Vortrag zum Paper „Results of the Active Learning Challenge“ von Guyon, et. al.

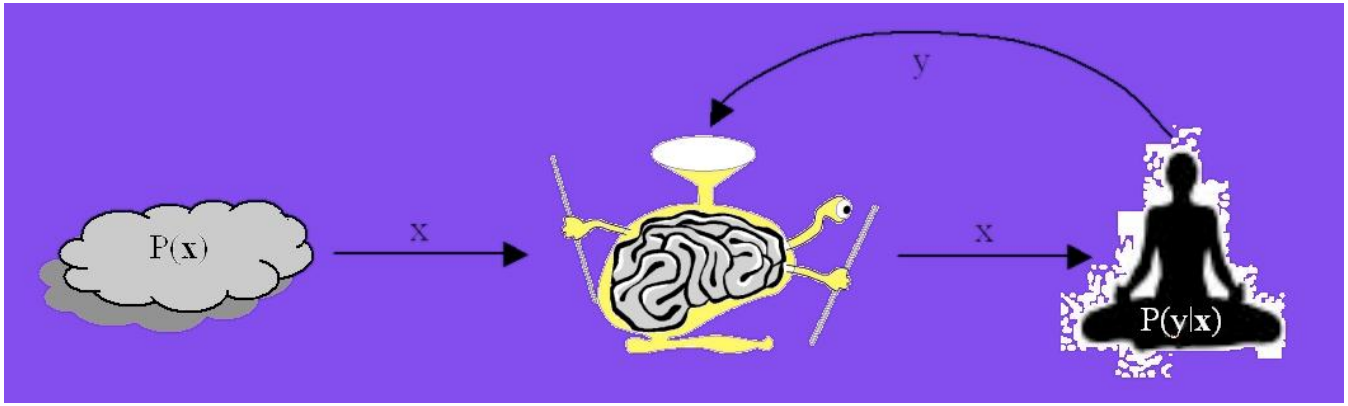
TECHNISCHE UNIVERSITÄT DARMSTADT

Sören Schmidt

Fachgebiet Knowledge Engineering

CHA LEARN

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning



Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

1) Einführung und Grundlagen
2) Beschreibung des Wettbewerbs
3) Übersicht der Ergebnisse
4) Eigenschaften der Ergebnisse
5) Fazit

Was ist Active Learning?

TECHNISCHE UNIVERSITÄT DARMSTADT

Passive Learning	▪ Trainings-Paare {x,y} verfügbar ▪ Anzahl Trainings-Beispiele normal	▪ Spam Filter ▪ Warenkorbanalysen ▪ uvm.
Active Learning	▪ Große Menge an Trainings-Beispielen ▪ Label y jedoch zunächst unbekannt ▪ Cost of labeling (Orakel)	▪ Astronomie (Galaxien erkennen) ▪ Handschrifterkennung (Alte Schriften) ▪ Marketingdaten (CRM) ▪ Datenanalyse (Newsgroups)

Auswahlstrategien: Learning Machine „sollte“ diejenigen Beispiele auswählen / kaufen, die am vielversprechendsten aussehen (d.h. die predictive performance des Modells verbessern)

Active Learning Varianten:

- **Pool-based active learning**: Großer „Pool“ (Menge) von Beispielen ist verfügbar zum Training
- **Stream-based active learning**: Neue Beispiele kommen kontinuierlich im Stream
- **De-novo query synthesis**: Der Lerner kann für beliebige Werte von x das Label y abfragen. X muss nicht in P(x) sein.

Wie evaluiert man einen Machine Learner?

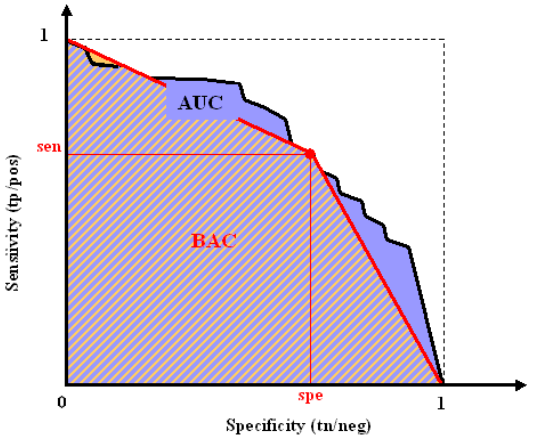
TECHNISCHE UNIVERSITÄT DARMSTADT

Confusion Matrix

	Classified as +	Classified as -	
Is +	true positives (tp)	false negatives(fp)	tp + fn = P
Is -	false positives (fp)	true negatives (tn)	fp + tn = N
	tp + fp	fn + tn	|E| = P + N

Eigenschaften / Definitionen:

- Sensitivity (tp rate) = tp/P
- Specificity (tn rate) = tn/N
- fp rate = 1 – spec



Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

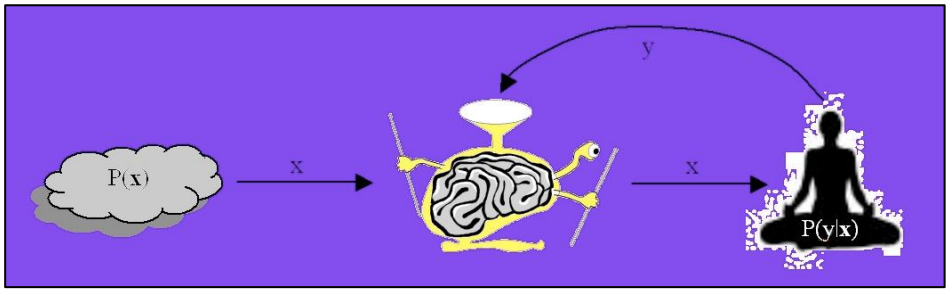
1) Einführung und Grundlagen
2) **Beschreibung des Wettbewerbs**
3) Übersicht der Ergebnisse
4) Eigenschaften der Ergebnisse
5) Fazit

Beschreibung des Wettbewerbs (1 von X)

Was ist die Active Learning Challenge?

TECHNISCHE UNIVERSITÄT DARMSTADT

- 2010 durchgeführter Machine Learning Wettbewerb
  - Annahme: labeling ist teuer (z.B. Expertenwissen)
  - Annahme: Große Mengen unlabeled data verfügbar
  - Annahme: Active learning zur Verbesserung der Vorhersage
- Modellierungs-Zielsetzung – predictive modeling: Gute Vorhersagen für fehlende Label
- Exponentieller Gewinner-Preis: USD $100*2^{(N-1)}$
  - Beim Gewinn von N = {1, 2, 3, 4, 5, 6} Datensätzen können USD 100, 200, 400, 800, 1600, respektive 3200 gewonnen werden.



Beschreibung des Wettbewerbs

Wie lief der Wettbewerb ab?

TECHNISCHE UNIVERSITÄT DARMSTADT

Challenge Protocol:

- Für jeden Datensatz haben die Teams ein Budget von „virtual cash“ (*experimental cash unit, ECU*)
- Jeder Datensatz hat ein einziges Seed-Example/Label, das kostenlos ist
- Teams können Labels der unlabeled Trainingsdaten kaufen – für je 1 ECU pro Label
- Zum Kaufen von Labels muss ein Query mit der Liste an Beispielen an den Server geschickt werden, für den die Labels gekauft werden sollen
  - Beliebige Anzahl von Labels pro Query (innerhalb des Gesamtbudgets)
  - Gesamtbudget = Anzahl von Trainingsdaten - 1
- Experiment endet, wenn Budget aufgebraucht oder Zeit abgelaufen

Monitoring:

- Die Teams mussten bei jedem Query die aktuellen Vorhersagen aller Labels abgeben (unknown training examples und test examples)

Trainingsdatensätze (Development Dataset)

TECHNISCHE UNIVERSITÄT DARMSTADT

Dataset	Feat. type	Feat. num.	Sparsity (%)	Missing (%)	Pos. lbls (%)	Tr & Te num.
ALEX	binary	11	0	0	72.98	5000
HIVA	binary	1617	90.88	0	3.52	21339
IBN SINA	mixed	92	80.67	0	37.84	10361
NOVA	binary	16969	99.67	0	28.45	9733
ORANGE	mixed	230	9.57	65.46	1.78	25000
SYLVA	mixed	216	77.88	0	6.15	72626
ZEBRA	continuous	154	0.04	0.0038	4.58	30744

Aufbau der Datensätze (Final Dataset)

TECHNISCHE UNIVERSITÄT DARMSTADT

	Ft. 1	Ft. 2	Ft. 3	...	...	...	...	...	Ft. n	Label
Trainingsdaten	0	1	0	0	0	1	1	0	0	1
	0	0	0	0	1	1	1	0	0	?
	1	1	0	1	0	0	1	0	0	?
	1	0	0	0	1	1	1	0	0	?
	...	...	...	...	...	...	...	...	...	...
	...	...	...	...	...	...	...	...	...	...
Testdaten	0	1	0	0	0	1	1	0	0	?
	0	0	0	0	1	1	1	0	0	?
	1	1	0	1	0	0	1	0	0	?
	1	0	0	0	1	1	1	0	0	?

Seed Label

Query -able

to be predicted

Performance Evaluation

global_score = (ALC-Arand)/(Amax-Arand)

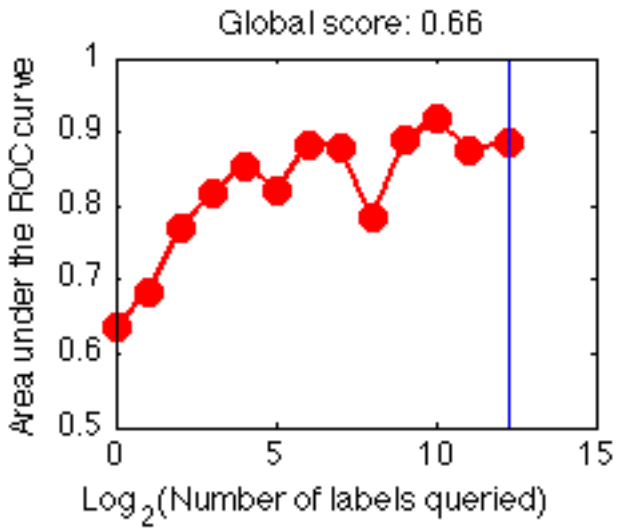
- **ALC = Area under the Learning Curve**
  - ALC trägt die AUC (Area under the ROC curve) für alle unbekannten Labels als Funktion der Anzahl gekaufter Labels ab
- **Amax = Area maximum**
  - Ideale Learning Curve, perfekte Vorhersagen (AUC = 1)
- **Arand = Area random**
  - Lazy Learning Curve, zufällige Vorhersagen (AUC ≈ 0.5)

Zwei Baseline Kurven

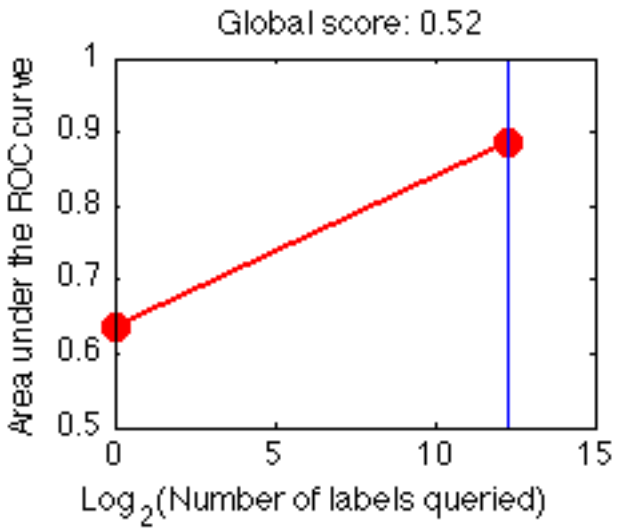
Beschreibung des Wettbewerbs

Beispiel-Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT



Random Sampling



Passive Learning

Eigenschaften

- Logarithmische X-Achse
- AUC > 0.5
- Ende bei log(# of Trainingsbeispiele)

Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

1) Einführung und Grundlagen
2) Beschreibung des Wettbewerbs
3) **Übersicht der Ergebnisse**
4) Eigenschaften der Ergebnisse
5) Fazit

Ergebnisse des Wettbewerbs

Baseline Resultate (Bechmarks / Referenzen)

TECHNISCHE UNIVERSITÄT DARMSTADT

Dataset	Experiment	Classifier	Strategy	AUC	ALC	Rank
HIVA	gcchiva4	Naïve Bayes	Bayesian	0.805504	0.328535	—
IBN_SINA	gccibnsina1	Linear KRR	Random	0.978585	0.813690	—
NOVA	gccnova1	Linear KRR	Random	0.991841	0.715582	—
ORANGE	gccorange1	Linear KRR	Random	0.814340	0.283319	—
SYLVA	gccsylva1	Linear KRR	Random	0.996240	0.921228	—
ZEBRA	gcczebra1	Linear KRR	Random	0.785913	0.416948	—
Avicena	gccA004v	Linear KRR	Random	0.883768	0.586001	3
Banana	gccb1	Linear KRR	Passive	0.720291	0.370762	3
Chemo	gccc4	Linear KRR	Random	0.814450	0.301776	5
Docs	gccd2	Linear KRR	Random	0.962951	0.651222	6
Embryo	gcce1	Linear KRR	Passive	0.773262	0.496610	5
Forest	gccf2	Linear KRR	Random	0.954557	0.821711	1

- Einfacher Ansatz / Klassifizierer
  - Komplexere nicht-lineare Methods (decision tree, SVM, naive Bayes,...) nicht verwendet
  - Gefahr von over-fitting; Linear KRR weniger over-fitting (Wahl des Rigide Parameter)
- Einfache Sampling Strategien
  - Passive learning (Query aller Labels auf einmal) und random learning bringen gute Resultate
- Sehr gute Resultate (avg. Rank = 3.833) – Selection-Bias

Ergebnisse des Wettbewerbs

Ergebnisse Dataset A

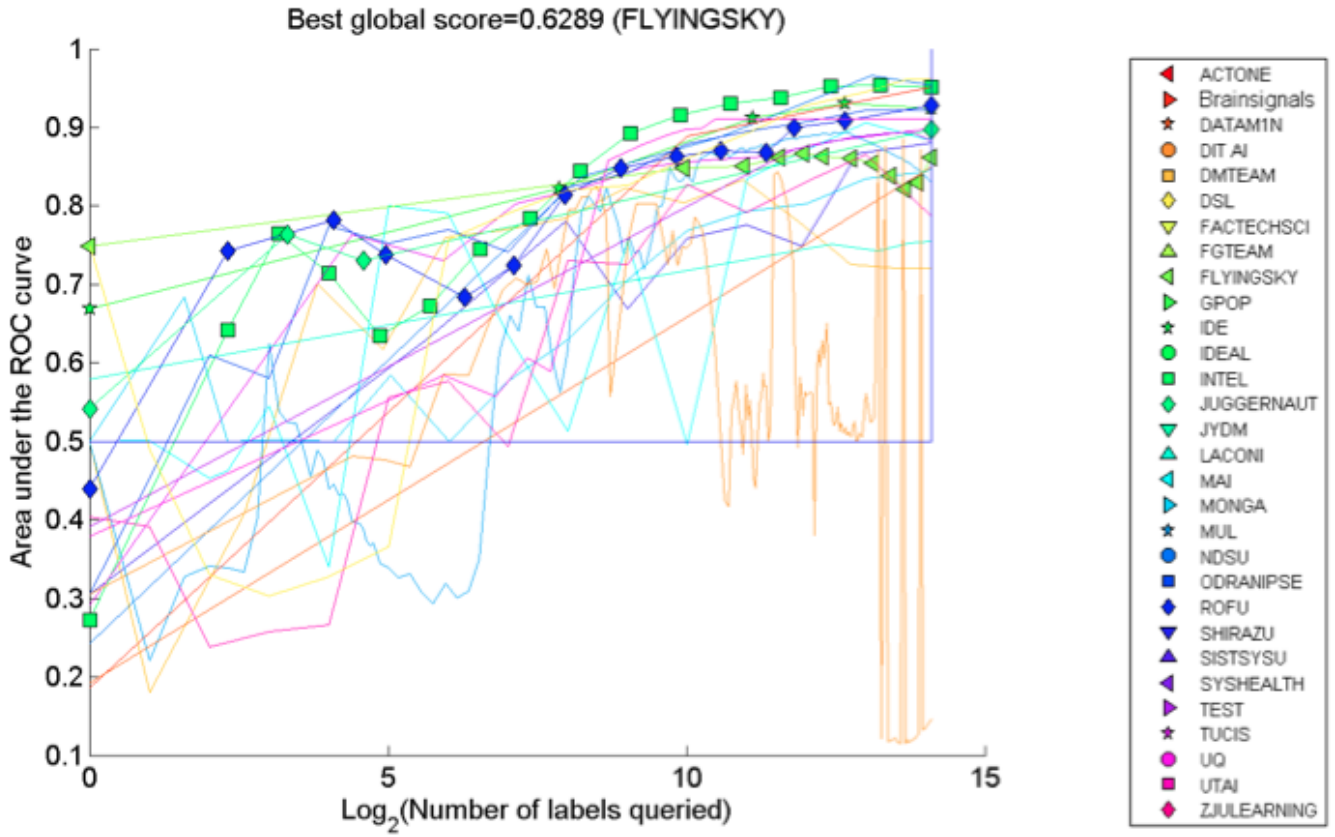


Figure 2: *Learning curves for dataset A.*

Ergebnisse Dataset B

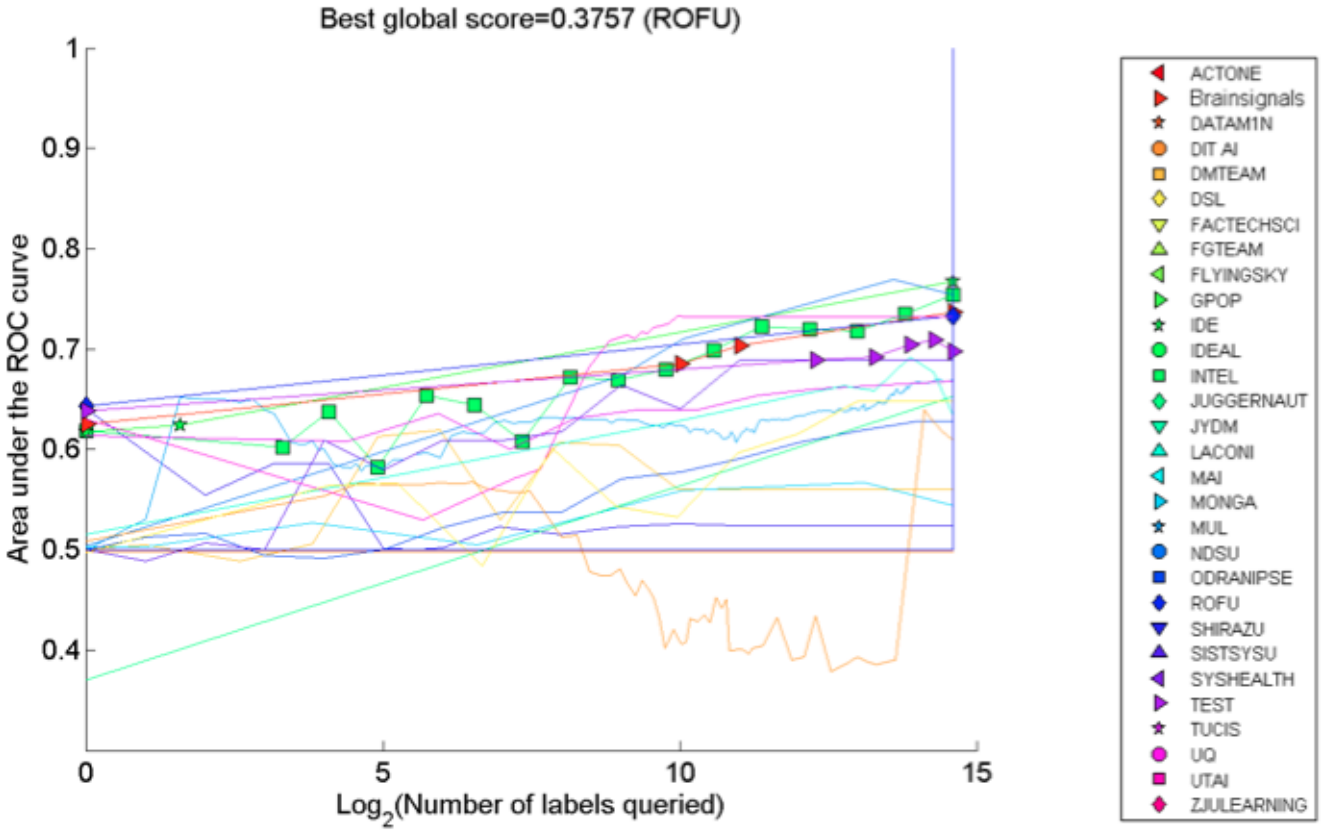


Figure 3: *Learning curves for dataset B.*

Ergebnisse Dataset C

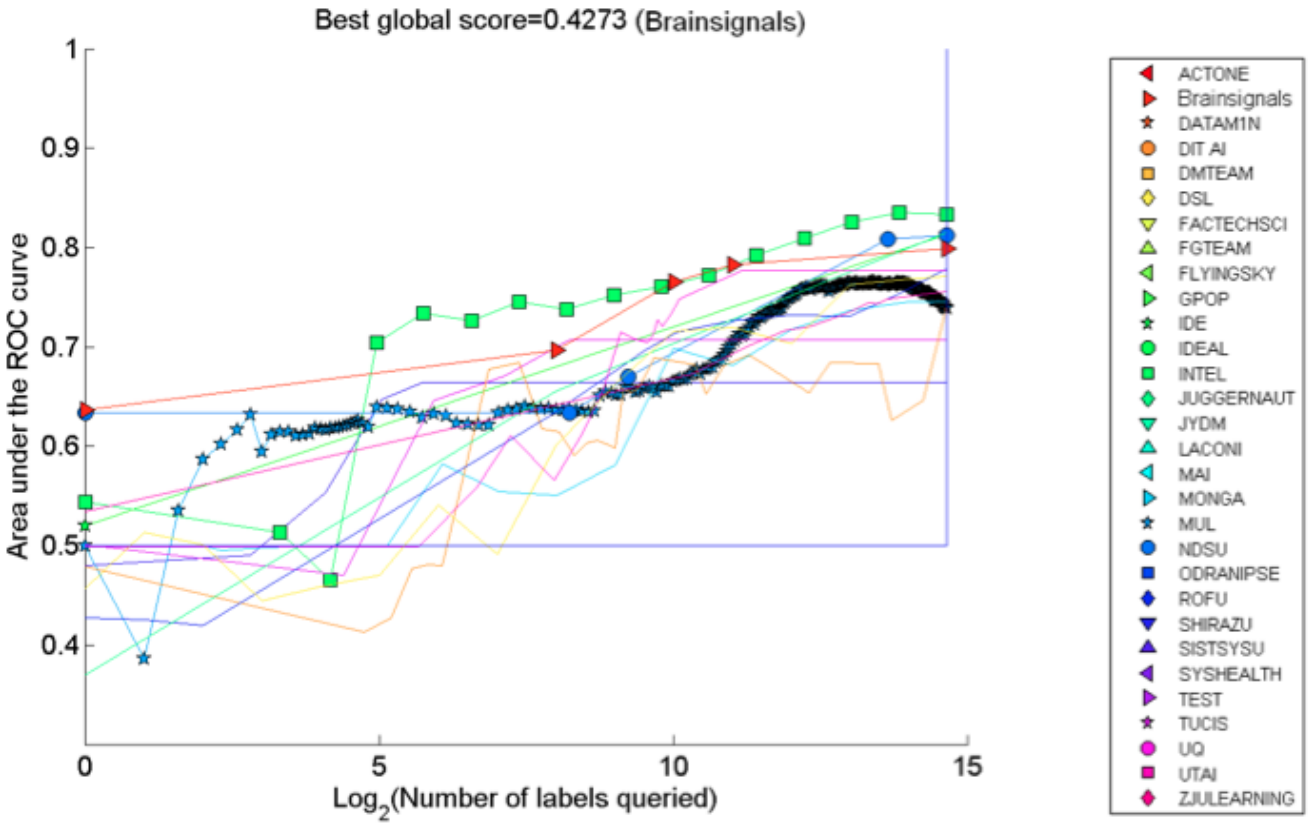


Figure 4: *Learning curves for dataset C.*

Ergebnisse Dataset D

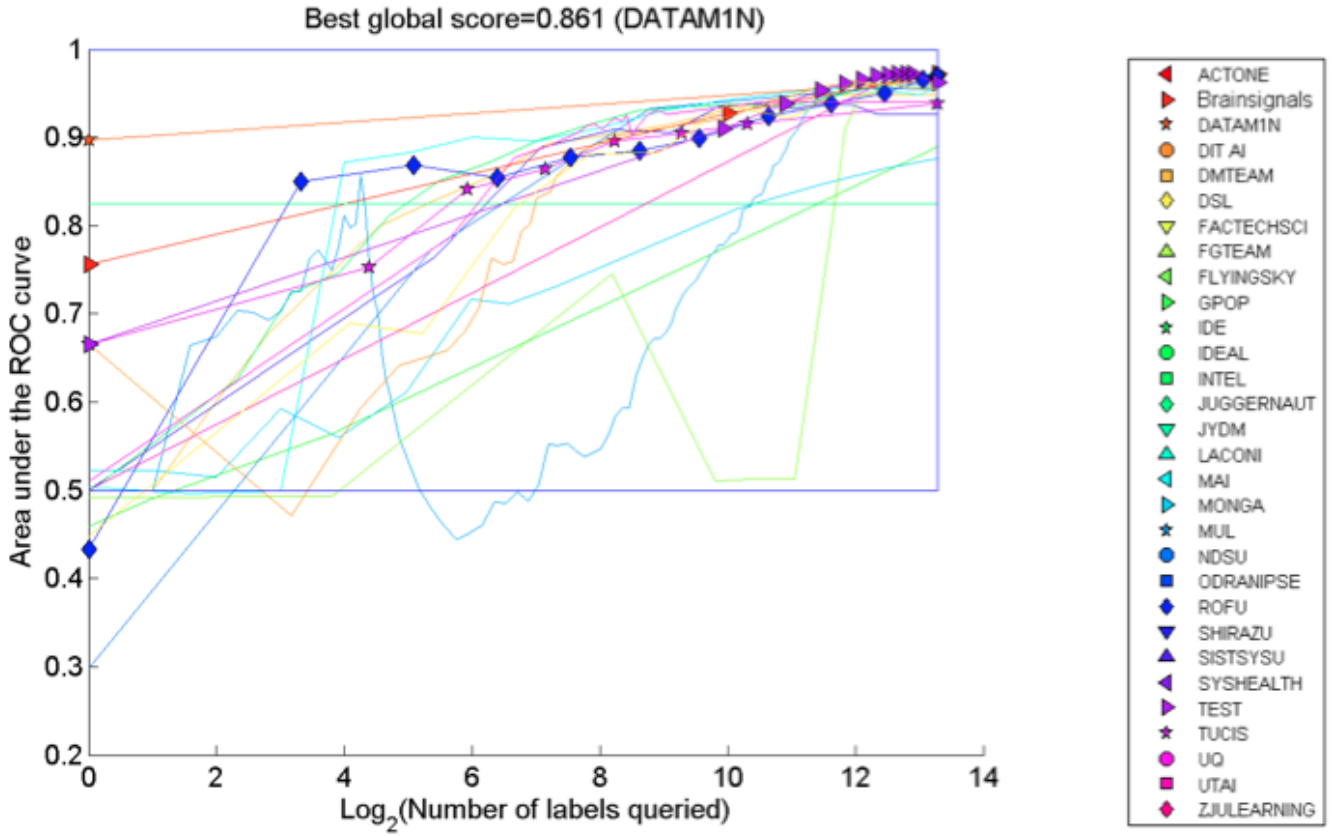


Figure 5: *Learning curves for dataset D.*

Ergebnisse Dataset E

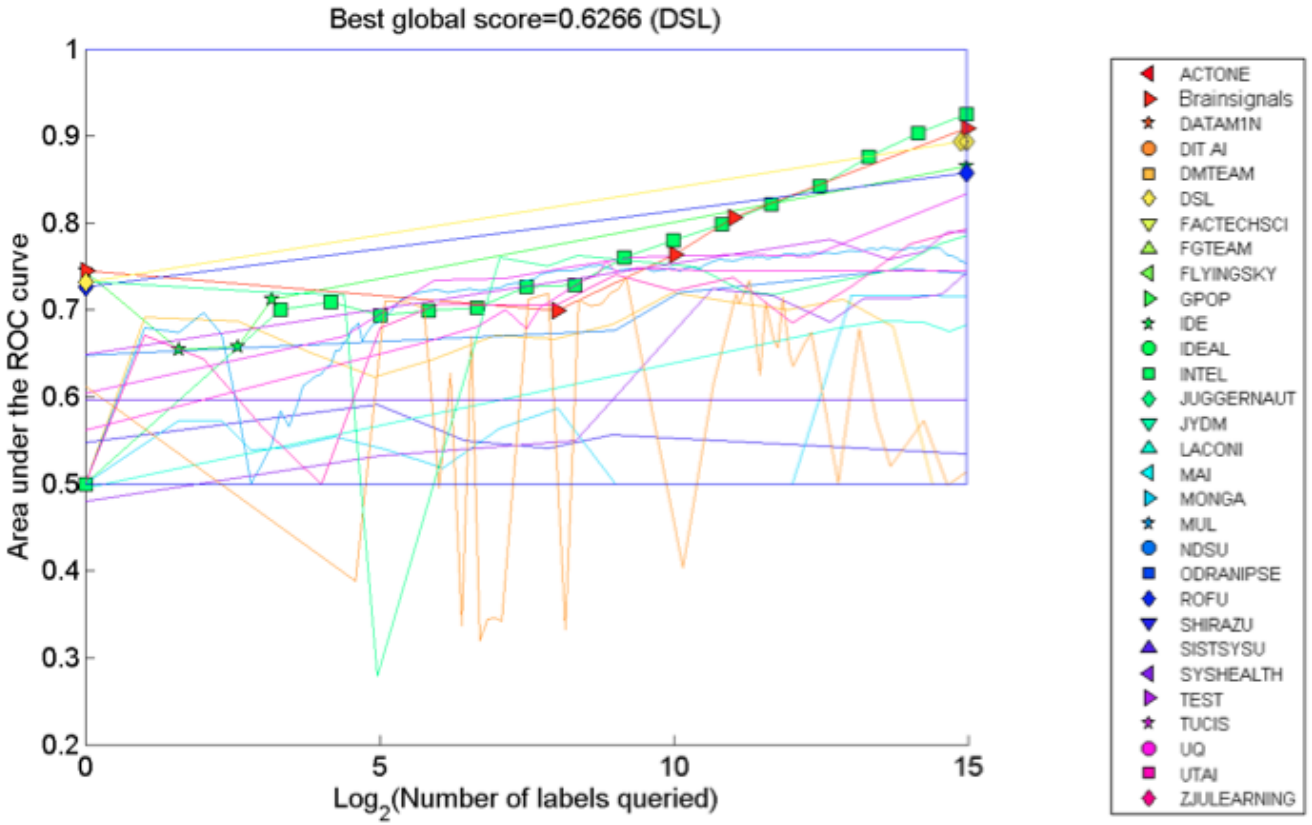


Figure 6: *Learning curves for dataset E.*

Ergebnisse des Wettbewerbs

Ergebnisse Dataset F

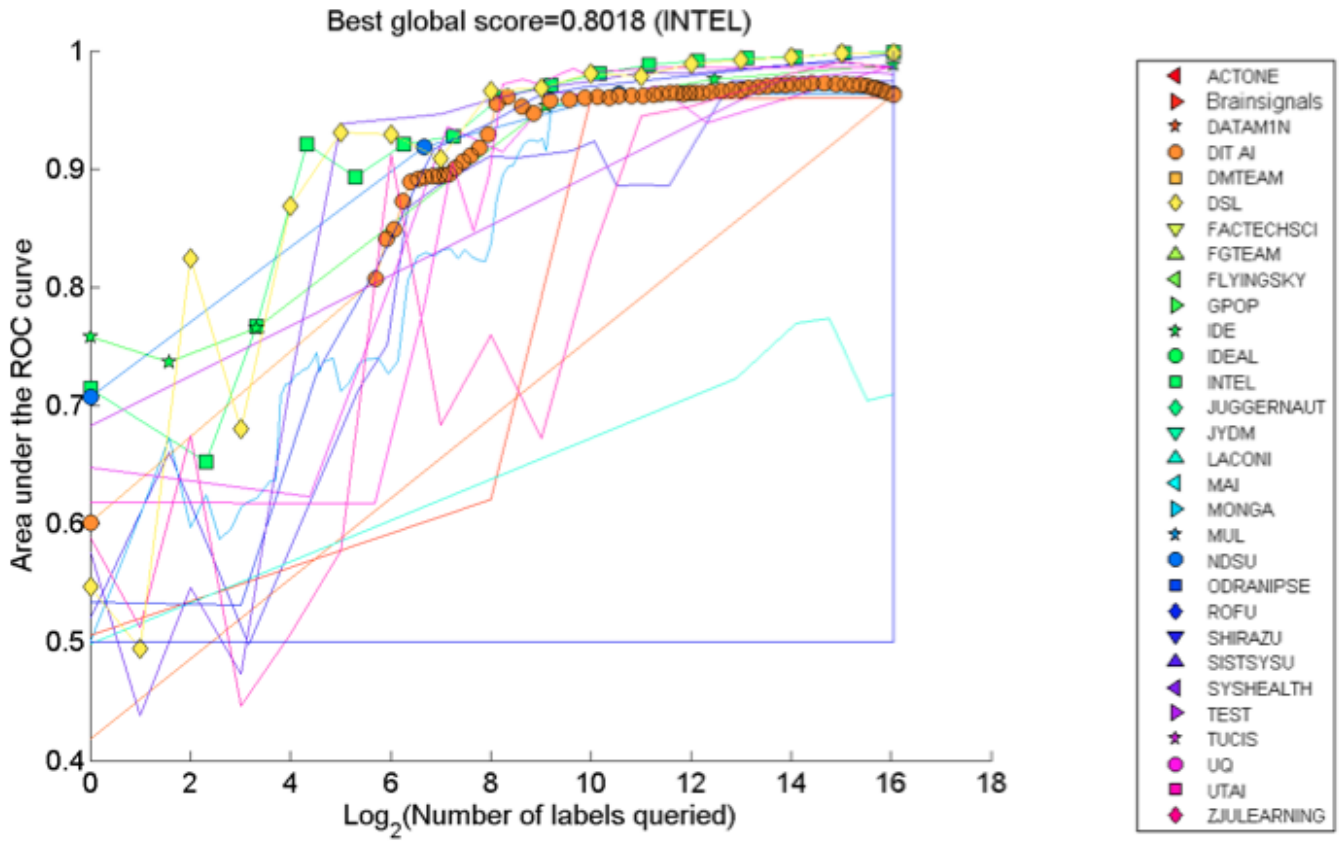


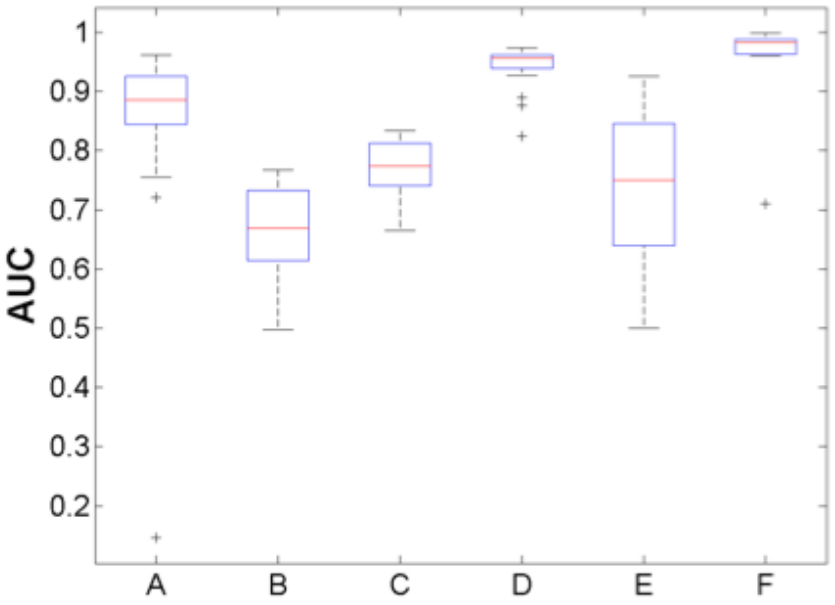
Figure 7: *Learning curves for dataset F.*

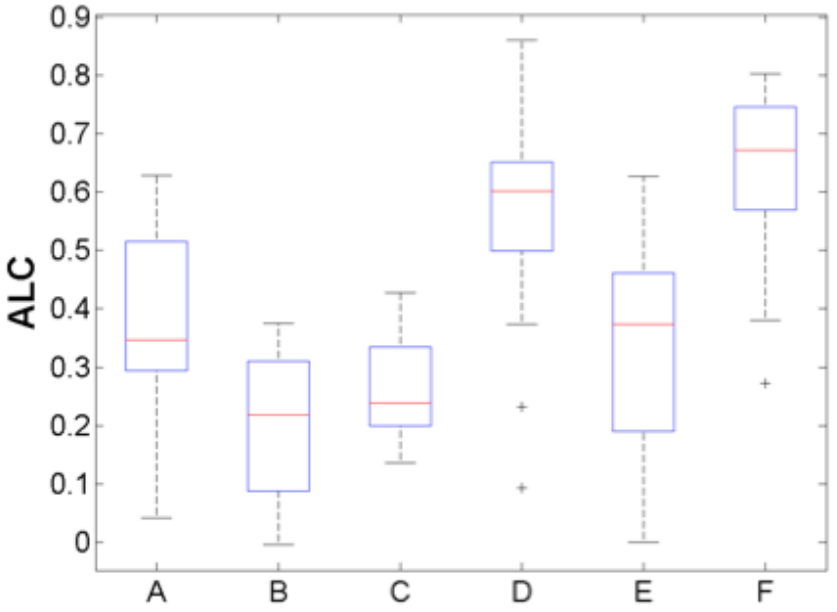
Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

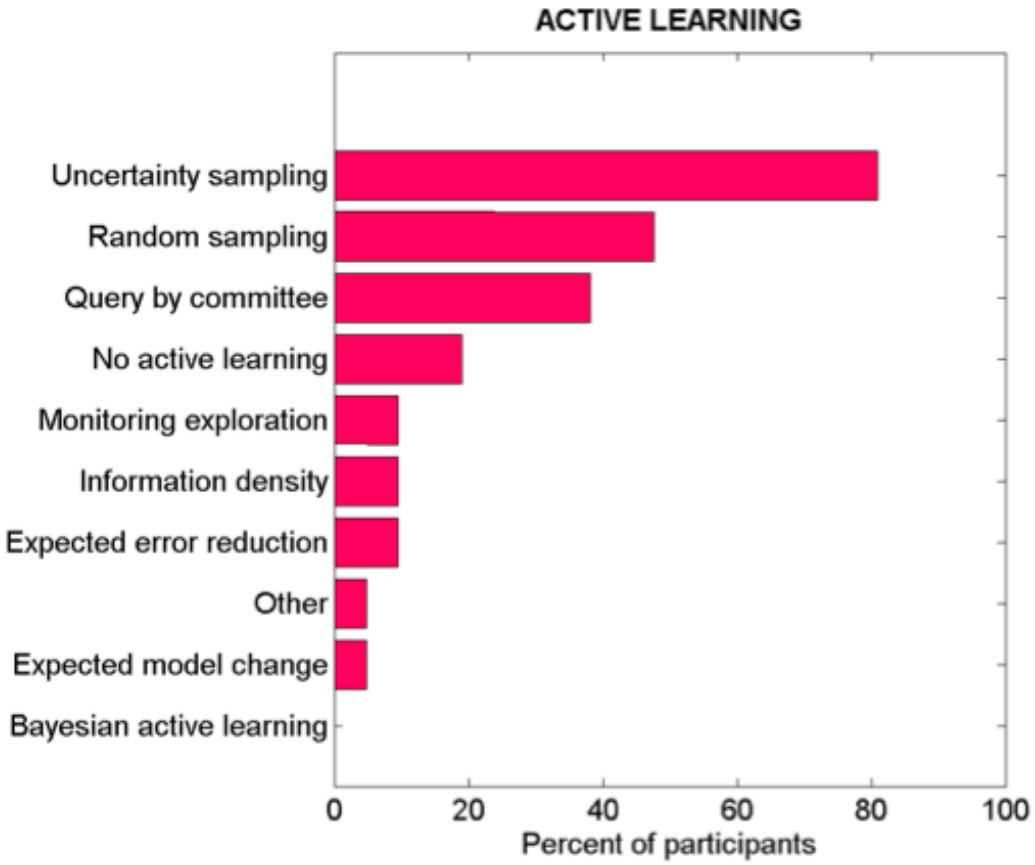
1) Einführung und Grundlagen
2) Beschreibung des Wettbewerbs
3) Übersicht der Ergebnisse
4) **Eigenschaften der Ergebnisse**
5) Fazit

Verteilung der Ergebnisse



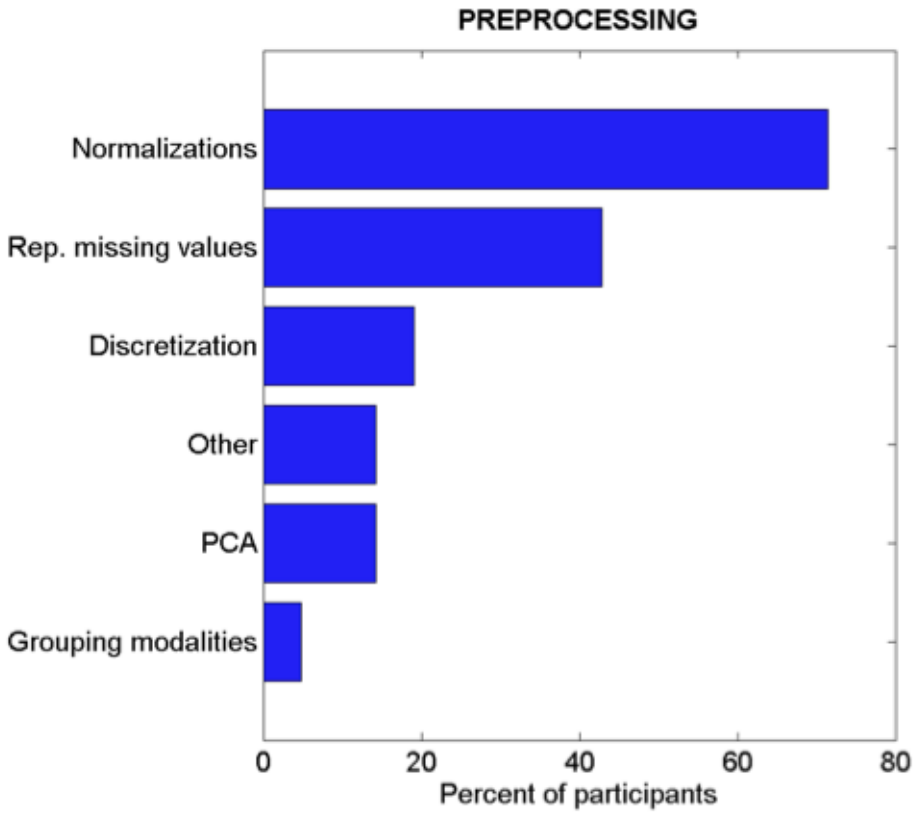


Active Learning Strategien



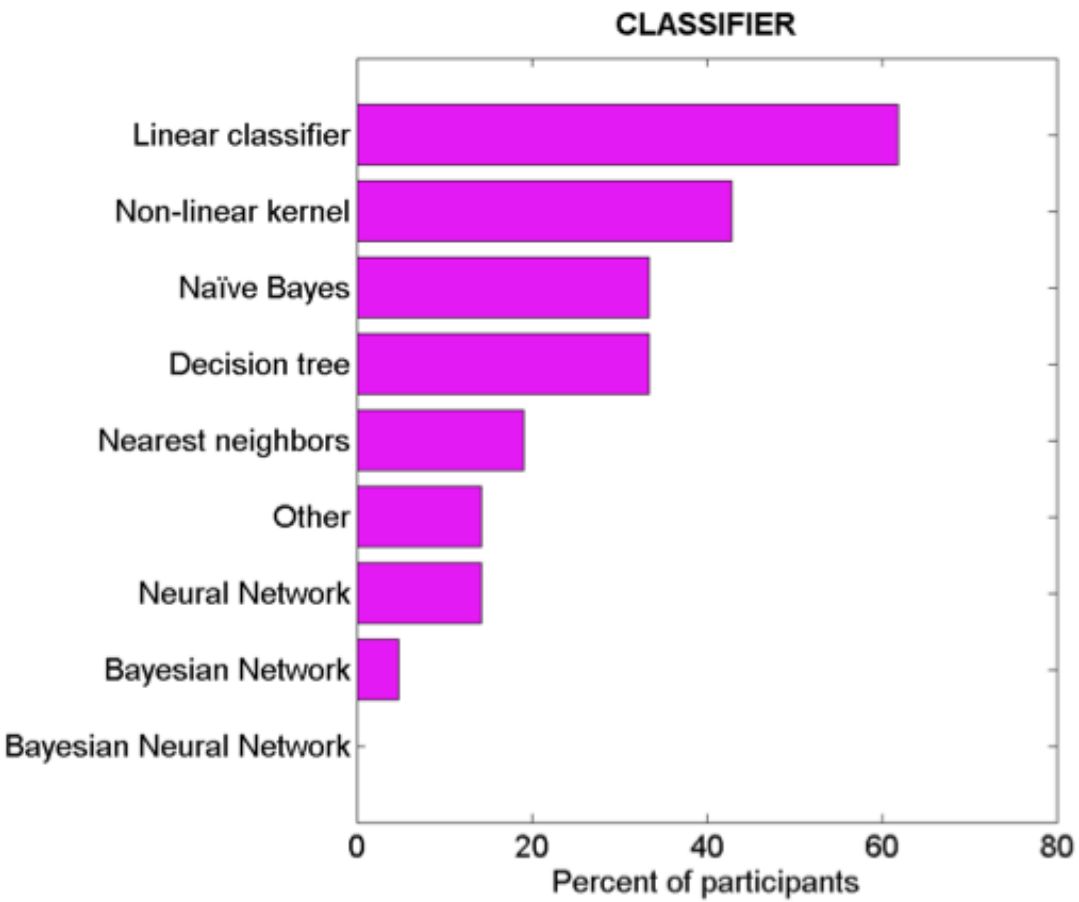
Eigenschaften der Ergebnisse

Pre-Processing



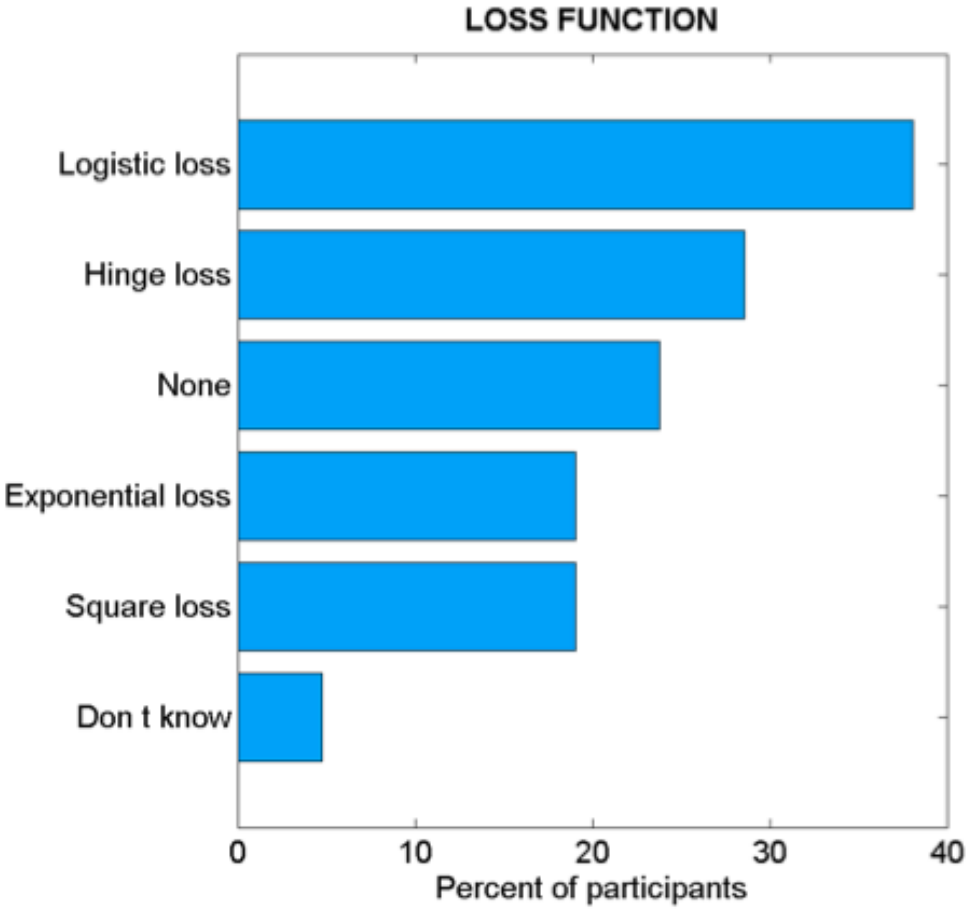
Eigenschaften der Ergebnisse

Klassifizierer



Eigenschaften der Ergebnisse

Loss Funktionen



Agenda

TECHNISCHE UNIVERSITÄT DARMSTADT

1) Einführung und Grundlagen
2) Beschreibung des Wettbewerbs
3) Übersicht der Ergebnisse
4) Eigenschaften der Ergebnisse
5) **Fazit**

Teilnahme und Gewinner

TECHNISCHE UNIVERSITÄT DARMSTADT

- Mehr als 300 Teilnehmer registriert (Development Phase)
- Mehr als 30 Teams für die Test-Phase
  - Teamgröße zwischen 1 und 20
  - 50% der Teams haben länger als 2 Wochen „programmiert“

- Kein Team hat mehr als ein Datenset gewonnen

- Overall ranking:
  - 1st: Intel team avg. rank 4.2
  - 2nd: RUFO team avg. rank 4.8
  - 3rd: IDE team avg. rank 5.7

Ergebnistabellen

Table 4: Result tables for the top ranking teams.

Dataset A		
Team	AUC (Ebar)	ALC
Flyingsky	0.8622 (0.0049)	0.6289
IDE	0.9250 (0.0044)	0.6040
ROFU	0.9281 (0.0040)	0.5533
JUGGERNAUT	0.8977 (0.0036)	0.5410
Intel	0.9520 (0.0045)	0.5273

Dataset B		
Team	AUC (Ebar)	ALC
ROFU	0.7327 (0.0034)	0.3757
IDE	0.7670 (0.0038)	0.3754
Brainsignals	0.7367 (0.0043)	0.3481
TEST	0.6980 (0.0044)	0.3383
Intel	0.7544 (0.0044)	0.3173

Dataset C		
Team	AUC (Ebar)	ALC
Brainsignals	0.7994 (0.0053)	0.4273
Intel	0.8333 (0.0050)	0.3806
NDSU	0.8124 (0.0050)	0.3583
IDE	0.8137 (0.0051)	0.3341
MUL	0.7387 (0.0053)	0.2840

Dataset D		
Team	AUC (Ebar)	ALC
DATAM1N	0.9641 (0.0033)	0.8610
Brainsignals	0.9717 (0.0033)	0.7373
ROFU	0.9701 (0.0032)	0.6618
TEST	0.9623 (0.0033)	0.6576
TUCIS	0.9385 (0.0037)	0.6519

Dataset E		
Team	AUC (Ebar)	ALC
DSL	0.8939 (0.0039)	0.6266
ROFU	0.8573 (0.0043)	0.5838
IDE	0.8650 (0.0042)	0.5329
Brainsignals	0.9090 (0.0039)	0.5267
Intel	0.9253 (0.0037)	0.4731

Dataset F		
Team	AUC (Ebar)	ALC
Intel	0.9990 (0.0009)	0.8018
NDSU	0.9634 (0.0018)	0.7912
DSL	0.9976 (0.0009)	0.7853
IDE	0.9883 (0.0013)	0.7714
DIT AI	0.9627 (0.0017)	0.7216

Fazit

Warum ist random sampling so gut?

TECHNISCHE UNIVERSITÄT DARMSTADT

Uncertainty sampling und query-by-commitee (ohne randomness) schneiden schlechter ab als random sampling

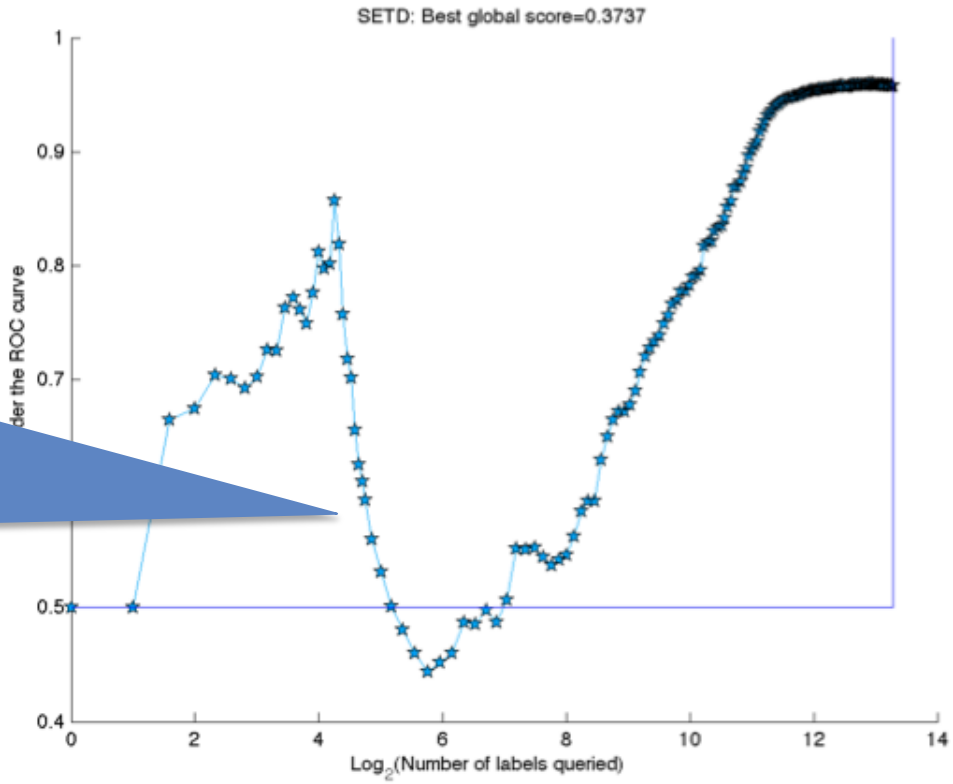


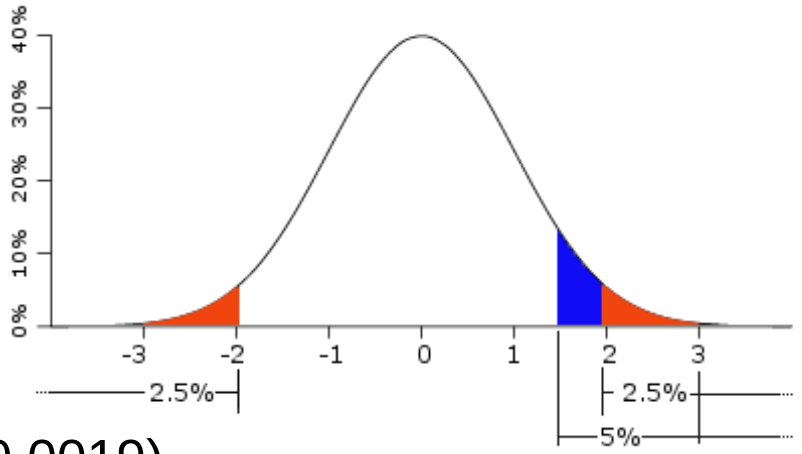
Figure 8: *Example of learning curves for dataset D using the uncertainty sampling strategy.*

Fazit

Gibt es signifikante Aussagen?

TECHNISCHE UNIVERSITÄT DARMSTADT

- **top ranked vs. lowest ranked**
- Friedman test
  - Avg. Platzierungen je Dataset
  - H0: Average ist gleich
  - Full-Score-Matrix notwendig



→ Hoch signifikant! (P-value: 0.0019)

- **Unterschiede zwischen Algorithmus A und allen anderen**
- Nemenyi test (post-hoc)
  - Nur das erste (Intel) und letzte (DIT AI) sind signifikant unterschiedlich

PostChallenge Experiments

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Ziel: Decoupling**
- Learning Strategie vom Klassifizierer unabhängig
  - Random sampling (passive learning)
- 10 Durchläufe (Varianzreduktion)

- **Resultat:**
  - Tree Klassifizierer eher schlecht (vor allem am Anfang)
  - Ensemble of Trees sind gut
  - Generativ (naive Bayes) besser als Discriminativ (am Anfang)

→ Wahl des Klassifizierers wohl wichtiger als Active Learning Strategie

Fazit

Zusammenfassung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Hohe Teilnahme, trotz Komplexität
- Viele verschiedene Techniken ausprobiert
- Eher explorativ: Learn-Strategie und Klassifizierer wurden vermischt
  - Müsstenxw getrennt betrachtet werden
- Wahl des Klassifizierers sehr wichtig
  - Strategien zur Klassifizierer-Wahl je nach Anzahl an Examples
- Semi-supervised learning am Anfang der Lernkurve (Seed)
- Randomizierung notwendig, um gute Resultate zu erreichen (over-fitting)

- → Active Learning Strategien müssen noch intensiver untersucht werden

Danke!

TECHNISCHE UNIVERSITÄT DARMSTADT

Danke für die Aufmerksamkeit!