Semiüberwachte Paarweise Klassifikation

Andriy Nadolskyy

Bachelor-Thesis

Betreuer: Prof. Dr. Johannes Fürnkranz
Dr. Eneldo Loza Mencía

Überblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation
- Grundbegriffe
- Einleitung
- Übersicht der Verfahren
- Datensätze
- Evaluation
- Zusammenfassung
- Ausblick

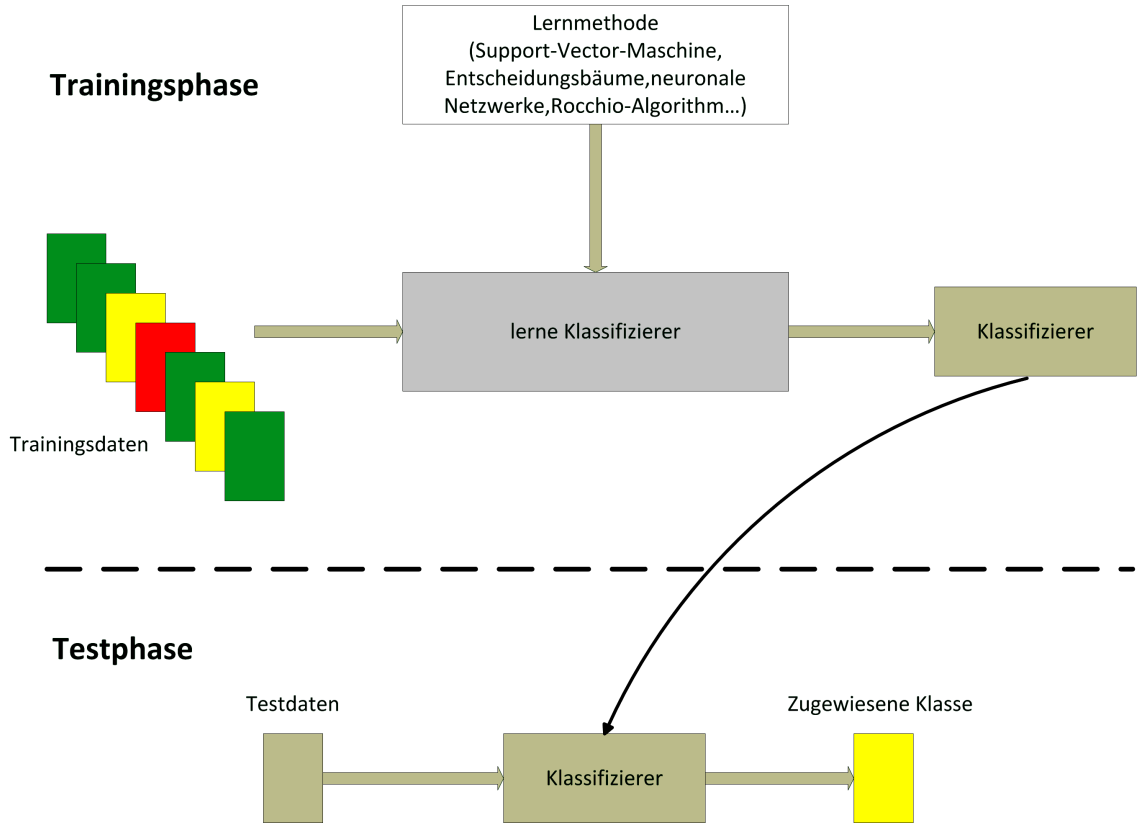
Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Riesige Menge an Informationen im Internet
- Die Fülle an Informationen benutzbar für den Endbenutzer zu machen

Grundbegriffe (1)

- Maschinelles Lernen
- Klassifizierung



Grundbegriffe (2)

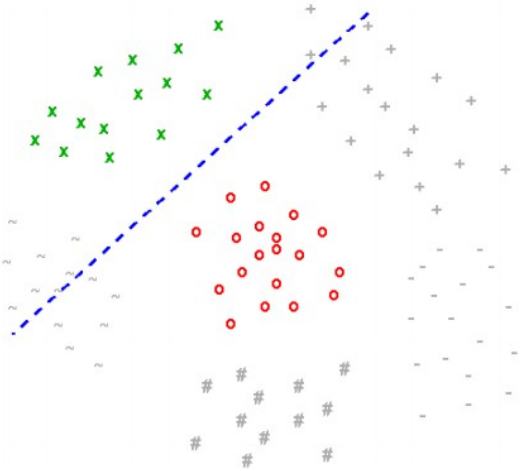
TECHNISCHE UNIVERSITÄT DARMSTADT

- überwachtes Lernen (supervised learning):
  - Klassenattribute sind für alle Instanzen bekannt
- unüberwachtes Lernen (unsupervised learning):
  - Klassenattribute sind nicht bekannt
- semiüberwachtes Lernen (semi-supervised learning):
  - Klassenattribute sind nur zum Teil bekannt

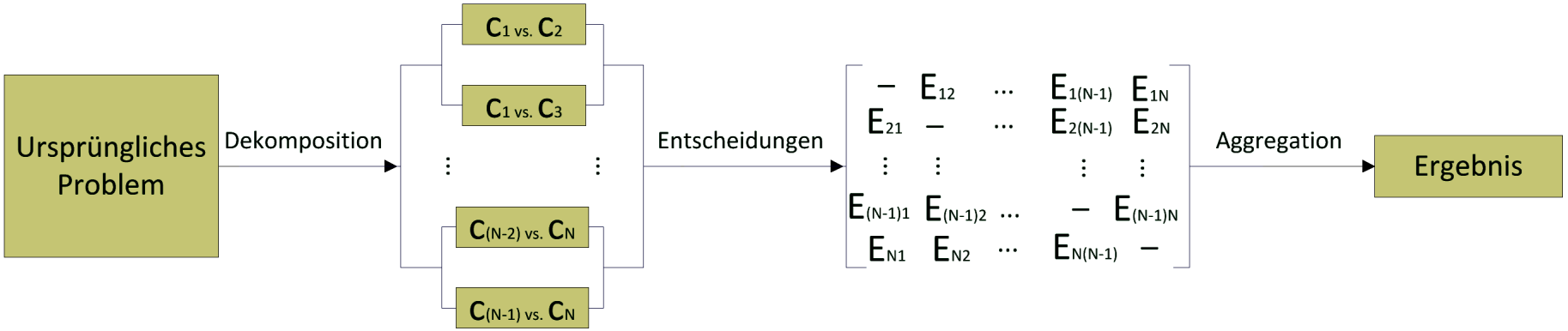
Grundbegriffe (3)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Multiklassen-Probleme



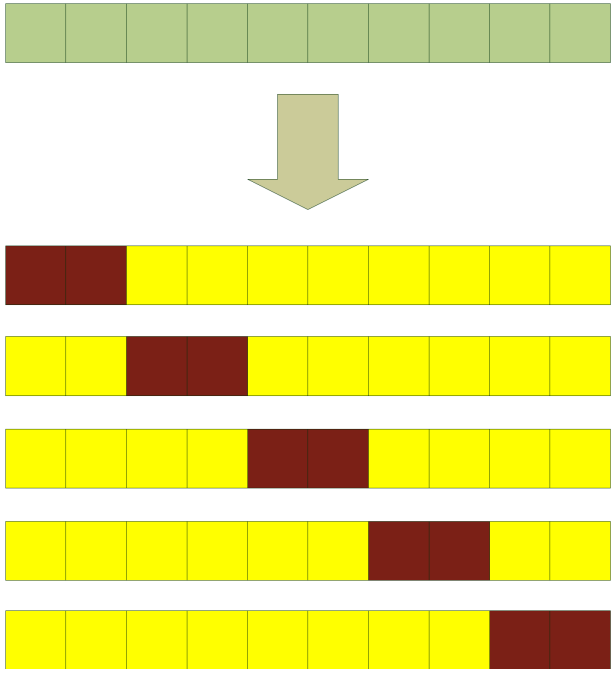
- paarweise Klassifizierung



Grundbegriffe (4)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Selbsttraining



- Kreuzvalidierung

Einleitung

TECHNISCHE UNIVERSITÄT DARMSTADT

- die Idee: überprüfen, inwieweit sich ein paarweiser Klassifizierer verbessern lässt
  - Information aus Rankings berücksichtigen
  - mehr Trainingsbeispiele durch Hinzunahme aus Ranking gewonnener Multiklass-Vorhersagen

- Trainingsdaten anpassen:
  - Präferenzen, die einer gelernten Reihung widersprechen, „korrigieren“ und dann nochmal trainieren
  - Präferenzen, die in den Trainingsdaten nicht auftreten, die aber in der ersten Iteration gelernt werden, ganz (oder zum Teil) dem ursprünglichen Trainingsset hinzufügen

Beispiel (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

- sei $\{c_0, c_1, c_2, c_3, c_4\}$ eine Menge von Labels
- für jedes Paar der Labels einen Klassifizierer trainieren
- es gibt also Klassifizierer für $c_0 > c_1$, $c_0 > c_2$, $c_0 > c_3$, $c_0 > c_4$, $c_1 > c_2$, $c_1 > c_3$, $c_1 > c_4$, $c_2 > c_3$, $c_2 > c_4$ und $c_3 > c_4$
- jeden dieser 10 Klassifizierer abfragen

Beispiel (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

$\{C_0, C_1, C_2, C_3, C_4\}$

$\{C_0 > C_1, C_0 > C_2, C_0 > C_3, C_0 > C_4, C_1 > C_2, C_1 > C_3, C_1 > C_4, C_2 > C_3, C_2 > C_4, C_3 > C_4\}$

Klassifizierer trainieren

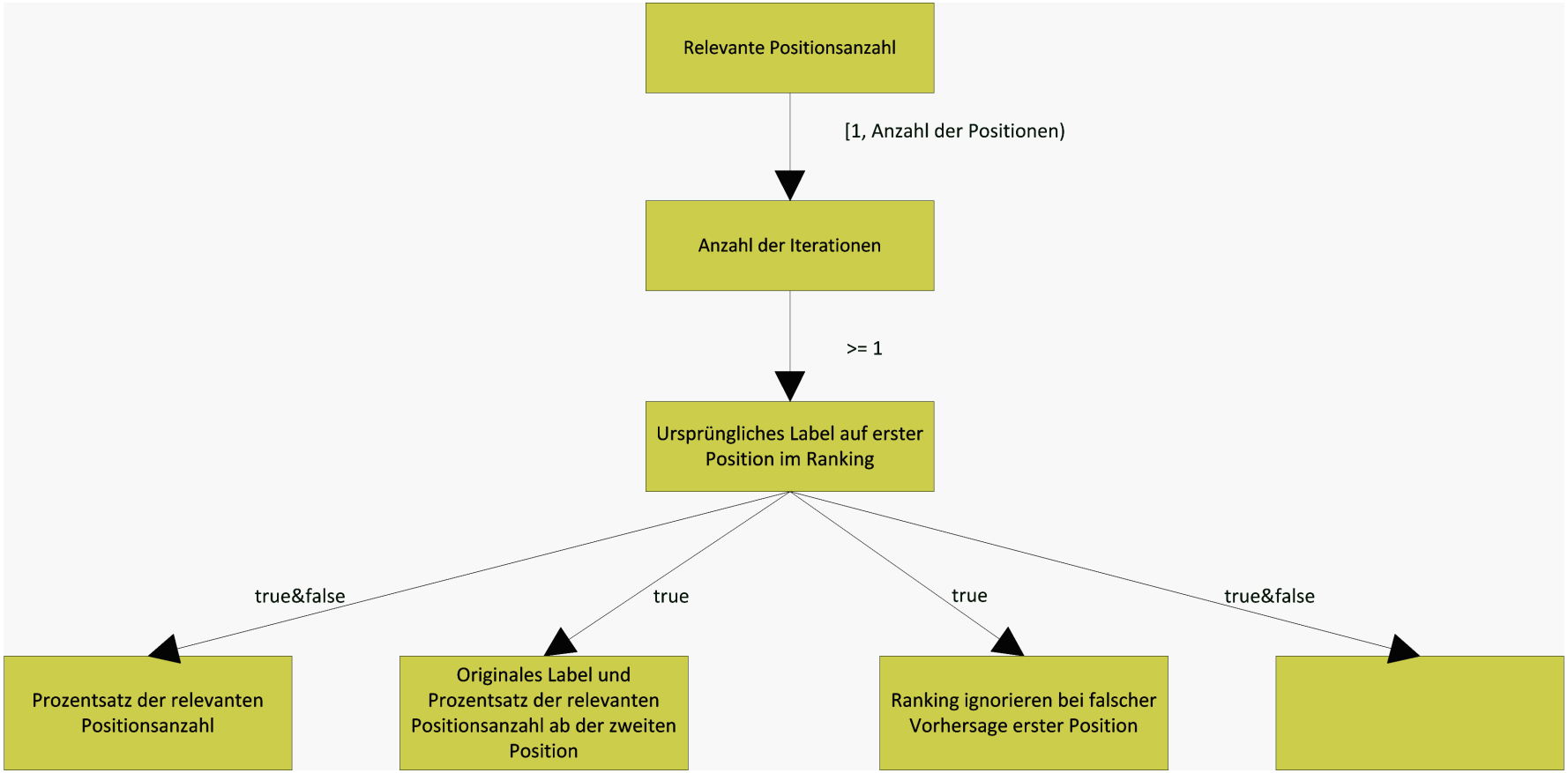
Klassifizierer abfragen

Beispiel (3)

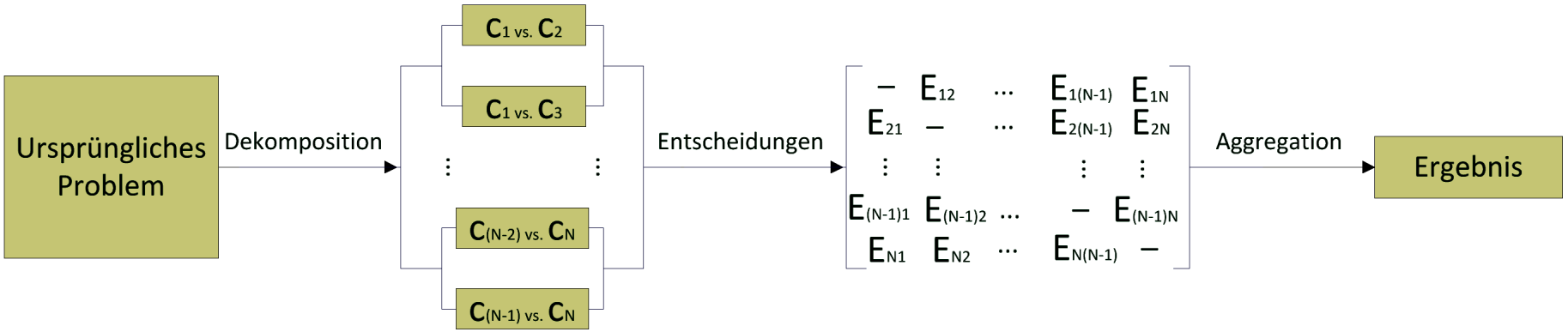
TECHNISCHE UNIVERSITÄT DARMSTADT

- endgültige Ordnung der Klassen durch Abstimmung bestimmen
- angenommen, c_0 hat 3, c_1 2, c_2 2, c_3 2 und c_4 1 Stimme
- eine Ordnung $c_0 > c_1, c_2, c_3 > c_4$ wird vorhergesagt
- binärer Klassifizierer für das Klassenpaar c_0 und c_4: $c_4 > c_0$
- es kann vorkommen, dass gegebene Präferenzen nach dem Lernen entgegengesetzt werden

Übersicht der Verfahren



Einstiegspunkt



- angepasstes Trainingsset
- das Gleiche noch mal

Verfahren (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

- ursprüngliches Label auf beliebiger Position im Ranking
- variable relevante Positionsanzahl (*RP*)

- sei $[c_0, c_1, c_2, c_3]$ gewonnene totale Ordnung einer Instanz
- bei *RP*=1 bekommen wir $c_1<c_0$, $c_2<c_0$, $c_3<c_0$
- bei *RP*=2 kommen noch $c_2<c_1$, $c_3<c_1$ dazu
- bei *RP*=3 kommt $c_3<c_2$ dazu

Verfahren (2)

TECHNISCHE UNIVERSITÄT DARMSTADT

- ursprüngliches Label auf beliebiger Position im Ranking
- variable relevante Positionsanzahl (*RP*)
- Prozentsatz der relevanten Positionsanzahl (*PSRP*)

- sei $[c_0, c_1, c_2, c_3]$ gewonnene totale Ordnung einer Instanz
- bei *RP*=2 bekommen wir $c_1<c_0$, $c_2<c_0$, $c_3<c_0$, $c_2<c_1$, $c_3<c_1$
- bei *PSRP*=60 werden z.B. Labels $c_1<c_0$, $c_3<c_0$, $c_2<c_1$ übernommen

Verfahren (3)

TECHNISCHE UNIVERSITÄT DARMSTADT

- ursprüngliches Label auf erster Position im Ranking
- variable relevante Positionsanzahl (*RP*)

- äquivalent zu dem entsprechenden obigen Verfahren
- … außer der Position des originalen Labels im Ranking
- angenommen, c_0 das ursprüngliche Label und $[c_1, c_2, c_0, c_3]$ gewonnene totale Ordnung
- Ranking wird angepasst: $[c_0, c_1, c_2, c_3]$

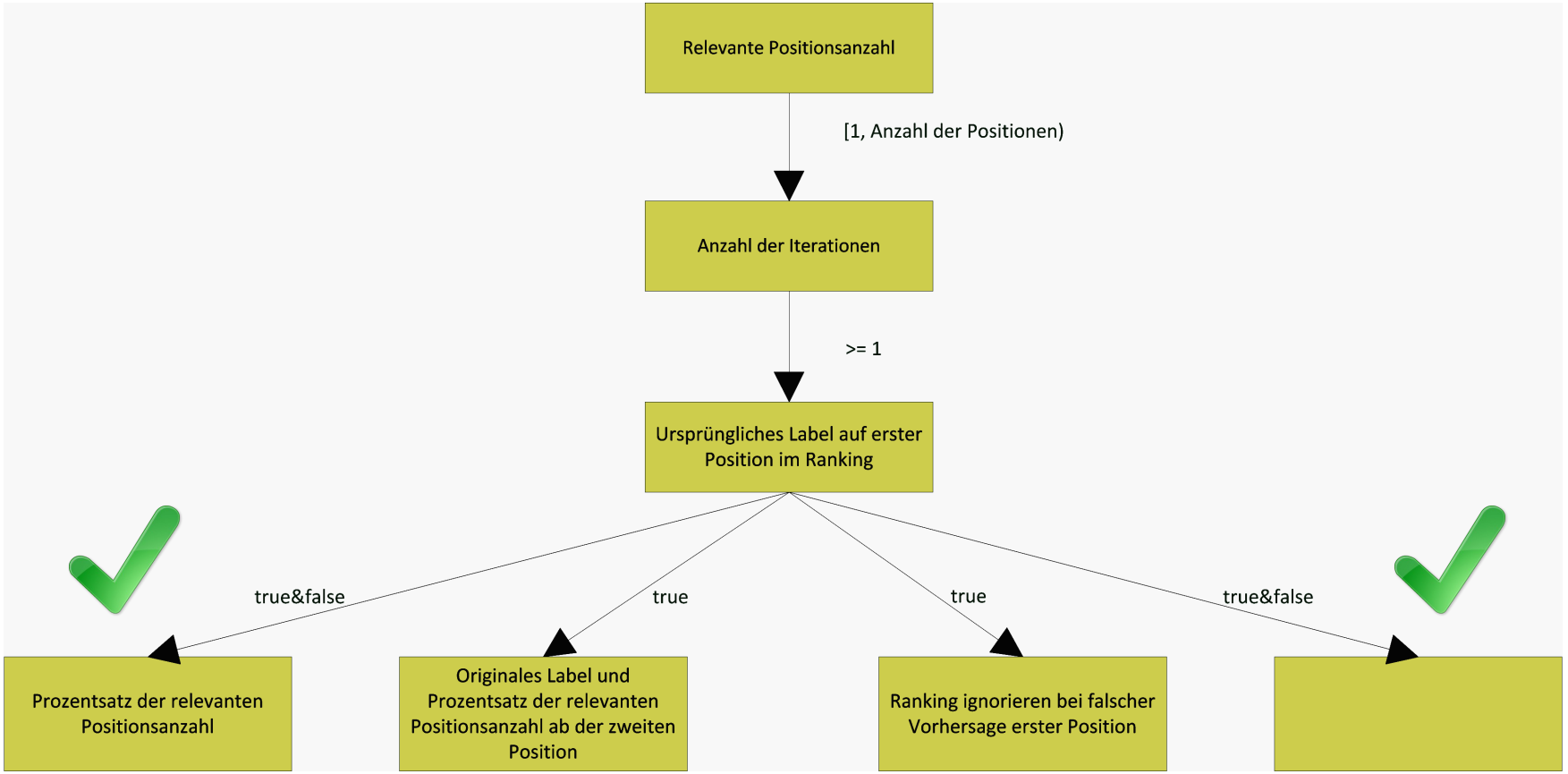
Verfahren (4)

TECHNISCHE UNIVERSITÄT DARMSTADT

- ursprüngliches Label auf erster Position im Ranking
- variable relevante Positionsanzahl (*RP*)
- Prozentsatz der relevanten Positionsanzahl (*PSRP*)

- äquivalent zu dem entsprechenden obigen Verfahren
- … außer der Position des originalen Labels im Ranking

Übersicht der Verfahren



Verfahren (5)

TECHNISCHE UNIVERSITÄT DARMSTADT

- ursprüngliches Label auf erster Position im Ranking
- variable relevante Positionsanzahl (*RP*)
- Prozentsatz der *RP* ab der zweiten Position

- sei $[c_0, c_1, c_2, c_3]$ gewonnene totale Ordnung und der Prozentsatz für die zweite Position ist 70
- $c_1<c_0$, $c_2<c_0$, $c_3<c_0$ werden komplett übernommen und von $c_2<c_1$, $c_3<c_1$ wird z.B. nur $c_3<c_1$ beibehalten

Verfahren (6)

TECHNISCHE UNIVERSITÄT DARMSTADT

- ursprüngliches Label auf erster Position im Ranking
- variable relevante Positionsanzahl (*RP*)
- Ranking bei falscher Vorhersage erster Position ignorieren

- angenommen, c_0 das ursprüngliche Label und $[c_1, c_2, c_0, c_3]$ gewonnene totale Ordnung
- Ranking wird verweigert
- paarweise Vergleiche für c_0 gebildet: $c_1 < c_0$, $c_2 < c_0$, $c_3 < c_0$

Evaluierungsmaße

TECHNISCHE UNIVERSITÄT DARMSTADT

- *Accuracy*: Prozentsatz korrekt klassifizierter Instanzen

$$\frac{N_{A,A} + N_{B,B}}{N}$$

		Classified as A	Classified as B	
True class	A	$N_{A,A}$	$N_{B,A}$	$N_{A,A} + N_{B,A}$
True class	B	$N_{A,B}$	$N_{B,B}$	$N_{A,B} + N_{B,B}$
		$N_{A,A} + N_{A,B}$	$N_{B,A} + N_{B,B}$	N

- *Positionsfehler*: Distanz zu der Position des originalen Labels im Ranking (normalisiert)

$$\frac{\sum_{i=1}^{n} positionError_i}{n}$$

Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

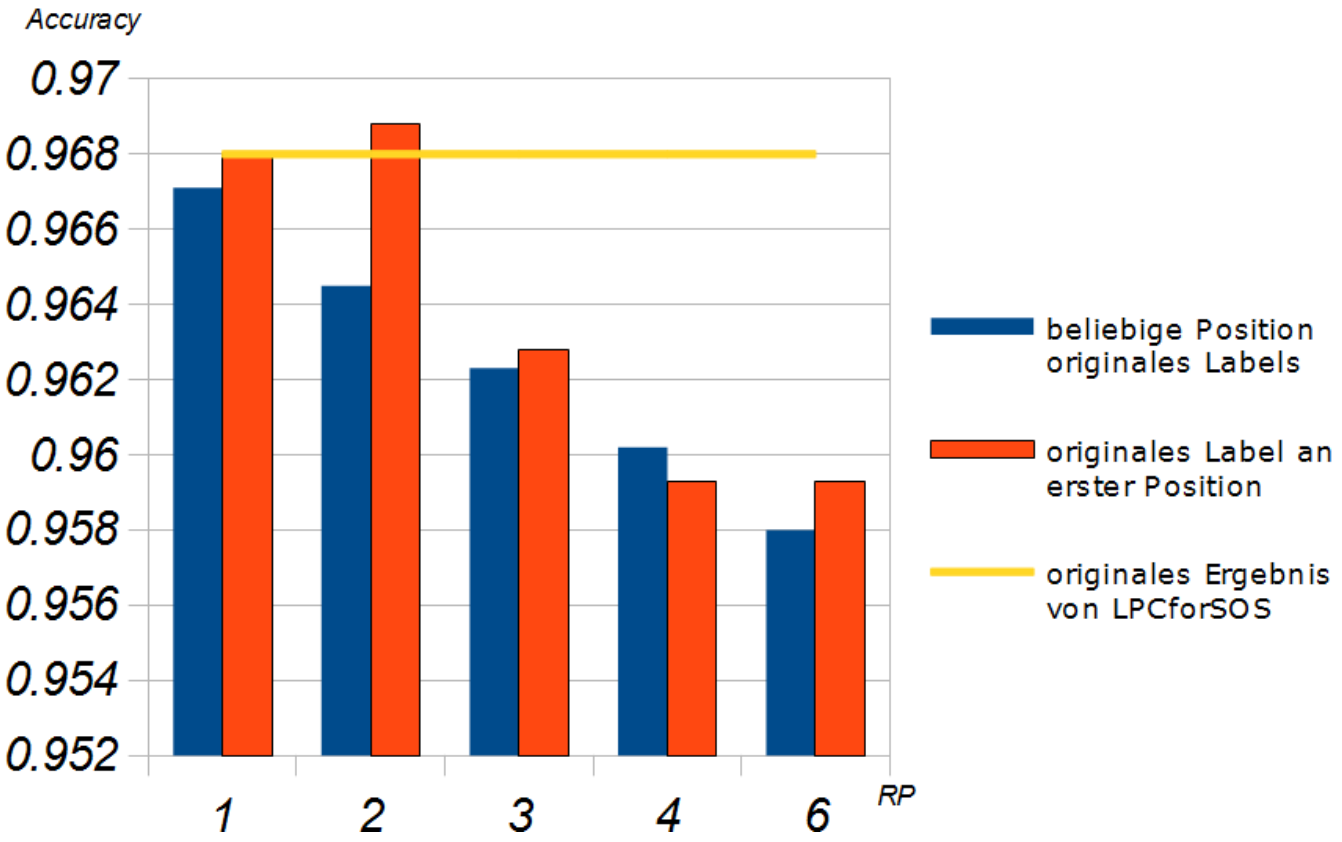
- stammen aus der UCI-Repository und 19 von George Forman gestifteten Multiklass-Textdatensätzen
- liegen im .arff-Format vor

Datensatz	Instanzen	Attribute	Klassen
letter	20000	17	26
mfeat-fourier	2000	77	10
optdigits	5620	65	10
segment	2310	20	7
fbis.wc	2463	2001	17
la1s.wc	3204	13196	6
la2s.wc	3075	12433	6
new3s.wc	9558	26833	44
oh0.wc	1003	3183	10
oh5.wc	918	3013	10
oh10.wc	1050	3239	10
oh15.wc	913	3101	10
ohscal.wc	11162	11466	10
re0.wc	1504	2887	13
re1.wc	1657	3759	25
tr11.wc	414	6430	9
tr12.wc	313	5805	8
tr21.wc	336	7903	6
tr23.wc	204	5833	6
tr31.wc	927	10129	7
tr41.wc	878	7455	10
tr45.wc	690	8262	10
wap.wc	1560	8461	20

Ergebnisse (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

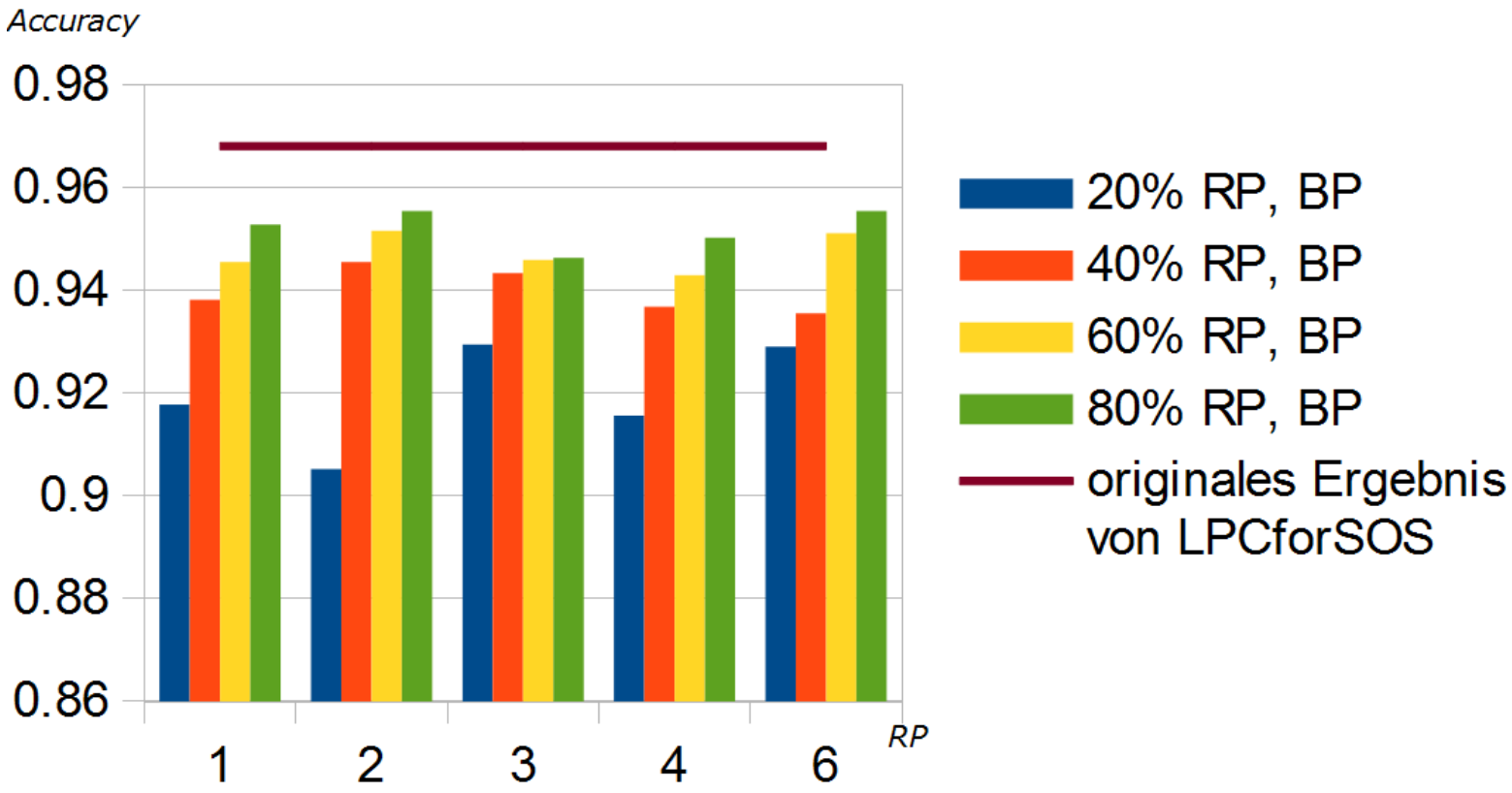
- variable *RP*: beliebige gegen erste Position originales Labels



Ergebnisse (2)

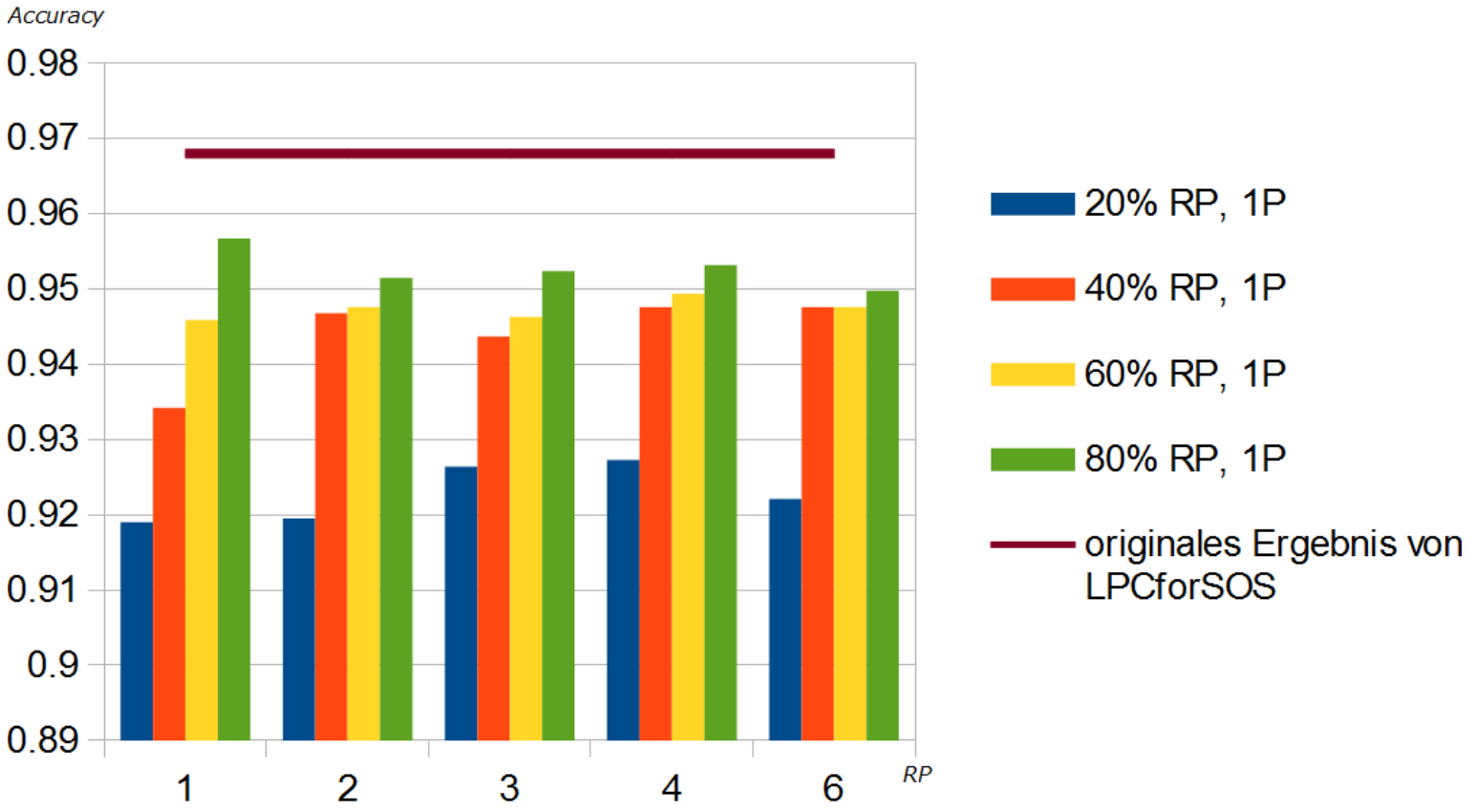
TECHNISCHE UNIVERSITÄT DARMSTADT

- variabler *PSRP:* originales Label an beliebiger Position



Ergebnisse (3)

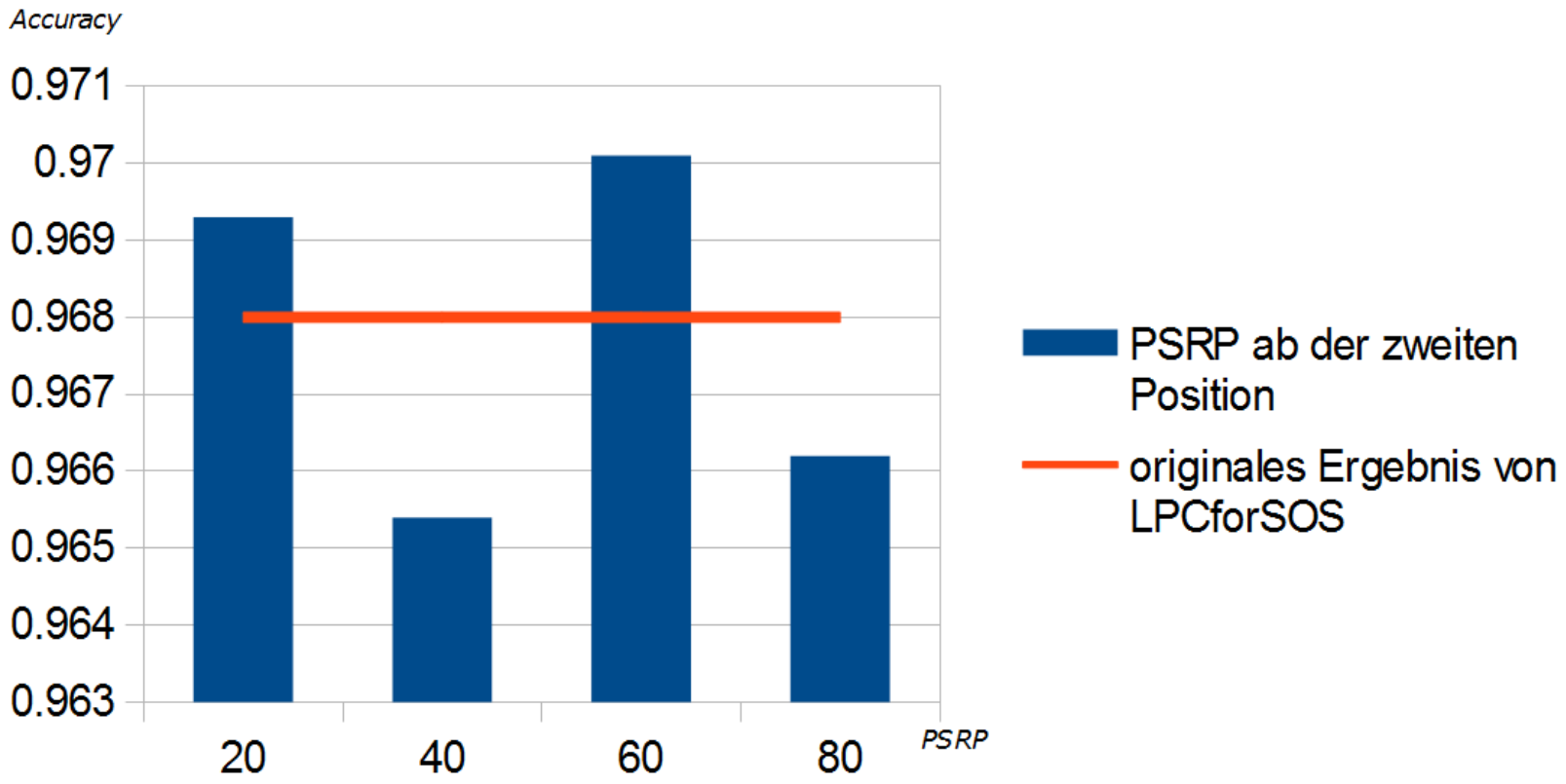
- variabler *PSRP:* originales Label an erster Position



Ergebnisse (4)

TECHNISCHE UNIVERSITÄT DARMSTADT

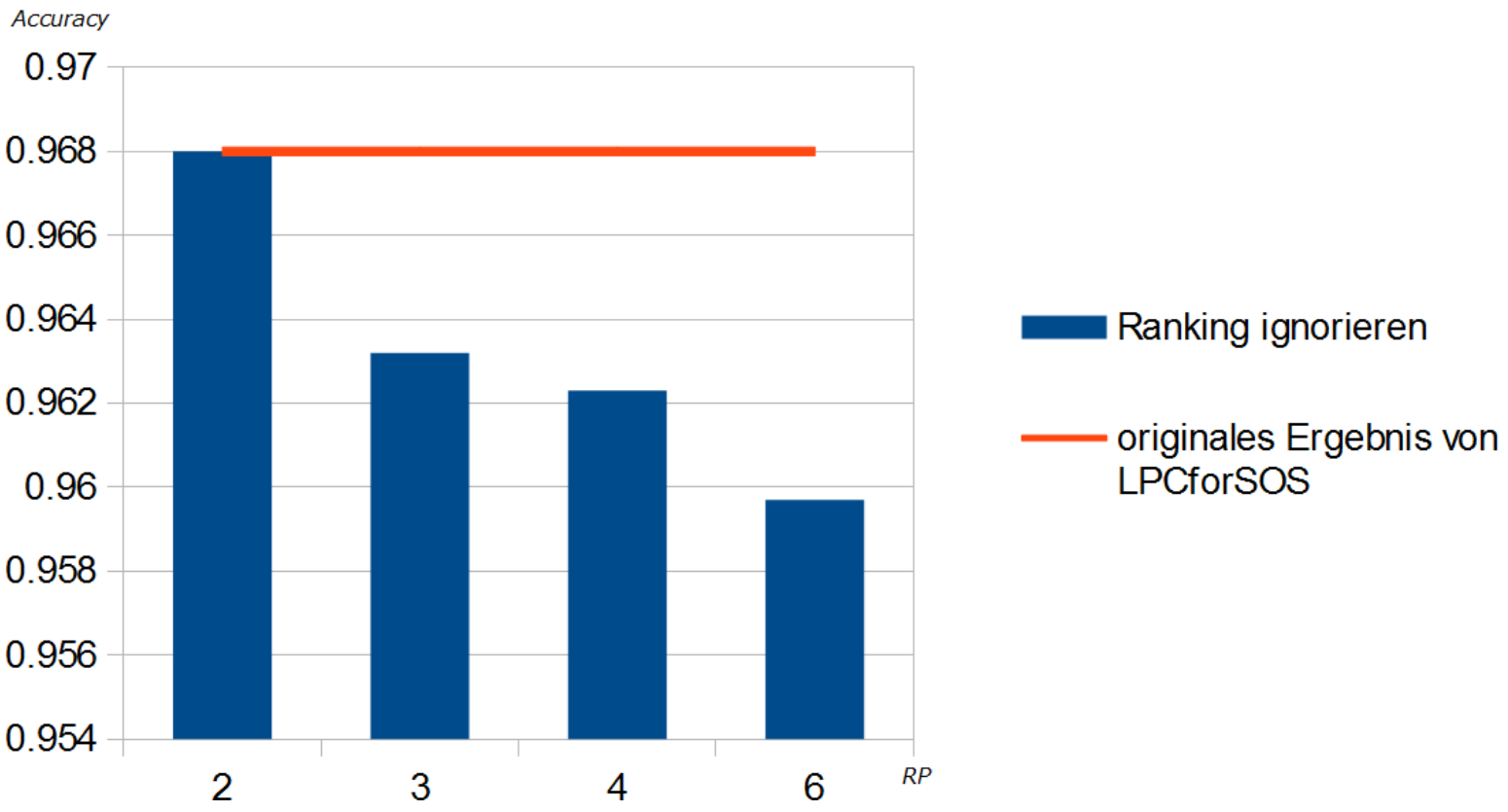
- variabler *PSRP* ab der zweiten Position*:* originales Label an erster Position



Ergebnisse (5)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ranking bei falscher Vorhersage erster Position ignorieren



Ergebnisse (6)

TECHNISCHE UNIVERSITÄT DARMSTADT

	RP, BP	PSRP, BP	RP, 1P	PSRP, 1P	PSRP ab 2P, 1P	ignore Ranking
Datensatz						
letter	0	0	0	0	0	0
mfeat-fourier	0	0	0	0	3	0
optdigits	0	0	0	0	0	0
segment	0	0	1	0	2	1
fbis.wc	-	-	-	-	-	-
la1s.wc	0	0	0	0	2	0
la2s.wc	0	0	0	0	2	0
new3s.wc	-	-	-	-	-	-
oh0.wc	0	0	0	0	0	0
oh5.wc	0	0	0	0	0	0
oh10.wc	0	0	0	0	0	0
oh15.wc	0	0	0	0	0	0
ohscal.wc	-	-	-	-	-	-
re0.wc	0	0	0	0	0	0
re1.wc	-	-	-	-	-	-
tr11.wc	0	0	0	0	0	0
tr12.wc	3	0	5	3	4	5
tr21.wc	5	0	4	0	3	4
tr23.wc	0	0	0	0	2	0
tr31.wc	0	0	1	0	4	0
tr41.wc	0	0	0	0	1	0
tr45.wc	0	0	0	0	0	0
wap.wc	0	0	0	0	0	0
Treffersumme	8	0	11	3	23	10

Zusammenfassung

TECHNISCHE UNIVERSITÄT DARMSTADT

- grundlegende Begriffe
- Erläuterung der Verfahren
- Evaluierungsmaße
- Datensätze
- Ergebnisse
- Schlussfolgerungen

Ausblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- ursprüngliches Label auf erster Position im Ranking
- weitere Maße für die Beschreibung nötig

- Distanz zwischen Labels im Ranking:
  - z.B. Abstand zwischen Votes für Labels
- gewichtete Labels

Vielen Dank für Ihre Aufmerksamkeit!