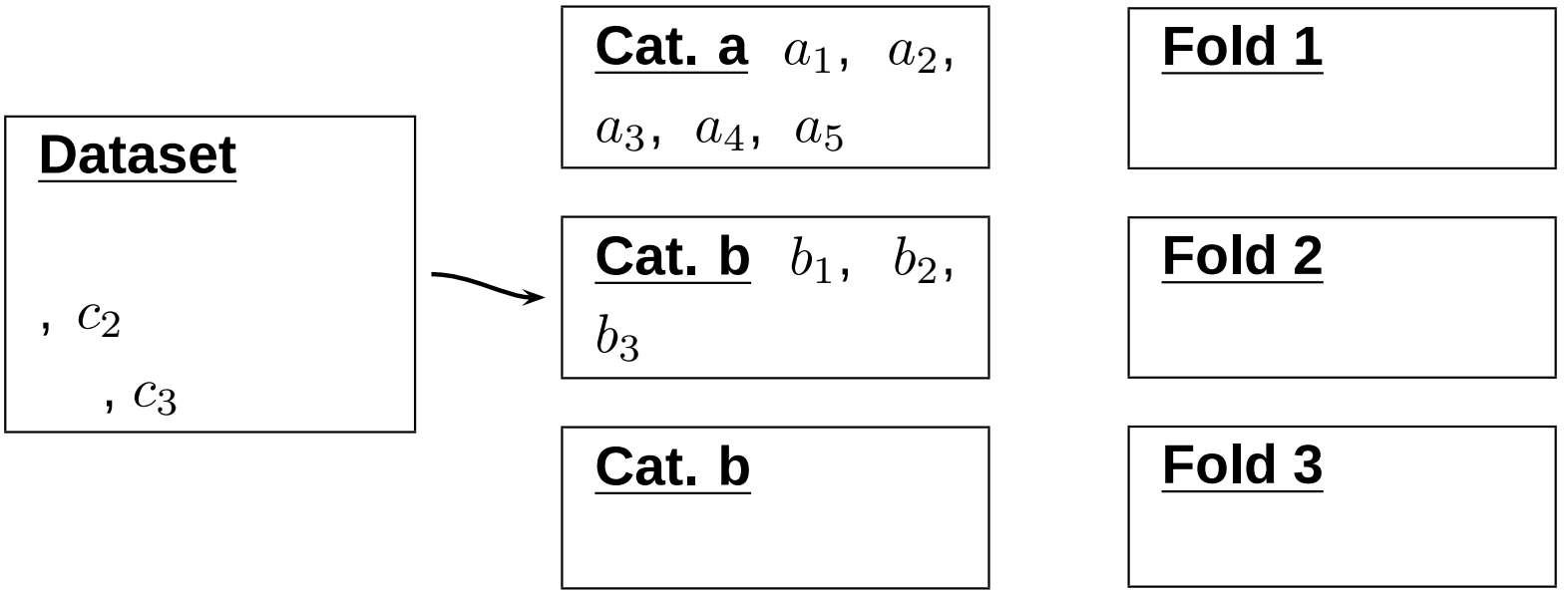
Dataset a_1, b_1 , a_2 , c_1 , a_3 , c_2 , b_2 , a_4 , b_3 , c_3 , a_5

Cat. a	Fold 1
Cat. b	Fold 2
Cat. b	Fold 3

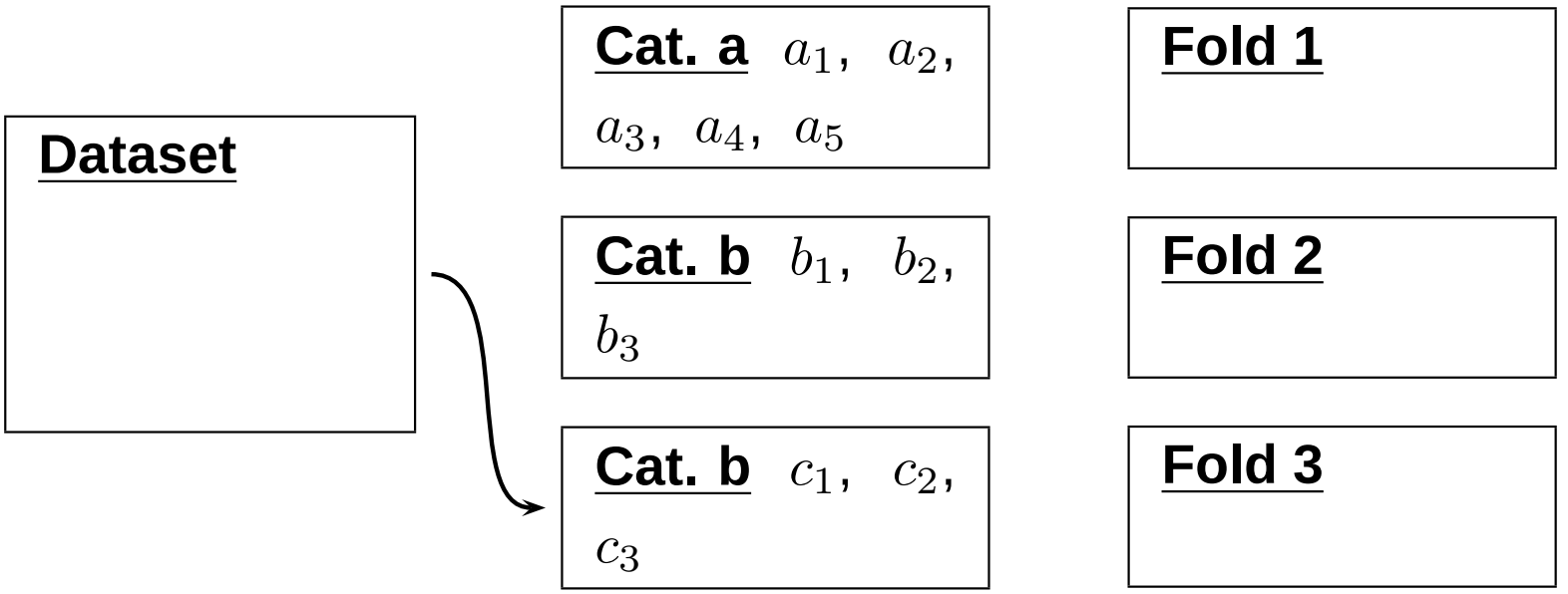
Stratified cross validation



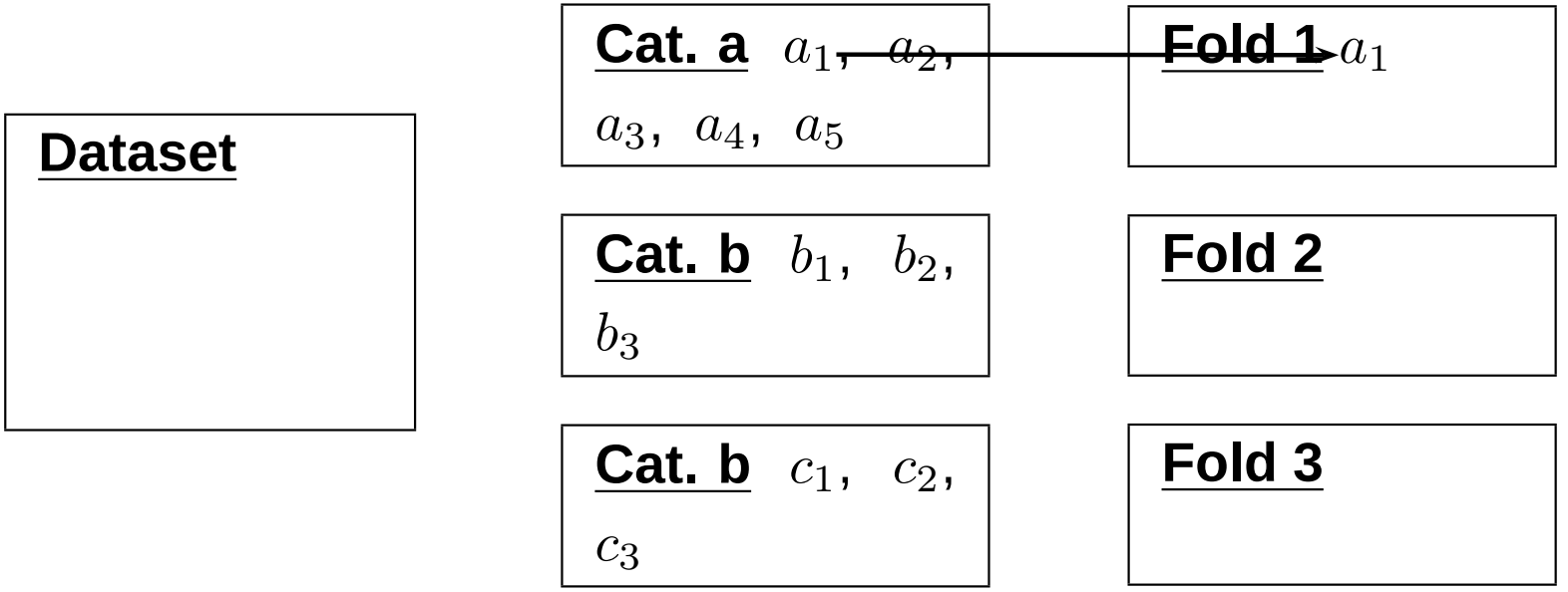
Stratified cross validation



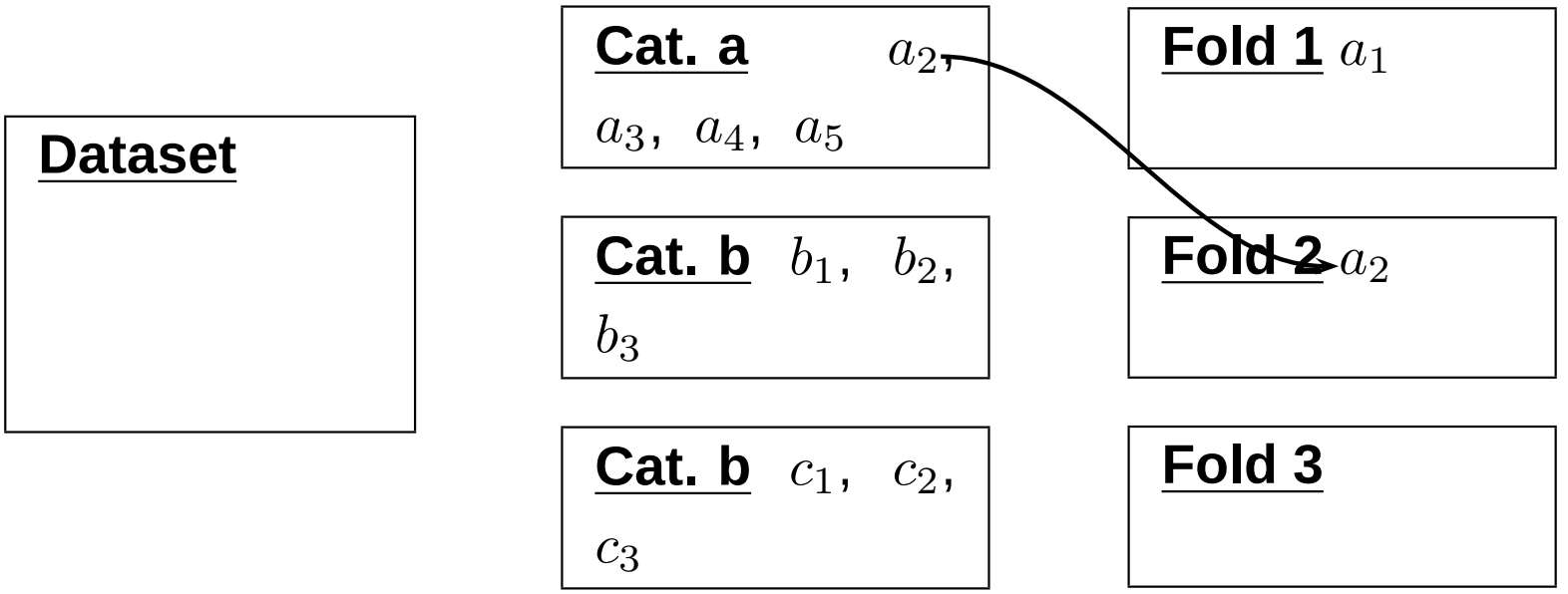
Stratified cross validation



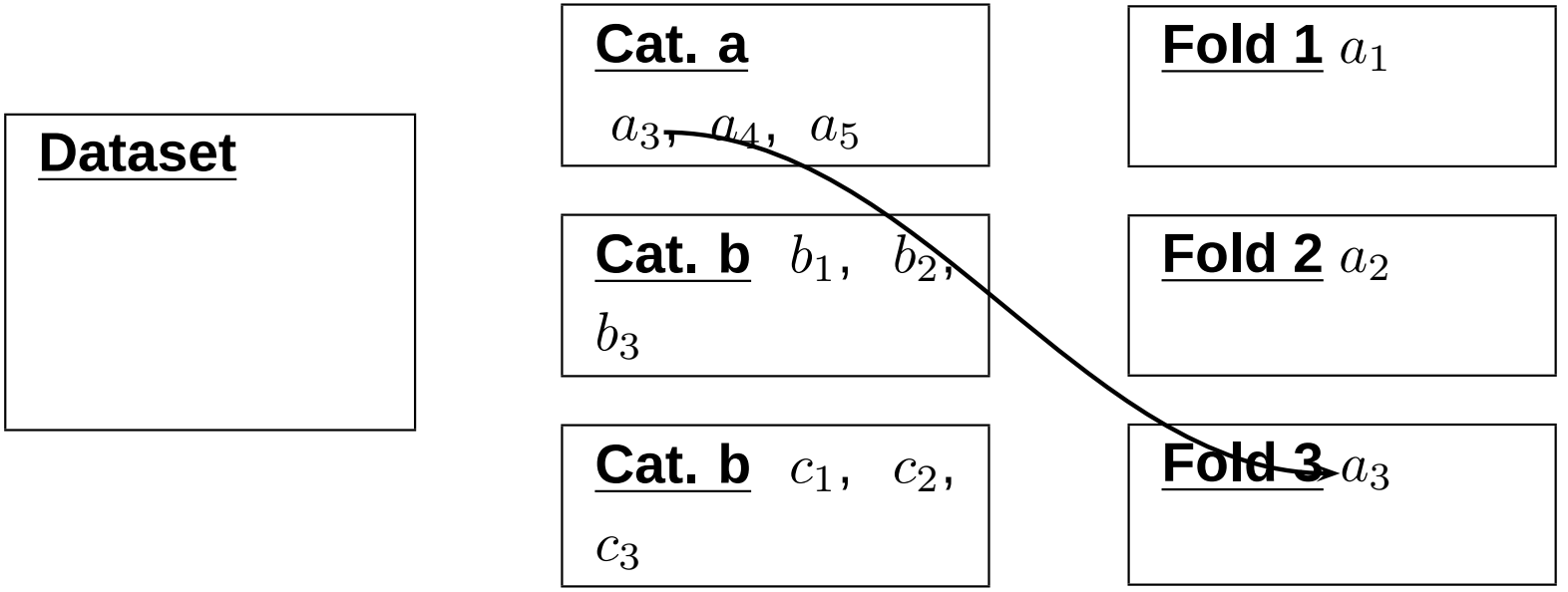
Stratified cross validation



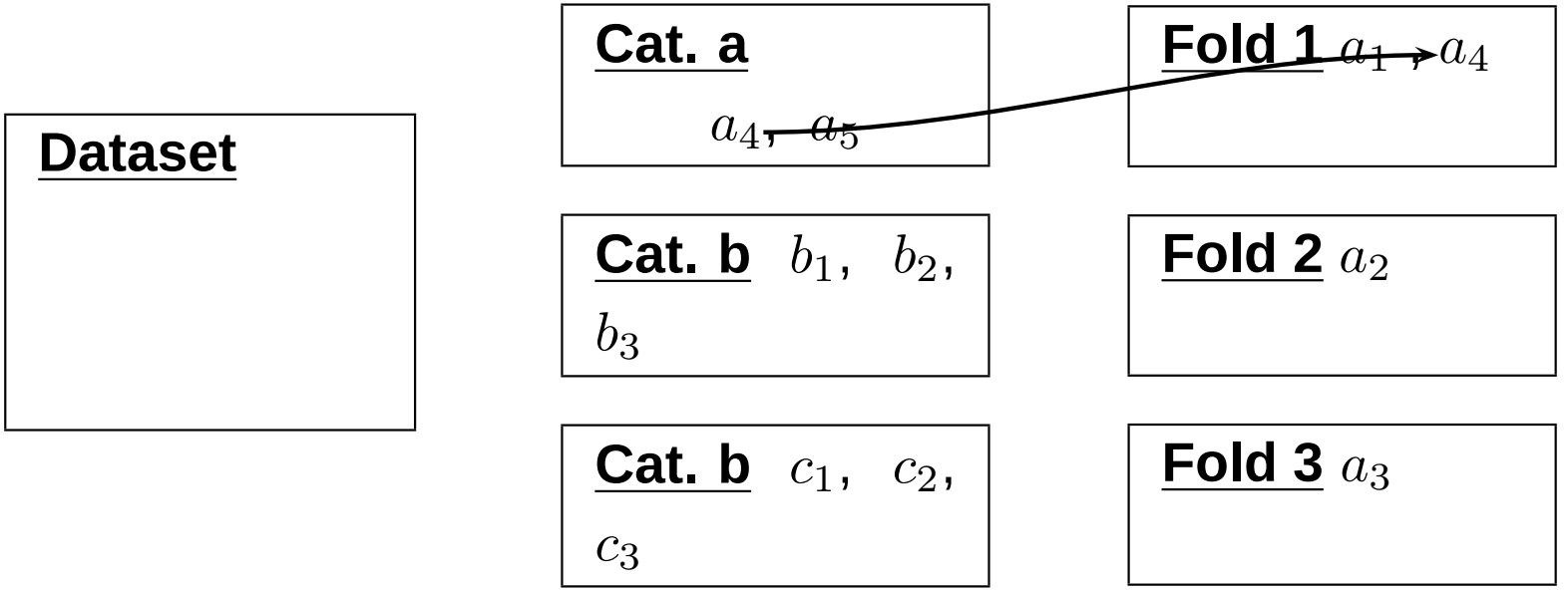
Stratified cross validation



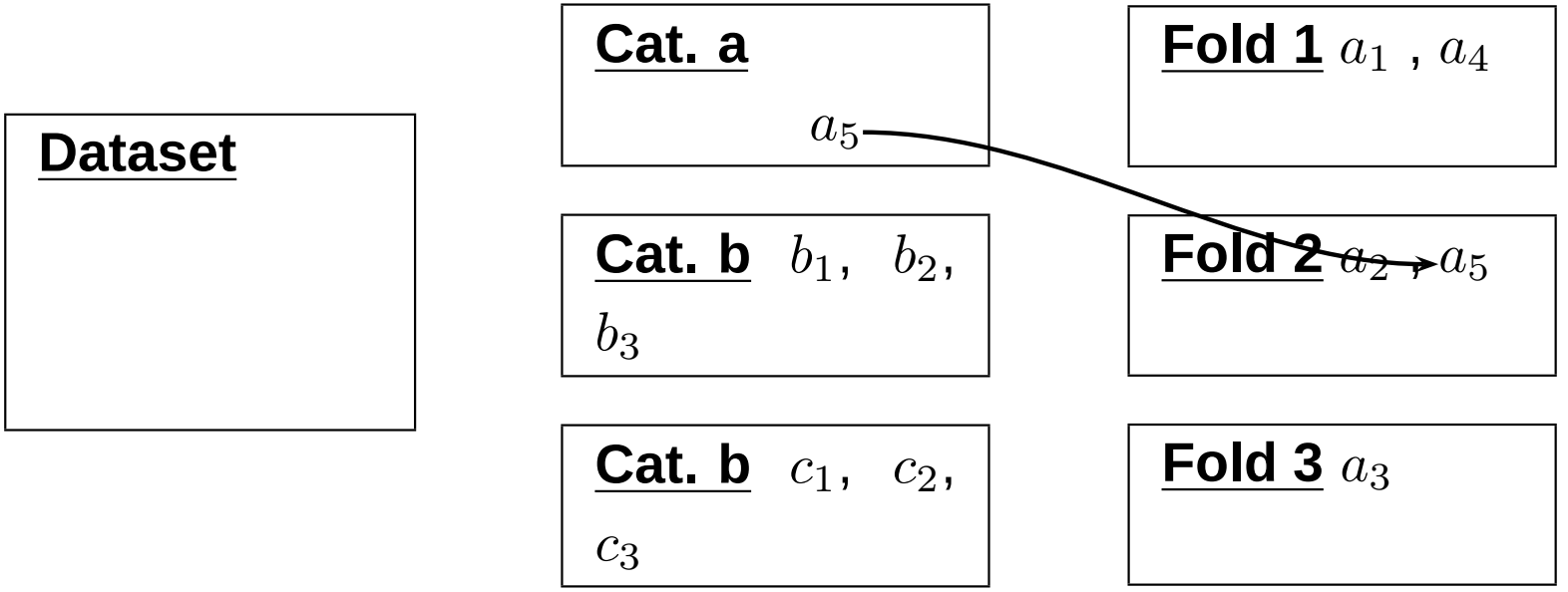
Stratified cross validation



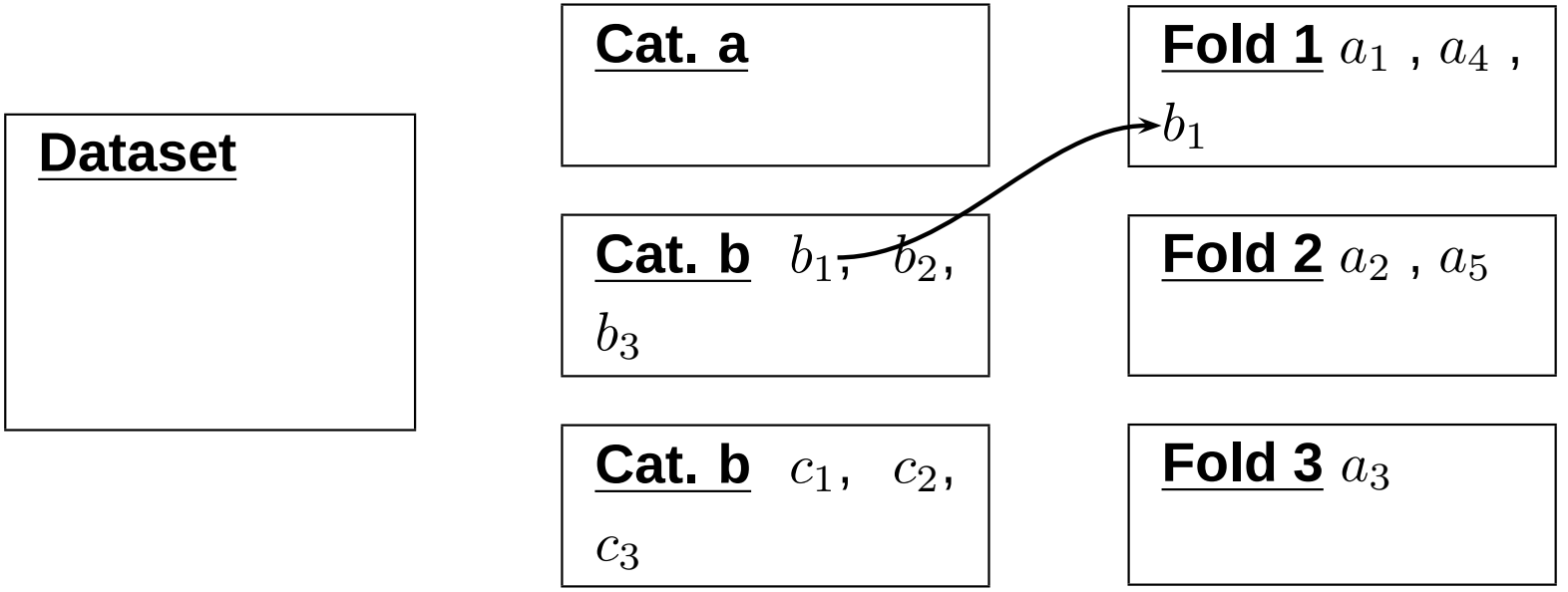
Stratified cross validation



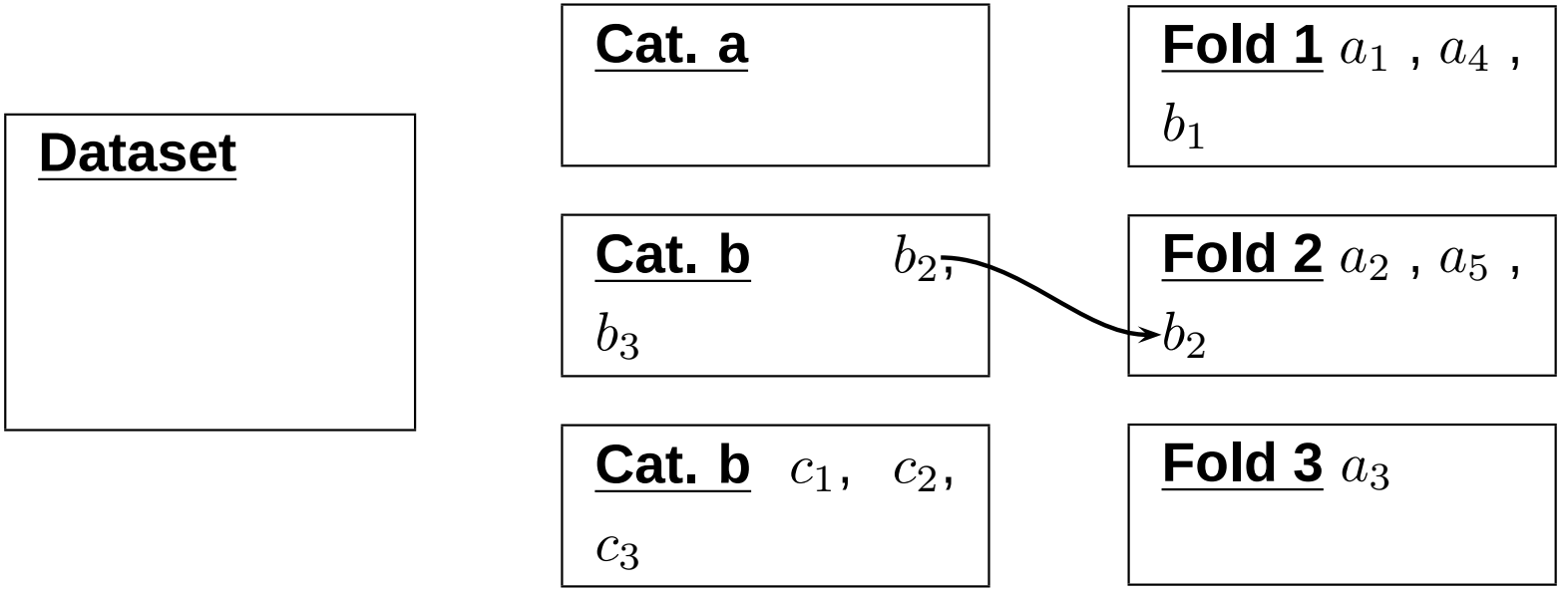
Stratified cross validation



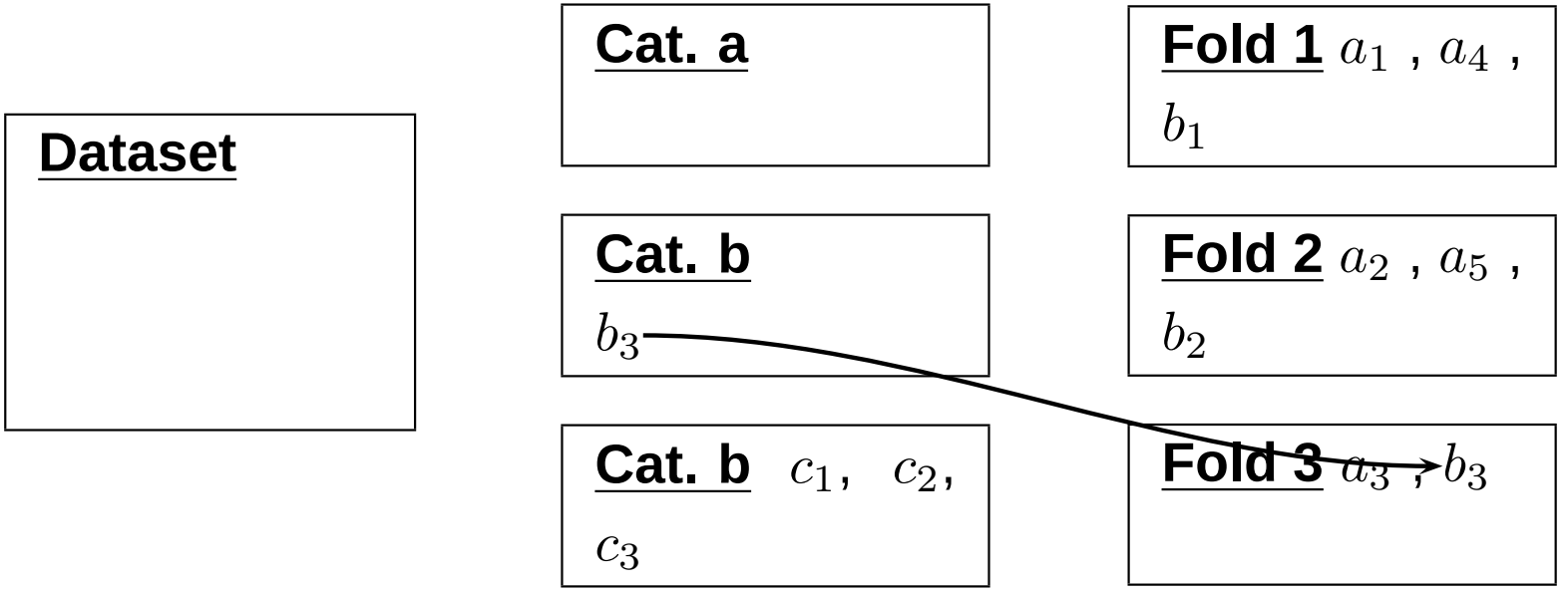
Stratified cross validation



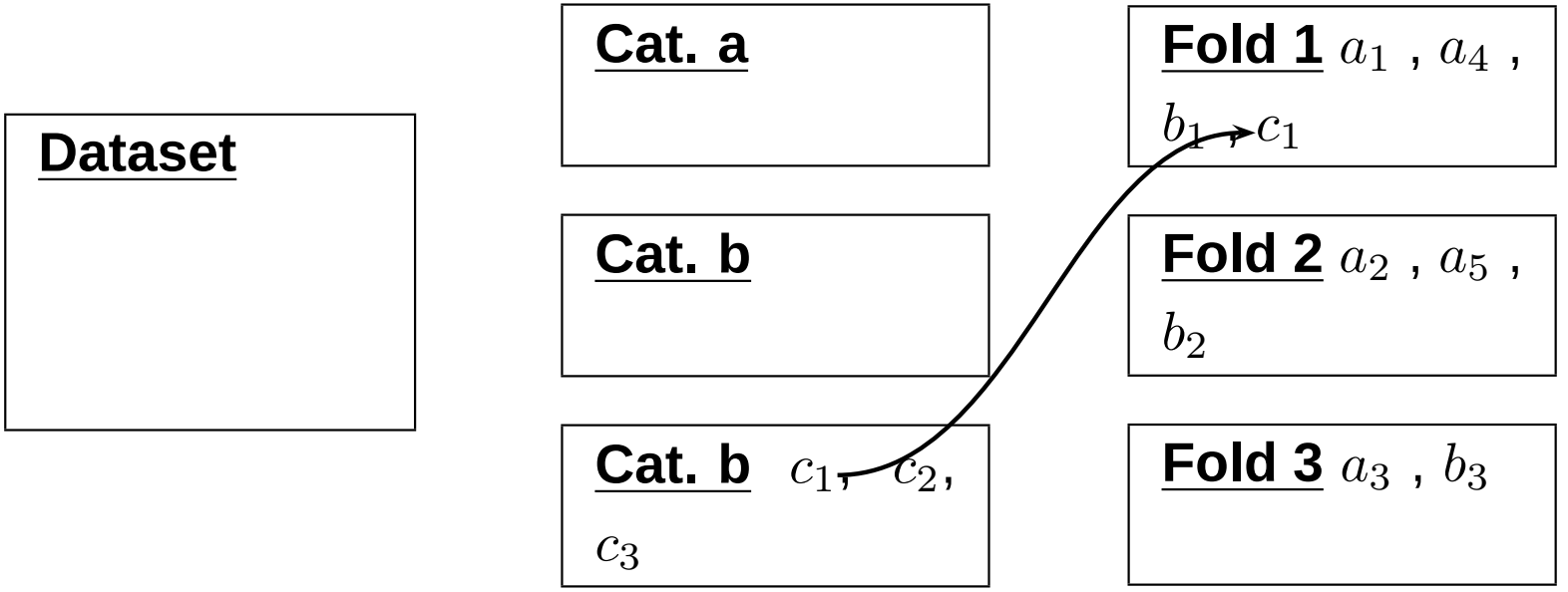
Stratified cross validation



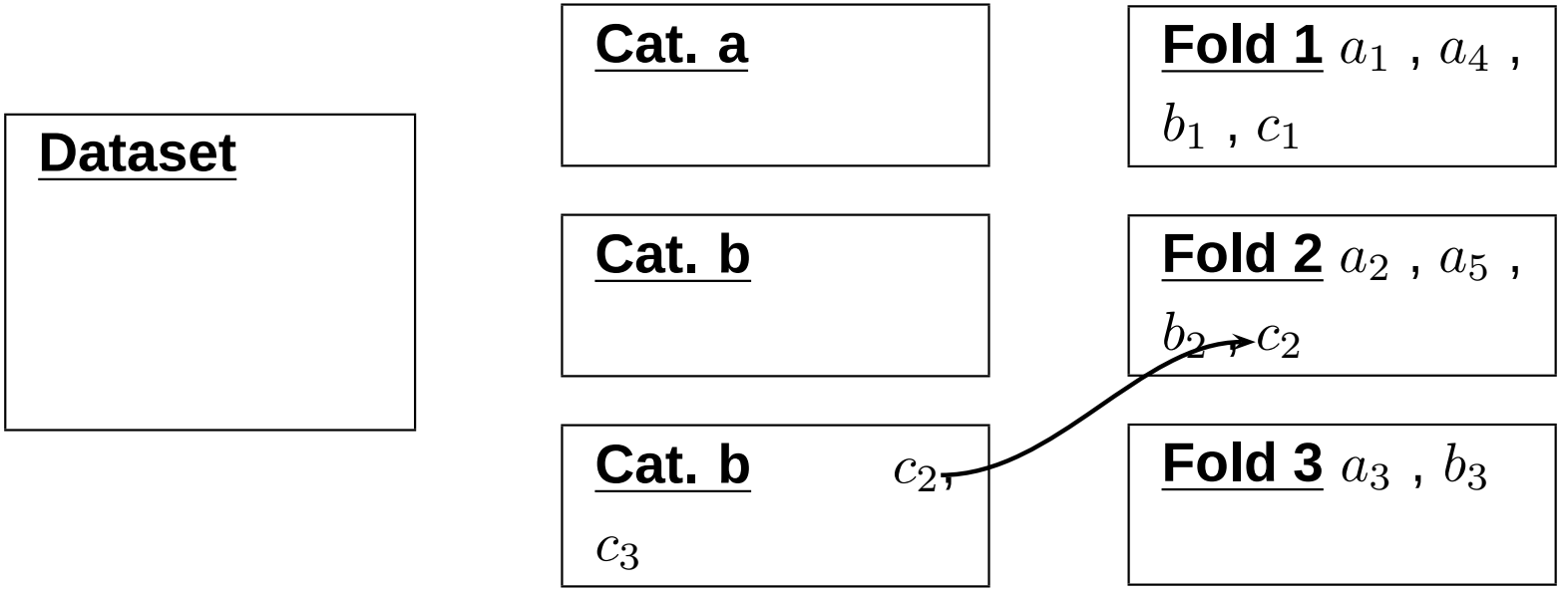
Stratified cross validation



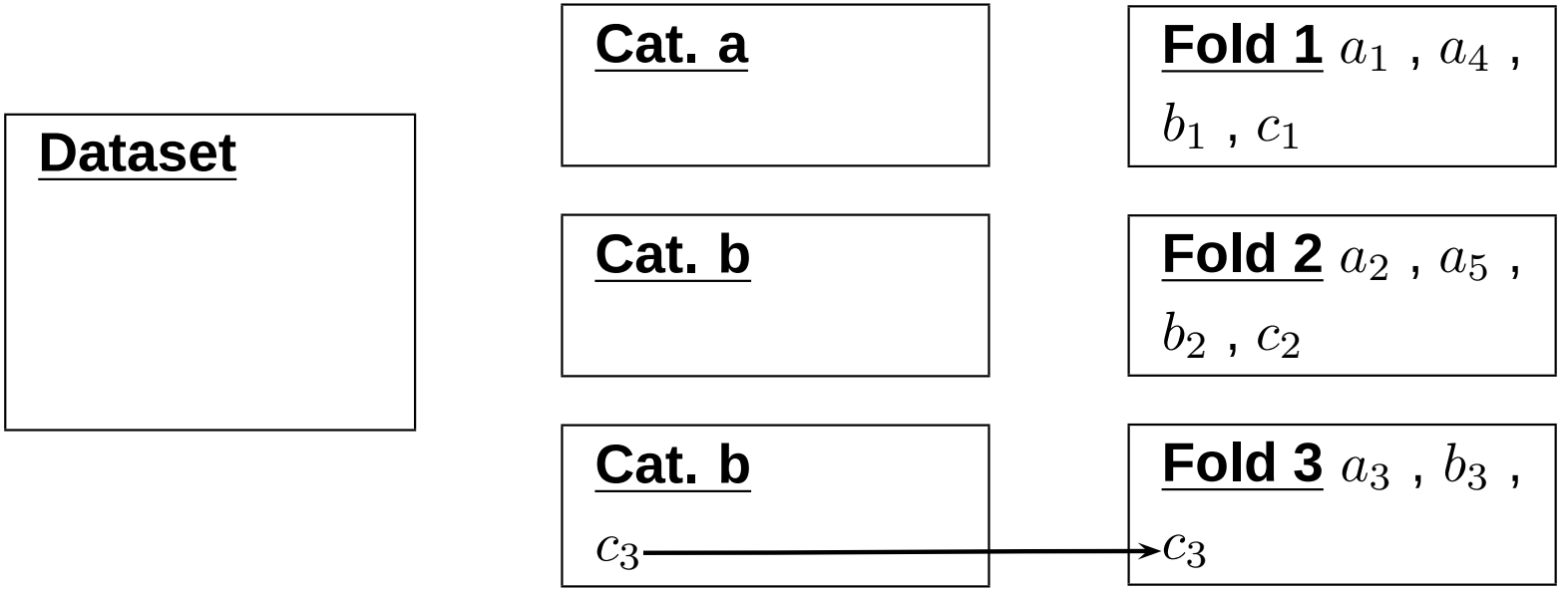
Stratified cross validation



Stratified cross validation



Stratified cross validation



Stratified cross validation

Dataset

Cat. a

Cat. b

Cat. b

Fold 1 a_1 , a_4 , b_1 , c_1

Fold 2 a_2 , a_5 , b_2 , c_2

Fold 3 a_3 , b_3 , c_3

Stratified cross validation

$$\begin{cases} \forall\, e\ example, \exists\ f\ fold, e \in f.test \\ \forall f_1, f_2\ folds, f_1 \neq f_2 \Rightarrow f_1.test \cap f_2.test = \emptyset \end{cases}$$

Presentation

- Hypertext Classification
- Related Work
- Our Model
- Implementation
- **Results**
- Conclusion

Presentation

- **Results**
  - **Evaluation of a single hypothesis**
  - Comparing several hypothesis
  - Results
    - One pattern
    - Combining two features
    - Meta Predecessor and Hyperlink Ensembles
    - Binarization: One against all or Round Robin
    - Merge or Tag
    - Combination of the features

Evaluation of a single hypothesis

Category c_i	Classified as positive	Classified as negative
Is positive	a	b
Is negative	c	d

- **Accuracy** $A = \frac{a+d}{a+b+c+d}$
- **Precision** $\pi = \frac{a}{a+c}$
- **Recall** $\rho = \frac{a}{a+b}$
- **F**$_\beta$ $F_\beta = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi+\rho}$

$$\lim_{\beta\to\infty}(F_\beta) = \pi$$

Macro Averaging

	as x	as y	as z
is x	1213	1	1
is y	352	33	0
is z	421	1	41

	as x	as !x
is x	1213	2
is !x	773	75

	as y	as !y
is y	33	352
is !y	2	1676

	as z	as !z
is z	41	422
is !z	1	1599

$$\pi_x = 0.61$$
$$\pi_x = 0.94$$
$$\pi_x = 0.98$$
$$\pi_{macro} = 0.84$$

Micro Averaging

	as x	as y	as z
is x	1213	1	1
is y	352	33	0
is z	421	1	41

	as +	as -
is +	1287	776
is -	776	3350

$$\pi_{micro} = 0.62$$

Micro Fold Averaging

Compute the macro-averaging contingency table for each fold, sum them and compute the precision.

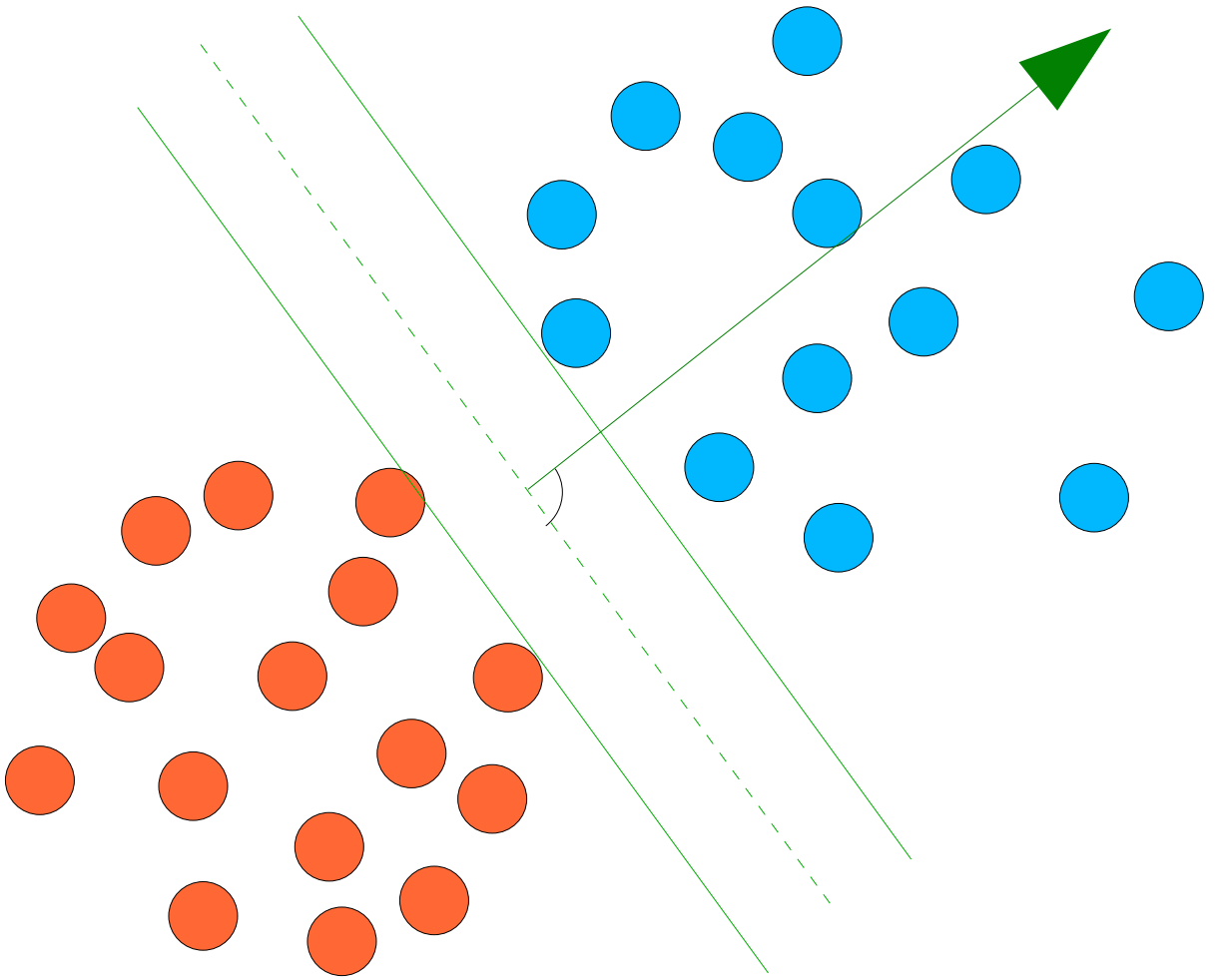
Choice of the evaluation function

- Finding documents relevant documents
  - Find all the relevant documents
  - Retrieved only relevant documents
- Huge number of Web documents
- Web crawlers index only a small subset of the Web
- **Precision**
- Don’t emphasize the WebKB hold all category `other`
- **Macro precision**

Presentation

- **Results**
  - Evaluation of a single hypothesis
  - **Comparing several hypothesis**
  - Results
    - One pattern
    - Combining two features
    - Meta Predecessor and Hyperlink Ensembles
    - Binarization: One against all or Round Robin
    - Merge or Tag
    - Combination of the features

Pattern ranking



Pattern ranking

Decision function

$$D(\vec{x}) = sign(\vec{w} \cdot \vec{x} + b)$$

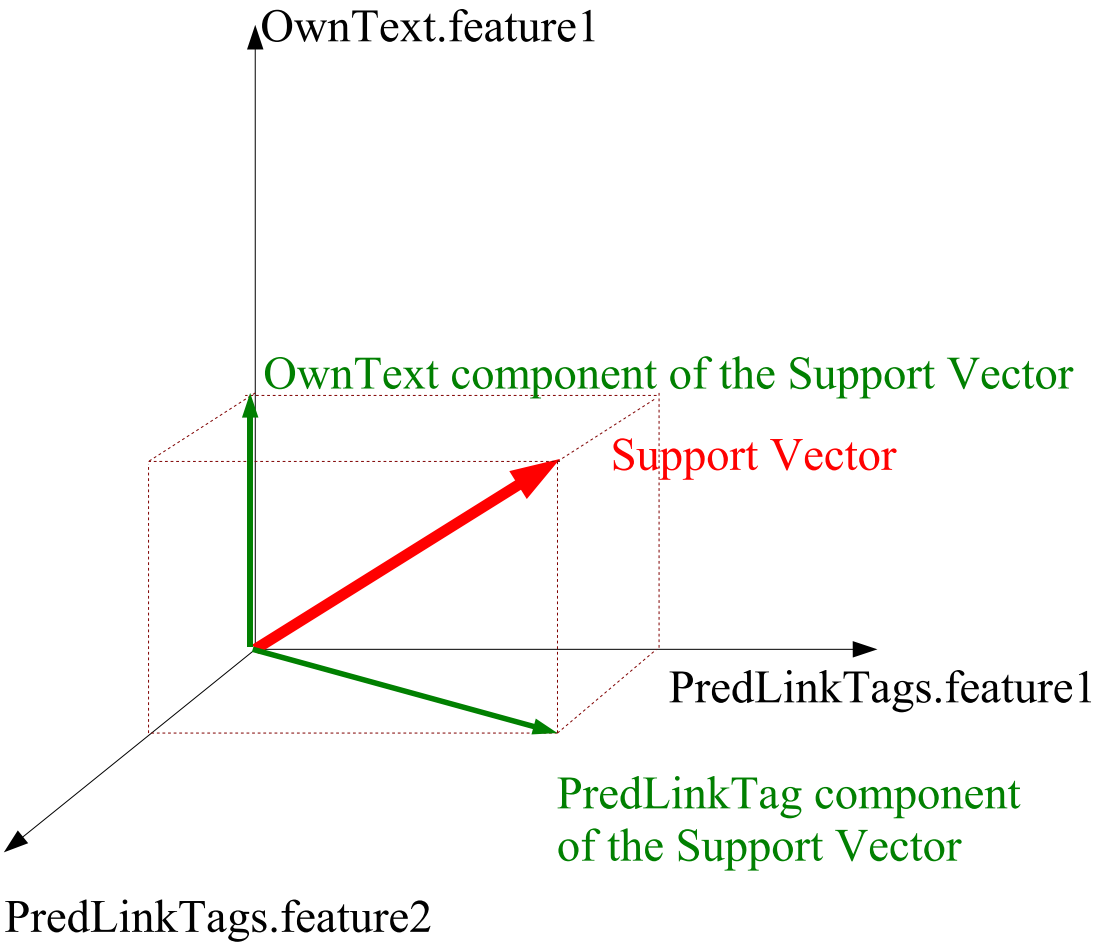
$\vec{w}$: orthogonal vector of the separation hyperplane

$$\vec{w} = \sum_{i=0}^{k-1} w_i \vec{j_i}$$

With $(\vec{j_i})_{i=0}^{k-1}$ orthonormal base

The bigger the component w_k of the vector $\vec{w}$, the stronger the influence of feature k on the classification.

Pattern ranking



Pattern ranking

$E = M_1 \oplus M_2 \oplus \cdots \oplus M_n$, with $\begin{cases} E & \text{the global vector space} \\ n & \text{number of mining methods} \\ M_i & \text{the subsets of features mined by the } i^{th} \text{ metho} \end{cases}$

$\vec{w}$ in $(\vec{w_i})_{i=1}^{n}$, with $\vec{w} = \sum_{i=1}^{n} \vec{w_i}$, $\forall i \in [1, n]\,, \vec{w_i} \in M_i$

Efficiency estimators

feature estimator $e_f(m) = \frac{e_g(m)}{|M|}$ Average information brought by one feature mined by the method m

mining method estimator $e_g(m) = |\vec{w_m}| = \sqrt{\sum_{f \in M} w_f^2}$ Information brought by all the features mined by the method m

Pattern ranking

Feature	*# features*
PredLinkParagraph	79588
PredNWordsAroundLink	41513
OwnText	37898
PredHeadings	32832
PredLinkTags	4211
PredListHeadings	4118

Pattern ranking

Feature	*Method component length*
PredLinkParagraph	51831
PredNWordsAroundLink	14360
PredHeadings	13070
OwnText	12658
PredListHeadings	4319
PredLinkTags	2594

pattern ranking

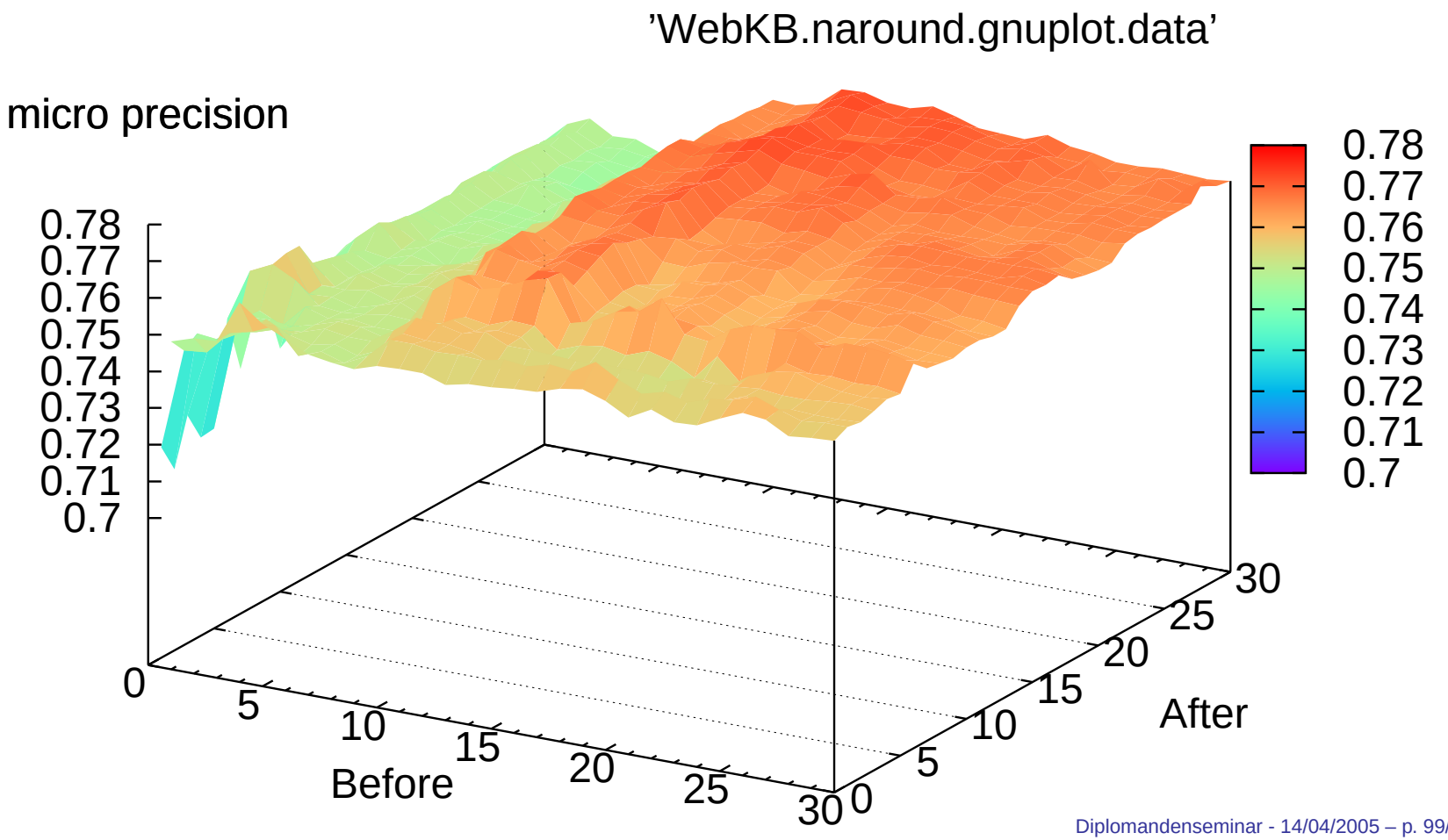
Feature	*average feature length (method component length/features count)*
PredListHeadings	1.05
PredNWordsAroundLink	0.65
PredLinkTags	0.62
PredHeadings	0.40
PredLinkParagraph	0.35
OwnText	0.33

Pattern ranking

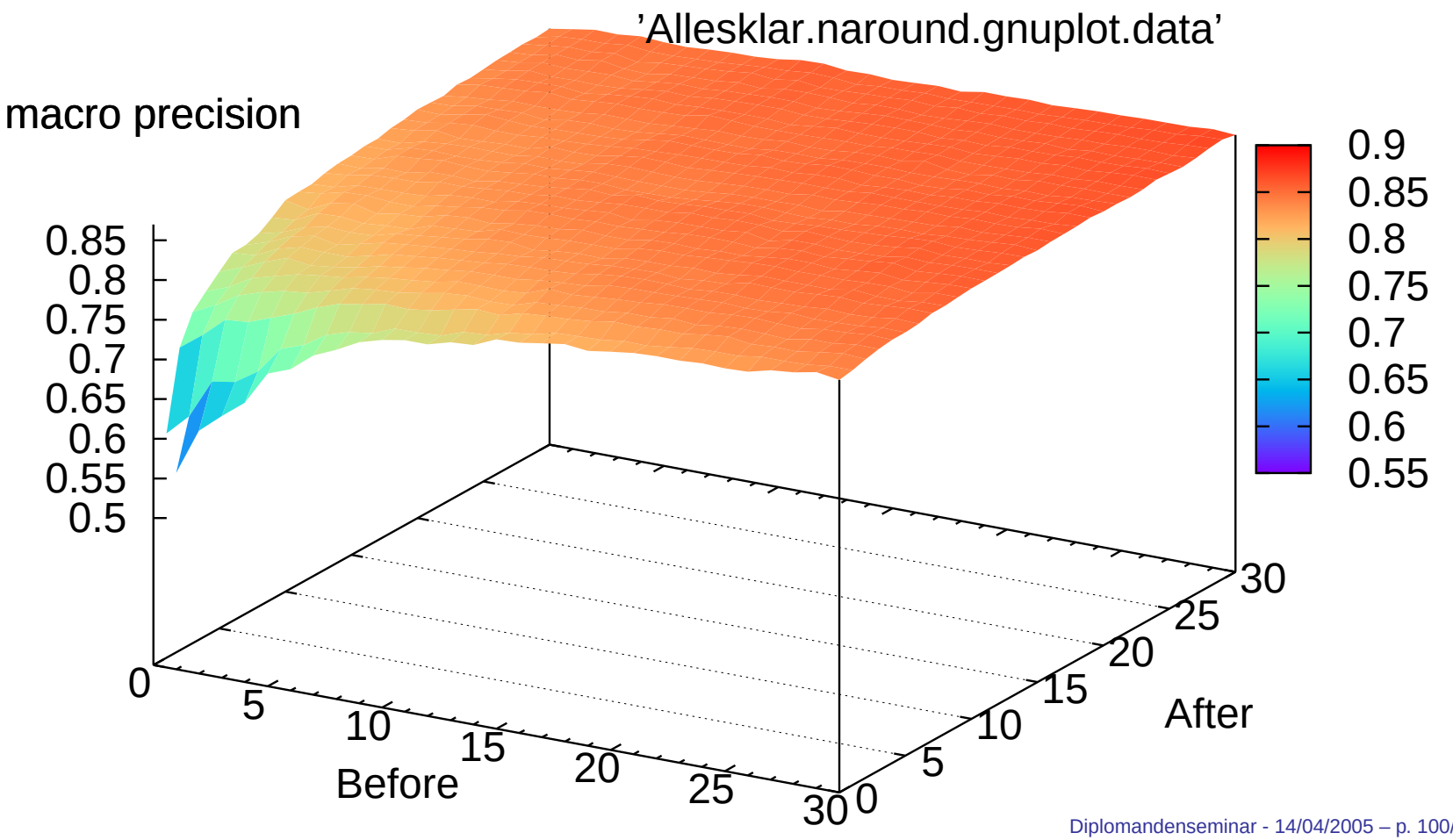
- PredLinkParagraph mines many features, but they and rather spurious
- Owntext is not targeted and thus mines spurious words

Neighborhood of the anchor

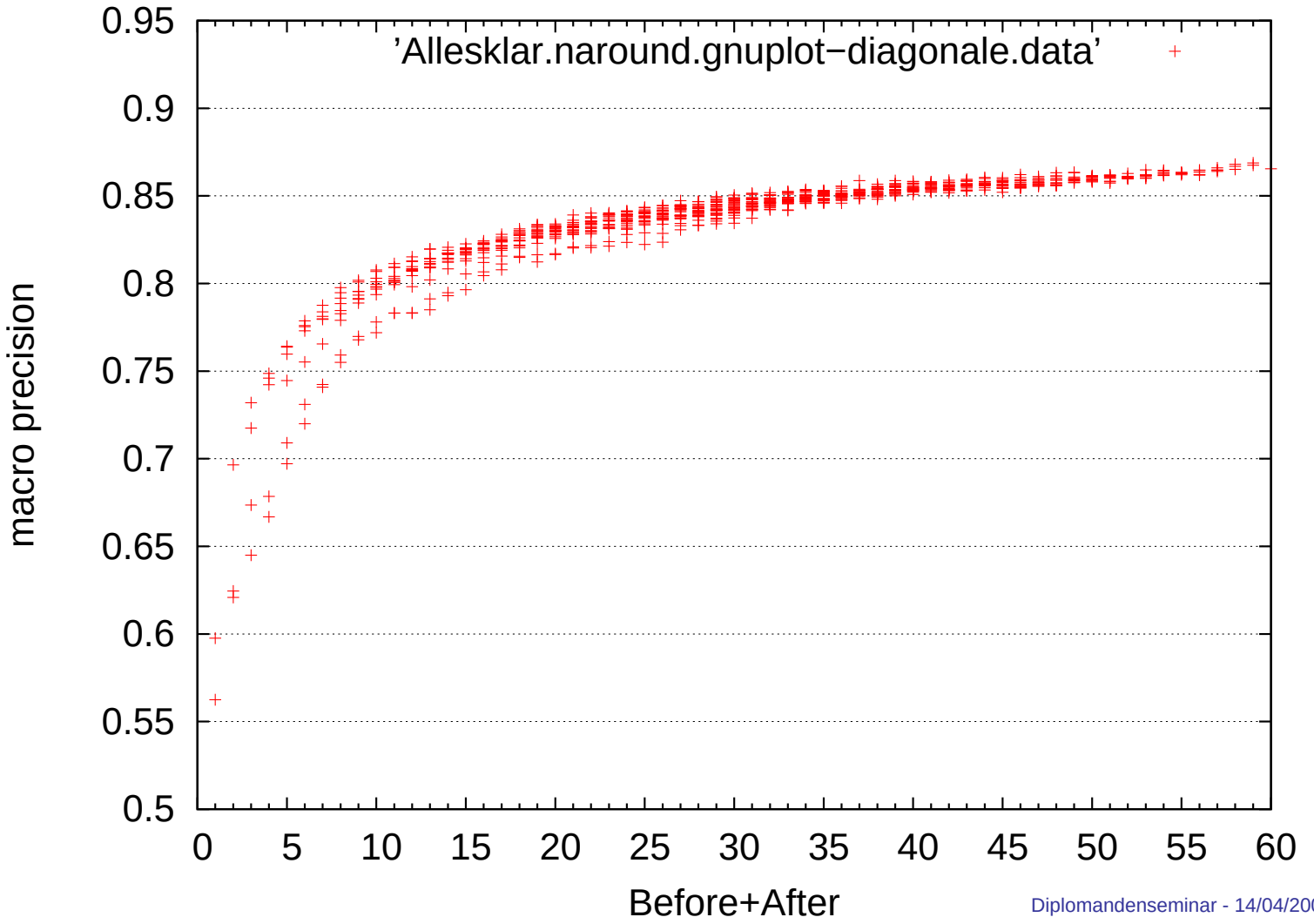
WebKB


'WebKB.naround.gnuplot.data'
micro precision
0.78
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.7
0
5
10
15
20
25
30
Before
After

Allesklar


'Allesklar.naround.gnuplot.data'
macro precision
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
5
10
15
20
25
30
Before
0
5
10
15
20
25
30
After
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55

Allesklar


'Allesklar.naround.gnuplot−diagonale.data'
macro precision
Before+After
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
5
10
15
20
25
30
35
40
45
50
55
60

Neighborhood of the anchor

$$precision(Words) \approx precision(Before, After)$$

, with $Words = After + Before$

Presentation

- **Results**
  - Evaluation of a single hypothesis
  - Comparing several hypothesis
  - **Results**
    - **One pattern**
    - Combining two features
    - Meta Predecessor and Hyperlink Ensembles
    - Binarization: One against all or Round Robin
    - Merge or Tag
    - Combination of the features

Using one feature

	Allesklar	WebKB
Words Around	83.40% 3664	39.49% 3007
Pred LinkTags	67.80% 3653	33.62% 2941
PredList Headings	51.57% 1870	21.78% 1644
Pred Headings	54.49% 2672	22.65% 2828
PredLink Paragraph	66.90% 2715	23.43% 1144
Own Text	58.15% 3831	40.96% 8277

Using one feature

	Allesklar	WebKB
Words Around	83.40% 3664	39.49% 3007
Pred LinkTags	67.80% 3653	33.62% 2941
PredList Headings	51.57% 1870	21.78% 1644
Pred Headings	54.49% 2672	22.65% 2828
PredLink Paragraph	66.90% 2715	23.43% 1144
Own Text	58.15% 3831	40.96% 8277

OwnText covers more examples than the other patterns. Non-local features are more often mined on Allesklar than on WebKB.

Using one feature

	Allesklar	WebKB
Words Around	83.40% 3664	39.49% 3007
Pred LinkTags	67.80% 3653	33.62% 2941
PredList Headings	51.57% 1870	21.78% 1644
Pred Headings	54.49% 2672	22.65% 2828
PredLink Paragraph	66.90% 2715	23.43% 1144
Own Text	58.15% 3831	40.96% 8277

The good connectivity of Allesklar confers to WordsAround and PredLinkTags a fast as good coverage as OwnText

Using one feature

	Allesklar	WebKB
Words Around	83.40% 3664	39.49% 3007
Pred LinkTags	67.80% 3653	33.62% 2941
PredList Headings	51.57% 1870	21.78% 1644
Pred Headings	54.49% 2672	22.65% 2828
PredLink Paragraph	66.90% 2715	23.43% 1144
Own Text	58.15% 3831	40.96% 8277

The slight coverage difference between WordsAround and PredlinkTags shows that not all the anchors have a description.

Using one feature

	Allesklar	WebKB
Words Around	83.40% 3664	39.49% 3007
Pred LinkTags	67.80% 3653	33.62% 2941
PredList Headings	51.57% 1870	21.78% 1644
Pred Headings	54.49% 2672	22.65% 2828
PredLink Paragraph	66.90% 2715	23.43% 1144
Own Text	58.15% 3831	40.96% 8277

PredListHeadings is difficult to mine because of its double condition.

Using one feature

	Allesklar	WebKB
Words Around	83.40% 3664	39.49% 3007
Pred LinkTags	67.80% 3653	33.62% 2941
PredList Headings	51.57% 1870	21.78% 1644
Pred Headings	54.49% 2672	22.65% 2828
PredLink Paragraph	66.90% 2715	23.43% 1144
Own Text	58.15% 3831	40.96% 8277

The classifier based on the neighborhood of the link outperforms the traditional text classifier by over 43%

Presentation

- **Results**
  - Evaluation of a single hypothesis
  - Comparing several hypothesis
  - **Results**
    - One pattern
    - Combining two features
    - **Meta Predecessor and Hyperlink Ensembles**
    - Binarization: One against all or Round Robin
    - Merge or Tag
    - Combination of the features

Combining two features

- Antagonist effects
  - increases the amount of information
  - increases the dimensionality of the classification problem
  - increases the number of examples to classify
- Combination helpful with
  - Nigh precisions
  - Disjunct patterns

Allesklar

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	85.83% 3664	85.63% 3678	**86.19%** 3665	85.44% 3665	84.41% 3667	83.26% 3898
Pred LinkTags	85.63% 3678	70.29% 3653	**71.96%** 3653	68.92% 3653	**71.63%** 3655	**72.8%** 3898
PredList Headings	**86.19%** 3665	**71.96%** 3653	52.68% 1870	56.74% 2744	65.64% 3013	**67.94%** 3864
Pred Headings	85.44% 3665	68.92% 3653	56.74% 2744	57.6% 2672	66.55% 3103	**69.87%** 3879
PredLink Paragraph	84.41% 3667	**71.63%** 3655	65.64% 3013	66.55% 3103	68.9% 2715	**70.53%** 3882
Own Text	83.26% 3898	**72.8%** 3898	**67.94%** 3864	**69.87%** 3879	**70.53%** 3882	65.72% 3831

WebKB

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	39.49% 3007	**46.62%** 3017	36.19% 3008	36.05% 3017	**42.35%** 3012	**41.73%** 8277
Pred LinkTags	**46.62%** 3017	33.62% 2941	24.22% 2942	30.5% 3002	32.86% 2955	**41.68%** 8277
PredList Headings	36.19% 3008	24.22% 2942	21.78% 1644	**25.7%** 2832	**28.91%** 2403	40.72% 8277
Pred Headings	36.05% 3017	30.5% 3002	**25.7%** 2832	22.65% 2828	**26.15%** 2912	40.72% 8277
PredLink Paragraph	**42.35%** 3012	32.86% 2955	**28.91%** 2403	**26.15%** 2912	23.43% 1144	40.92% 8277
Own Text	**41.73%** 8277	**41.68%** 8277	40.72% 8277	40.72% 8277	40.92% 8277	40.96% 8277

Presentation

- **Results**
  - Evaluation of a single hypothesis
  - Comparing several hypothesis
  - **Results**
    - One pattern
    - Combining two features
    - **Meta Predecessor and Hyperlink Ensembles**
    - Binarization: One against all or Round Robin
    - Merge or Tag
    - Combination of the features

Meta Predecessor and Hyperlink Ensembles

Allesklar

Hyperlink Ensembles Meta Predecessor

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	72.49% **85.83%** 3664	71.91% **85.63%** 3678	71.9% **86.19%** 3665	61.54% **85.44%** 3665	70.33% **84.41%** 3667	67.15% **83.26%** 3898
Pred LinkTags	71.91% **85.63%** 3678	61.26% **70.29%** 3653	63.3% **71.96%** 3653	57.15% **68.92%** 3653	59.51% **71.63%** 3655	60.17% **72.8%** 3898
PredList Headings	71.9% **86.19%** 3665	63.3% **71.96%** 3653	47.29% **52.68%** 1870	46.38% **56.74%** 2744	54.07% **65.64%** 3013	63.66% **67.94%** 3864
Pred Headings	61.54% **85.44%** 3665	57.15% **68.92%** 3653	46.38% **56.74%** 2744	48.27% **57.6%** 2672	47.69% **66.55%** 3103	58.2% **69.87%** 3879
PredLink Paragraph	70.33% **84.41%** 3667	59.51% **71.63%** 3655	54.07% **65.64%** 3013	47.69% **66.55%** 3103	58.23% **68.9%** 2715	60.84% **70.53%** 3882
Own Text	67.15% **83.26%** 3898	60.17% **72.8%** 3898	63.66% **67.94%** 3864	58.2% **69.87%** 3879	60.84% **70.53%** 3882	65.72% 65.72% 3831

WebKB

Hyperlink Ensembles Meta Predecessors

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	**40%** 39.49% 3007	**53.7%** 52.04% 3017	30.19% **35.66%** 3008	25.99% **36.05%** 3017	36.94% **37.96%** 3012	38.37% **41.73%** 8277
Pred LinkTags	**53.7%** 52.04% 3017	**38.79%** 33.62% 2941	**41.04%** 35.65% 2942	**36.39%** 30.5% 3002	**35.97%** 30.13% 2955	37.21% **41.68%** 8277
PredList Headings	30.19% **35.66%** 3008	**41.04%** 35.65% 2942	**24.1%** 21.78% 1644	**26.62%** 25.7% 2832	**28.21%** 23.48% 2403	39.62% **40.72%** 8277
Pred Headings	25.99% **36.05%** 3017	**36.39%** 30.5% 3002	**26.62%** 25.7% 2832	**26.61%** 22.65% 2828	25.14% **26.15%** 2912	34.87% **40.72%** 8277
PredLink Paragraph	36.94% **37.96%** 3012	**35.97%** 30.13% 2955	**28.21%** 23.48% 2403	25.14% **26.15%** 2912	**27.73%** 23.43% 1144	**41.01%** 40.92% 8277
Own Text	38.37% **41.73%** 8277	37.21% **41.68%** 8277	39.62% **40.72%** 8277	34.87% **40.72%** 8277	**41.01%** 40.92% 8277	40.96% 40.96% 8277

Hyperlink Ensembles

- Disappointing result
- Apparent contradiction with Chakrabarti's and Getoor's results
- Reason for this poor efficiency
  - Contradiction between the feature sets sizes and the dimensionality of the learning problem

Hyperlink Ensembles

Solutions

- Learn on Meta Predecessors to enlarge the feature sets
- Reduce the dimensionality of the problem
  - Stems
  - Synonyms
  - Abbreviations expansion

Allesklar

Learns with MP, classifies with HE Learns and classifies with MP

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	**87.49%** 85.83% 3664	75.64% **86.46%** 3678	74.32% **85.85%** 3665	69.64% **85.42%** 3665	71.93% **85.28%** 3667	68.66% **83.26%** 3898
Pred LinkTags	75.64% **86.46%** 3678	**71.44%** 70.29% 3653	58.11% **72.91%** 3653	55.04% **68.92%** 3653	56.9% **72.75%** 3655	57.77% **72.8%** 3898
PredList Headings	74.32% **85.85%** 3665	58.11% **72.91%** 3653	51.53% **52.68%** 1870	40.61% **56.74%** 2744	47.27% **66.38%** 3013	62.71% **67.94%** 3864
Pred Headings	69.64% **85.42%** 3665	55.04% **68.92%** 3653	40.61% **56.74%** 2744	**58.83%** 57.6% 2672	39.57% **66.54%** 3103	68.93% **69.87%** 3879
PredLink Paragraph	71.93% **85.28%** 3667	56.9% **72.75%** 3655	47.27% **66.38%** 3013	39.57% **66.54%** 3103	**69.6%** 68.9% 2715	70.1% **70.54%** 3882
Own Text	68.66% **83.26%** 3898	57.77% **72.8%** 3898	62.71% **67.94%** 3864	68.93% **69.87%** 3879	70.1% **70.54%** 3882	65.72% 65.72% 3831

WebKB

WebKB-HE-LT-TS-merged-oneagainstall WebKB-MP-merge-oneagainstall

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	39.19% **39.49%** 3007	44.78% **46.62%** 3017	34.84% **36.19%** 3008	22.03% **36.05%** 3017	41.22% **42.35%** 3012	39.54% **41.73%** 8277
Pred LinkTags	44.78% **46.62%** 3017	30.35% **33.62%** 2941	**24.63%** 24.22% 2942	29.29% **30.5%** 3002	**33.22%** 32.86% 2955	40.44% **41.68%** 8277
PredList Headings	34.84% **36.19%** 3008	**24.63%** 24.22% 2942	21.46% **21.78%** 1644	25.21% **25.7%** 2832	**29.77%** 28.91% 2403	39.54% **40.72%** 8277
Pred Headings	22.03% **36.05%** 3017	29.29% **30.5%** 3002	25.21% **25.7%** 2832	**26.47%** 22.65% 2828	25.41% **26.15%** 2912	39.24% **40.72%** 8277
PredLink Paragraph	41.22% **42.35%** 3012	**33.22%** 32.86% 2955	**29.77%** 28.91% 2403	25.41% **26.15%** 2912	23.33% **23.43%** 1144	40.05% **40.92%** 8277
Own Text	39.54% **41.73%** 8277	40.44% **41.68%** 8277	39.54% **40.72%** 8277	39.24% **40.72%** 8277	40.05% **40.92%** 8277	40.96% 40.96% 8277

Presentation

- **Results**
  - Evaluation of a single hypothesis
  - Comparing several hypothesis
  - **Results**
    - One pattern
    - Combining two features
    - Meta Predecessor and Hyperlink Ensembles
    - **Binarization: One against all or Round Robin**
    - Merge or Tag
    - Combination of the features

Binarization

Allesklar-MP-notmerged-oneagainstall Allesklar-MP-notmerged-roundrobin

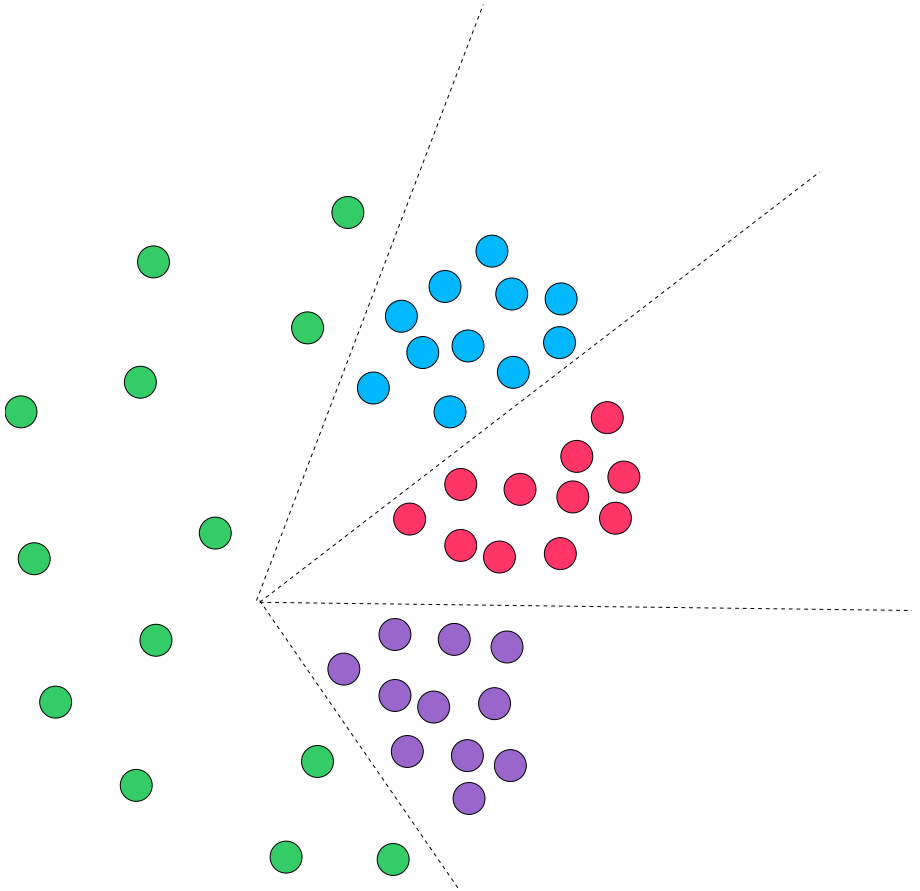
	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	**85.83%** 83.4% 3664	**85.63%** 83.47% 3678	**86.19%** 83.28% 3665	**85.44%** 81.87% 3665	**84.41%** 79.89% 3667	**83.26%** 79.04% 3898
Pred LinkTags	**85.63%** 83.47% 3678	**70.29%** 67.8% 3653	**71.96%** 68.11% 3653	**68.92%** 64.67% 3653	**71.63%** 67.31% 3655	**72.8%** 65.47% 3898
PredList Headings	**86.19%** 83.28% 3665	**71.96%** 68.11% 3653	**52.68%** 51.57% 1870	**56.74%** 56.61% 2744	**65.64%** 61.19% 3013	**67.94%** 61.88% 3864
Pred Headings	**85.44%** 81.87% 3665	**68.92%** 64.67% 3653	**56.74%** 56.61% 2744	**57.6%** 54.49% 2672	**66.55%** 59.52% 3103	**69.87%** 63.17% 3879
PredLink Paragraph	**84.41%** 79.89% 3667	**71.63%** 67.31% 3655	**65.64%** 61.19% 3013	**66.55%** 59.52% 3103	**68.9%** 66.9% 2715	**70.53%** 63.03% 3882
Own Text	**83.26%** 79.04% 3898	**72.8%** 65.47% 3898	**67.94%** 61.88% 3864	**69.87%** 63.17% 3879	**70.53%** 63.03% 3882	**65.72%** 58.15% 3831

Binarization

WebKB-MP-notmerged-oneagainstall WebKB-MP-notmerged-roundrobin

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	**39.49%** 39.42% 3007	**52.04%** 51.79% 3017	**35.66%** 31.99% 3008	36.05% **38.7%** 3017	**37.96%** 34.82% 3012	41.73% **42.22%** 8277
Pred LinkTags	**52.04%** 51.79% 3017	**33.62%** 32.1% 2941	35.65% **37.43%** 2942	30.5% **32.72%** 3002	30.13% **31.41%** 2955	41.68% **42.46%** 8277
PredList Headings	**35.66%** 31.99% 3008	35.65% **37.43%** 2942	21.78% **22.67%** 1644	**25.7%** 22.65% 2832	23.48% **25.03%** 2403	40.72% **41.65%** 8277
Pred Headings	36.05% **38.7%** 3017	30.5% **32.72%** 3002	**25.7%** 22.65% 2832	22.65% **24.89%** 2828	26.15% **26.72%** 2912	40.72% **41.54%** 8277
PredLink Paragraph	**37.96%** 34.82% 3012	30.13% **31.41%** 2955	23.48% **25.03%** 2403	26.15% **26.72%** 2912	23.43% **26.1%** 1144	40.92% **41.31%** 8277
Own Text	41.73% **42.22%** 8277	41.68% **42.46%** 8277	40.72% **41.65%** 8277	40.72% **41.54%** 8277	40.92% **41.31%** 8277	40.96% **41.88%** 8277

Sticky classes



Sticky classes

- class that is not as specific as the others
- most populated class
- hold on class

Sticky classes of Allesklar

source Allesklar-MP-notmerged-roundrobin/Allesklar-0-0-predlinktags-.

1 : Gesellschaft-Politik
2 : Bildung-Wissenschaft
3 : Immobilien-Wohnen
4 : Freizeit-Lifestyle
5 : Arbeit-Beruf

	as 1	as 2	as 3	as 4	as 5	recall	F1
is 1	584	84	126	16	8	0.713	0.643
is 2	132	504	122	24	10	0.636	0.649
is 3	115	52	525	42	12	0.703	0.587
is 4	69	55	138	474	5	0.639	0.7
is 5	95	65	128	55	213	0.383	0.529
Prec.	0.586	0.663	0.505	0.775	0.858		

macro_precision : 0.678006900282544

Presentation

- **Results**
  - Evaluation of a single hypothesis
  - Comparing several hypothesis
  - **Results**
    - One pattern
    - Combining two features
    - Meta Predecessor and Hyperlink Ensembles
    - Binarization: One against all or Round Robin
    - **Merge or Tag**
    - Combination of the features

Allesklar

Allesklar-MP-merged-oneagainstall Allesklar-MP-notmerged-oneagainstall

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	85.83% 85.83% 3664	**86.46%** 85.63% 3678	85.85% **86.19%** 3665	85.42% **85.44%** 3665	**85.28%** 84.41% 3667	83.26% 83.26% 3898
Pred LinkTags	**86.46%** 85.63% 3678	70.29% 70.29% 3653	**72.91%** 71.96% 3653	68.92% 68.92% 3653	**72.75%** 71.63% 3655	72.8% 72.8% 3898
PredList Headings	85.85% **86.19%** 3665	**72.91%** 71.96% 3653	52.68% 52.68% 1870	**56.74%** 56.74% 2744	**66.38%** 65.64% 3013	67.94% 67.94% 3864
Pred Headings	85.42% **85.44%** 3665	68.92% 68.92% 3653	**56.74%** 56.74% 2744	57.6% 57.6% 2672	66.54% **66.55%** 3103	69.87% 69.87% 3879
PredLink Paragraph	**85.28%** 84.41% 3667	**72.75%** 71.63% 3655	**66.38%** 65.64% 3013	66.54% **66.55%** 3103	68.9% 68.9% 2715	**70.54%** 70.53% 3882
Own Text	83.26% 83.26% 3898	72.8% 72.8% 3898	67.94% 67.94% 3864	69.87% 69.87% 3879	**70.54%** 70.53% 3882	65.72% 65.72% 3831

WebKB

WebKB-MP-merge-oneagainstall WebKB-MP-notmerged-oneagainstall

	Words Around	Pred LinkTags	PredList Headings	Pred Headings	PredLink Paragraph	Own Text
Words Around	39.49% 39.49% 3007	46.62% **52.04%** 3017	**36.19%** 35.66% 3008	36.05% 36.05% 3017	**42.35%** 37.96% 3012	41.73% 41.73% 8277
Pred LinkTags	46.62% **52.04%** 3017	33.62% 33.62% 2941	24.22% **35.65%** 2942	30.5% 30.5% 3002	**32.86%** 30.13% 2955	41.68% 41.68% 8277
PredList Headings	**36.19%** 35.66% 3008	24.22% **35.65%** 2942	21.78% 21.78% 1644	25.7% 25.7% 2832	**28.91%** 23.48% 2403	40.72% 40.72% 8277
Pred Headings	36.05% 36.05% 3017	30.5% 30.5% 3002	25.7% 25.7% 2832	22.65% 22.65% 2828	26.15% 26.15% 2912	40.72% 40.72% 8277
PredLink Paragraph	**42.35%** 37.96% 3012	**32.86%** 30.13% 2955	**28.91%** 23.48% 2403	26.15% 26.15% 2912	23.43% 23.43% 1144	40.92% 40.92% 8277
Own Text	41.73% 41.73% 8277	41.68% 41.68% 8277	40.72% 40.72% 8277	40.72% 40.72% 8277	40.92% 40.92% 8277	40.96% 40.96% 8277

Merge or Tag ?

- Merging outperforms Tagging
  - when the feature patterns may mine the same features
  - It reinforces the weight of a feature mined by two different patterns
- Tagging outperforms Merging
  - when the feature patterns mine on disjunct fields of the data
  - when the patterns mine features of similar purities

Presentation

- **Results**
  - **Evaluation of a single hypothesis**
  - Comparing several hypothesis
  - **Results**
    - One pattern
    - Combining two features
    - Meta Predecessor and Hyperlink Ensembles
    - Binarization: One against all or Round Robin
    - Merge or Tag
    - **Combination of the features**

Combination of the features

- Anchor group
  - PredLinkTags
  - WordsAround
- Headings group
  - PredHeadings
  - PredListHeadings
- Simple words group
  - OwnText
  - PredLinkParagraph

Combination of the features

- Allesklar
  - precision:85.46%
  - precision of the text only classifier:65.72% (+30.5%)
  - of Around Anchor alone: 85.83% (-0.5%)
- WebKB
  - precision:84.73%
  - precision of the text only classifier:40.96% (+106.9%)
  - of Around Anchor alone: 39.49% (+114.6%)

Presentation

- Hypertext Classification
- Related Work
- Our Model
- Implementation
- Results
- **Conclusion**

Conclusion

- We proposed a model of hypertext classifiers based on both local and non-local features.
- Our model outperforms by up to 115% traditional text classifiers.

Conclusion

- Despite negative results with Hyperlink ensembles, we believe that this model would outperform the Meta Predecessor with a powerful dimension reduction.
- The Round Robin binarization should prevail when the problem doesn't contain a sticky class.
- The best method between Merging and Tagging depends on the nature of the features to mutualize.

Vielen Dank für Ihre Aufmerksamkeit !!!