Separate and Conquer Framework und disjunktive Regeln

Matthias Thiel

Überblick

- Begriffe und Notation
- CN2 und BEXA
- Ausgangsproblem als Motivation für SeCo-Framework
- Beschreibung des SeCo-Framework
- Fallstudie: Hypothesensprache

Ein Lernproblem

Zu erlernender Begriff:
“Es wird morgen nicht mehr regnen.”

Attribute

Name	**Type**	**Domain**
outlook	nominal	{Sunny, overcast, rain}
autumn	nominal	{yes,no}
temp	linear	{15..35}

Extension/Umfang

$X_{TS}(sunny)=\{1,4,5,10,12\}$

Training Set

#	**outlook**	**autumn**	**temp**	**class**
1	sunny	yes	17	-
2	overcast	no	18	-
3	rain	yes	16	-
4	sunny	yes	22	-
5	sunny	no	29	-
6	overcast	yes	30	-
7	overcast	no	35	-
8	rain	yes	23	-
9	rain	no	27	-
10	sunny	yes	28	+
11	overcast	no	23	+
12	sunny	no	27	+
13	rain	no	23	+

Klassische Verfeinerung (CN2)

Beginne mit $[\,]\Rightarrow +$

$X_{TS}()=\{1...13\}$

Wähle Attribut *autumn=no*

$X_{TS}(autumn=no)=\{2,5,7,9,11,12,13\}$

Wähle Attribut temp=23

$X_{TS}(no,23)=\{11,13\}$

Gelernte Regel

$[autumn=no][temp=23]\Rightarrow +$

Training Set

#	outlook	autumn	temp	class
1	sunny	yes	17	-
2	overcast	no	18	-
3	rain	yes	16	-
4	sunny	yes	22	-
5	sunny	no	29	-
6	overcast	yes	30	-
7	overcast	no	35	-
8	rain	yes	23	-
9	rain	no	27	-
10	sunny	yes	28	+
11	overcast	no	23	+
12	sunny	no	27	+
13	rain	no	23	+

Verfeinerung im BEXA

Spezialisierungsprozeß

Konjunktion	Xn	Xp
$[A \in \{a,b,c\}][B \in \{x,y\}]$	{2,5}	{1,3,4,6}
$[A \in \{b,c\}][B \in \{x,y\}]$	{5}	{3,4,6}
$[A \in \{b,c\}][B = y]$	{}	{4,6}

Training Set

#	**A**	**B**	**Class**
1	A	x	+
2	A	y	-
3	b	x	+
4	b	y	+
5	c	x	-
6	c	y	+

$A \in \{a,b,c\}$

$B \in \{x,y\}$

Unterschiede zwischen CN2 und BEXA

- Hauptunterschied: **Hypothesensprache**
- Andere Unterschiede:
  - Restriktionen bei Bedingungssuche
    - prevent-empty-conjunction
    - uncover-new-negative
    - irredundancy-restriction
  - stop-growth-test
    - Übergeneralisierung, Kompensation von “Rauschen”
  - Post processing

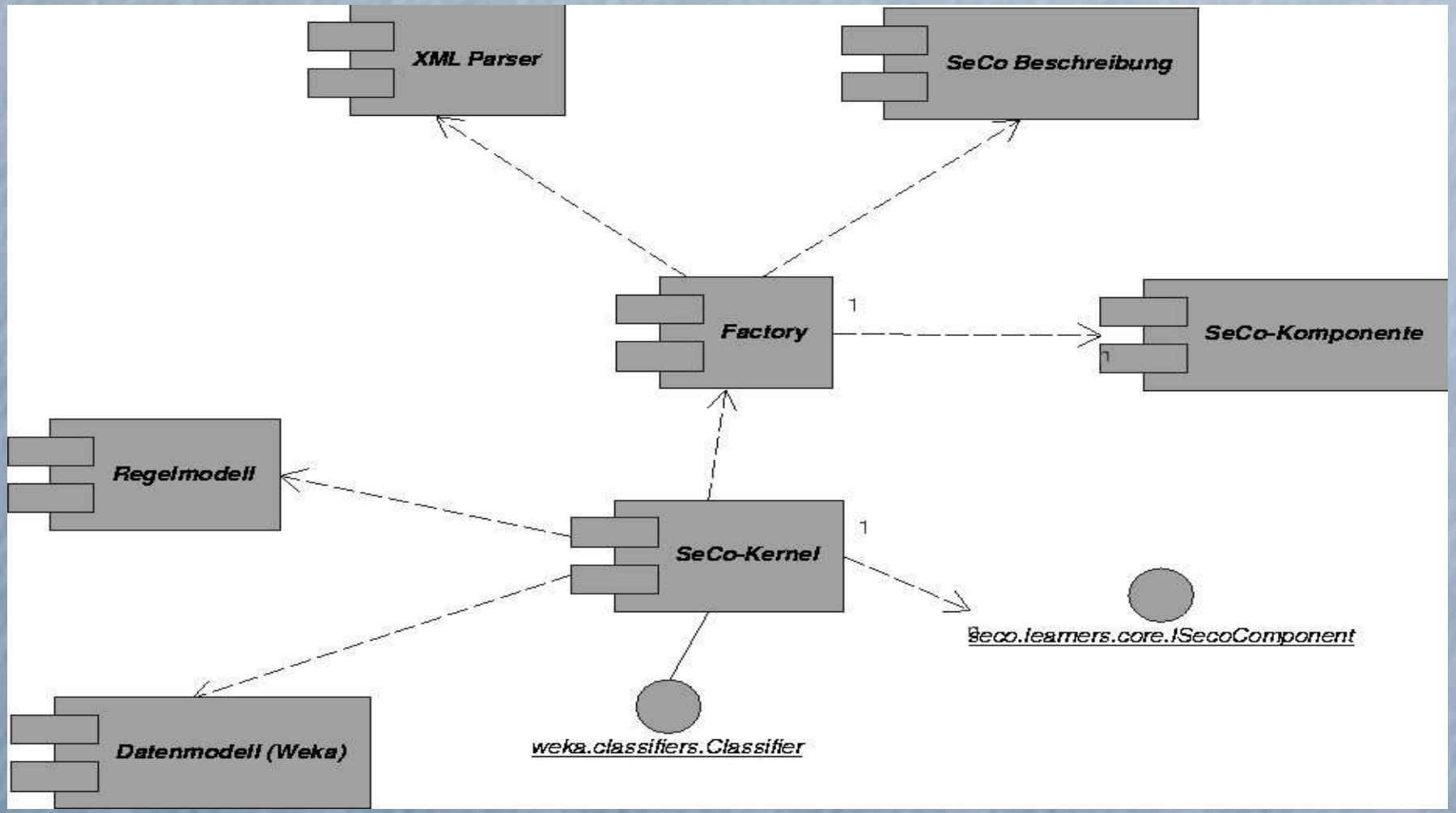
Problemstellung für Vergleich

- Ziel: Vergleich der Hypothesensprachen
- CN2 und BEXA unterscheiden sich in mehreren Punkten
- Art der Implementationen ist verschieden und könnte Meßergebnisse verfälschen.
- => Benötigt wird eine Implementation, die sich nur im interessierenden Punkt unterscheidet.

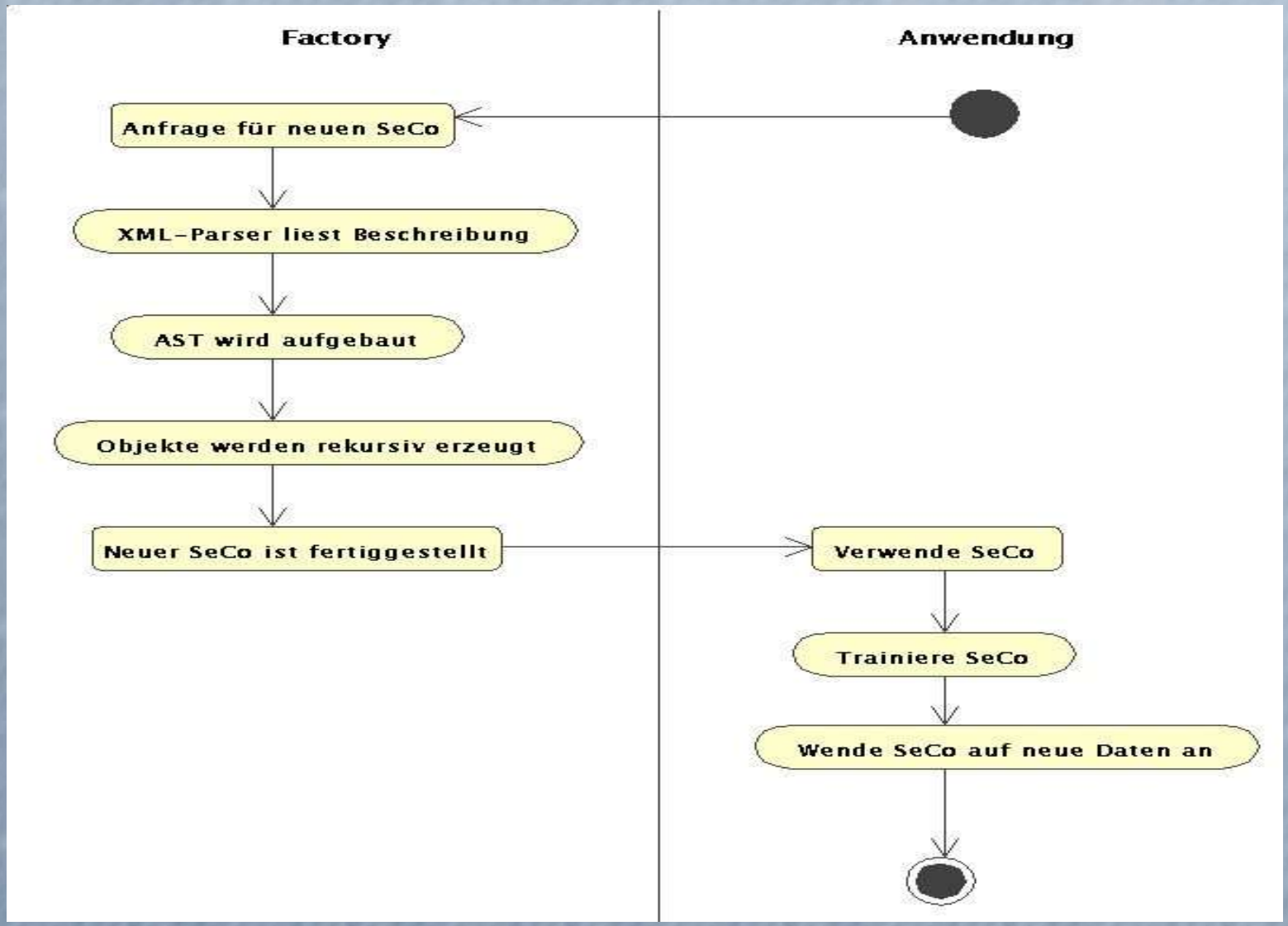
SeCo-Komponenten

• RuleInitializer	Bestimmung der Startregel
• RuleEvaluator	Bewertung der Regeln
• CandidateSelector	Auswahl für Verfeinerung
• RuleRefiner	Verfeinerung einer Regel
• StoppingCriterion	Verfeinerung stoppen ?
• RuleFilter	Filter für Kandidatenregeln
• RuleStoppingCriterion	Lernprozeß beenden ?
• PostProcessor (optional)	Nachbearbeitung der Regeln

Bestandteile des Frameworks



Verwendung der Factory



XML Beschreibung des BEXA

Quelltext 5.2 Beschreibung des **BEXA-Algorithmus**

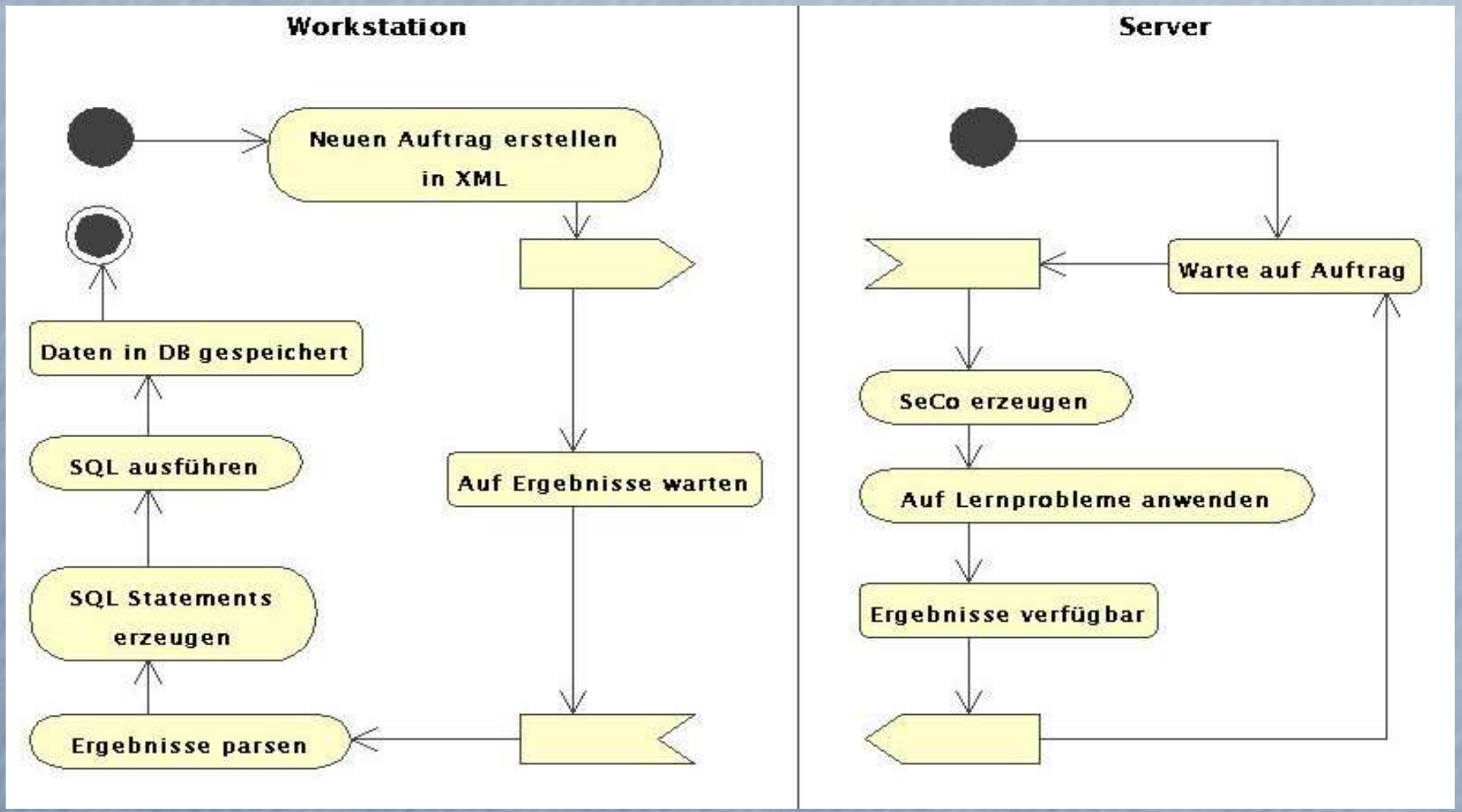
```
<seco>
  <secomp interface="ruleevaluator"
    classname="DefaultRuleEvaluator">
     <jobject package="seco.heuristics" classname="Laplace"
       setter="heuristic"/>
  </secomp>

  <secomp interface="rulerefiner" classname="BexaRefinerTopDown"
   package="seco.learners.bexa"/>
  <secomp interface="selector" classname="DefaultSelector"/>

  <secomp interface="stopcriterion" classname="LikelihoodRatio">
    <property name="threshold" value="0.9"/>
  </secomp>
  <secomp interface="rulefilter" classname="MultiRuleFilter">
    <jobject classname="ChiSquareFilter" setter="filter">
      <property name="threshold" value="0.9"/>
    </jobject>
    <jobject classname="BeamWidthFilter" setter="filter">
      <property name="beamwidth" value="3"/>
    </jobject>
  </secomp>

</seco>
```

SeCo-Factory im Serverbetrieb



Fallstudie: Hypothesensprache

- Ziel: Vergleich des konjunktiven Regellernens (=) mit disjunktivem Regellernen (!=).
- Was gemessen wird
  - Bekannte Kriterien
    - Korrektheit
    - Anzahl der Regeln
  - Neue Kriterien
    - Anzahl der referenzierten Attribute
    - Normalisierte Regellänge

Korrektheit

Problem	BEXA disj.	BEXA konj.	Diff.	LRS	χ^2
anneal	98.44	99.78	-1.34	0.9	0.99
anneal.ORIG	95.55	95.43	.12	0.995	0
audiology	69.91	69.03	.88	0.7	0.7
autos	76.1	80.98	-4.88	0	0.99
breast-cancer	74.83	73.43	1.40	0.99	0
breast-w	95.71	95.99	-.28	0	0.7
colic	78.8	79.62	-.82	0.99	0.99
colic.ORIG	81.25	74.46	6.79	0.99	0.99
credit-a	84.06	84.78	-.72	0.99	0.9
credit-g	72	74.4	-2.40	0.99	0.9
heart-c	76.57	77.89	-1.32	0.995	0.995
heart-h	77.21	79.25	-2.04	0.995	0.995
hepatitis	83.87	85.16	-1.29	0.99	0
hypothyroid	97.14	97.38	-.24	0.995	0.9
kr-vs-kp	99.19	99.31	-.12	0.99	0.7
labor	89.47	92.98	-3.51	0	0.7
lymph	86.49	81.76	4.73	0.9	0.995
mushroom	100	100	.00	0	0
primary-tumor	38.94	39.23	-.29	0.9	0.995
sick	97.75	97.48	.27	0.995	0.7
soybean	89.31	90.63	-1.32	0	0.7
splice	89.53	52.92	36.61	0.7	0.995
tic-tac-toe	98.54	98.33	.21	0.99	0
vote	96.55	96.09	.46	0.99	0.99
vowel	51.52	57.17	-5.65	0.7	0.995
zoo	92.08	87.13	4.95	0.9	0
Durchschnitt	84.26	83.10	1.16		
Gewonnen	10	15			

Anzahl der Regeln/Bedingungen

	Anzahl der Regeln			Anzahl der Bedingungen		
Problem	disj.	konj.	$\frac{disj.}{konj.}$	disj.	konj.	$\frac{disj.}{konj}$
anneal	27	31	0.87	62	51	1.21
anneal.ORIG	35	46	0.76	96	132	0.72
audiology	94	94	1.00	261	204	1.27
autos	69	65	1.06	141	122	1.15
breast-cancer	19	20	0.95	124	56	2.21
breast-w	37	39	0.94	95	102	0.93
colic	31	48	0.64	84	118	0.71
colic.ORIG	62	193	0.32	206	198	1.04
credit-a	64	72	0.88	203	209	0.97
credit-g	125	139	0.89	432	525	0.82
heart-c	28	25	1.12	71	60	1.18
heart-h	29	34	0.85	72	80	0.90
hepatitis	15	15	1.00	22	22	1.00
hypothyroid	61	52	1.17	177	142	1.24
kr-vs-kp	73	57	1.28	301	225	1.33
labor	10	12	0.83	18	18	1.00
lymph	18	15	1.20	55	35	1.57
mushroom	19	22	0.86	47	28	1.67
primary-tumor	63	77	0.81	315	294	1.07
sick	62	62	1.00	171	164	1.04
soybean	87	95	0.91	381	313	1.21
splice	55	2185	0.02	521	2202	0.23
tic-tac-toe	45	15	3.00	233	45	5.17
vote	17	14	1.21	49	39	1.25
vowel	250	324	0.77	616	743	0.82
zoo	11	26	0.42	22	29	0.75
Mittel[48]	54.07	145.26	0.78	183.65	236.76	1.08

Normalisierte Regellänge

	Referenzierte Attribute			Normalisierte Regellänge		
Problem	disj.	konj.	$\frac{disj.}{konj.}$	disj.	konj.	$\frac{disj.}{konj.}$
anneal	52	48	1.08	62	51	1.21
anneal.ORIG	75	111	0.67	96	125	0.76
audiology	247	204	1.21	250	204	1.22
autos	128	120	1.06	141	122	1.15
breast-cancer	98	56	1.75	117	56	2.08
breast-w	90	97	0.92	95	102	0.93
colic	78	114	0.68	84	118	0.71
colic.ORIG	151	198	0.76	206	198	1.04
credit-a	167	185	0.90	203	209	0.97
credit-g	315	419	0.75	432	525	0.82
heart-c	69	58	1.18	71	60	1.18
heart-h	66	74	0.89	71	80	0.88
hepatitis	22	22	1.00	22	22	1.00
hypothyroid	149	119	1.25	177	142	1.24
kr-vs-kp	298	225	1.32	298	225	1.32
labor	18	18	1.00	18	18	1.00
lymph	50	35	1.42	53	35	1.51
mushroom	45	28	1.60	46	28	1.64
primary-tumor	308	294	1.04	308	294	1.04
sick	149	147	1.01	171	164	1.04
soybean	363	313	1.15	369	313	1.17
splice	367	2202	0.16	468	2202	0.21
tic-tac-toe	190	45	4.22	190	45	4.22
vote	49	39	1.25	49	39	1.25
vowel	465	597	0.77	616	743	0.82
zoo	22	29	0.75	22	29	0.75
Mittel[49]	155.03	222.96	1.00	178.26	236.50	1.06

Spezialfall: Splice

BEXA konj.:

```
Class = EI :- attribute_50 = N.  [3|0] Val: 0.8
Class = EI :- Instance_name = HUMALBGC-DONOR-17044.  [2|0]
Val: 0.75
Class = EI :- Instance_name = HUMMYLCA-DONOR-2559.  [2|0]
Val: 0.75
```

BEXA disj.:

```
Class = EI :- attribute_35 != C, attribute_35 != T,
attribute_35 != A, attribute_32 != C, attribute_34 != T,
attribute_32 != G, attribute_32 != A, attribute_34 != G,
attribute_31 != T, attribute_53 != A, attribute_2 != G,
attribute_44 != A, attribute_17 != A.  [216|0] Val: 0.995

Class = EI :- attribute_32 != A, attribute_32 != C,
attribute_32 != G, attribute_31 != C, attribute_31 != A,
attribute_31 != T, attribute_18 != T, attribute_35 != C,
attribute_35 != T, attribute_35 != A, attribute_24 != C,
attribute_6 != T, attribute_41 != A.  [198|0] Val: 0.995
```

Ergebnisse der Experimente

- Korrektheit
  - BEXA disj. verliert bei meisten Problemen
  - Differenz ist bei meisten Problemen jedoch gering
- Größe der Regelmengen
  - BEXA disj. erzeugt tendenziell weniger Regeln
  - aber mit mehr Bedingungen
- BEXA konj. versagt bei Problem Splice, während BEXA disj. gute Ergebnisse liefert.

Schlußfolgerung

- SeCo-Framework ist für Umsetzung von SeCo Algorithmen geeignet
  - Wiederverwendbarkeit von Quelltext
  - Isolierte Betrachtung von Modifikationen
  - komfortable Konfiguration
  - Versuche sind besser nachvollziehbar
- Fallstudie: Hypothesensprache
  - Eignung der Sprache ist stark von Lernproblem abhängig
  - Man sollte Experimente nicht auf UCI Repository beschränken.
  - Eigenschaften von Lernproblemen sollten genauer untersucht und formal beschrieben werden.

Offene Punkte

- SeCo-Framework
  - Optimierung der Rechenzeit
  - Verallgemeinerung
  - Entwurf eines Meta-Algorithmus
- Fallstudie: Hypothesensprache
  - Leicht prüfbares Kriterium zur Wahl von disj/konj
  - Heuristik für BEXA mixed.
  - Gibt es eine Heuristik, welche die Schwäche von konj. bei kritischen Lernproblemen ausgleichen kann?