

Effiziente Multilabel-Klassifizierung durch paarweises Lernen

Verteidigung der Doktorarbeit
Efficient Pairwise Multilabel Classification
24.07.2012

Einleitung

- Multilabel-Klassifizierung

- Zerlegung

- Herausforderungen

Hohe Dimensionalität des Klassenraums

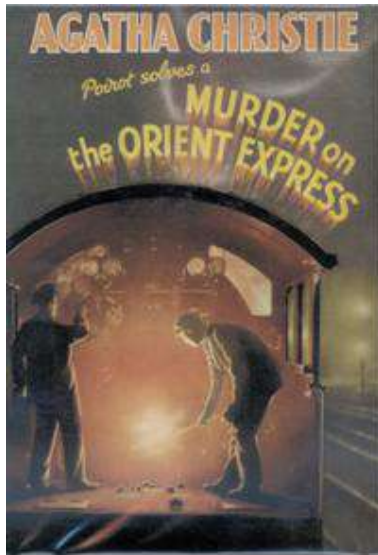
- Quick Voting

- EUR-Lex Datensatz und Dual MLPP

- Paarweises HOMER

Zusammenfassung

Multilabel-Szenario



Genres:

Crime
Mystery
Thriller

- Zuordnung eines Objekts x zu beliebig vielen Labels aus einer Labelmenge Y
- im Gegensatz zu
 - *Multiclass* Klassifizierung: Zuordnung zu genau einer Klasse
 - *Binäre* Klassifizierung: Zuordnung zu einer von zwei möglichen Klassen

Summary: Returning from an important case in Syria, Hercule Poirot boards the Orient Express in Istanbul. The train is unusually crowded for the time of year. Poirot secures a berth only with difficulty.

▪ **Text:** Indexierung von Nachrichten, Webseiten, Blogs, ... mit Schlüsselwörtern, Themen, Genres, Autoren, Sprachen, ...

▪ **Multimedia:** Erkennung von Szenen/Objekten (Bilder), Instrumente, Emotionen, Musikstile (Audio)

▪ **Biologie:** Klassifizierung von Funktionen von Genomen und Proteinen

Lernen von Klassifikationen: Formale Definition

Gegeben:

- eine Menge von Trainingsobjekten $\{x_1, \dots, x_m\}$, x_i Vektoren in \mathbb{R}^a
- eine Menge von Label-Zuordnungen $\{y_1, \dots, y_m\}$
 - binär: $y_i \in Y = \{-1, 1\}$ z.B.: $y_i = \{non_thriller\}, \{thriller\}$
 - multiclass: $y_i \in Y = \{\lambda_1, \dots, \lambda_n\}$ $y_i = \{thriller\}, \{mystery\}$
 - multilabel: $y_i \subseteq Y = \{\lambda_1, \dots, \lambda_n\}$ $y_i = \{thriller, mystery\}, \{\}, Y$

Ziel:

- Finde eine Funktion $h: \mathbb{R}^a \rightarrow Y$ welche x_i auf y_i abbildet
 - so akkurat wie möglich
 - so effizient wie möglich

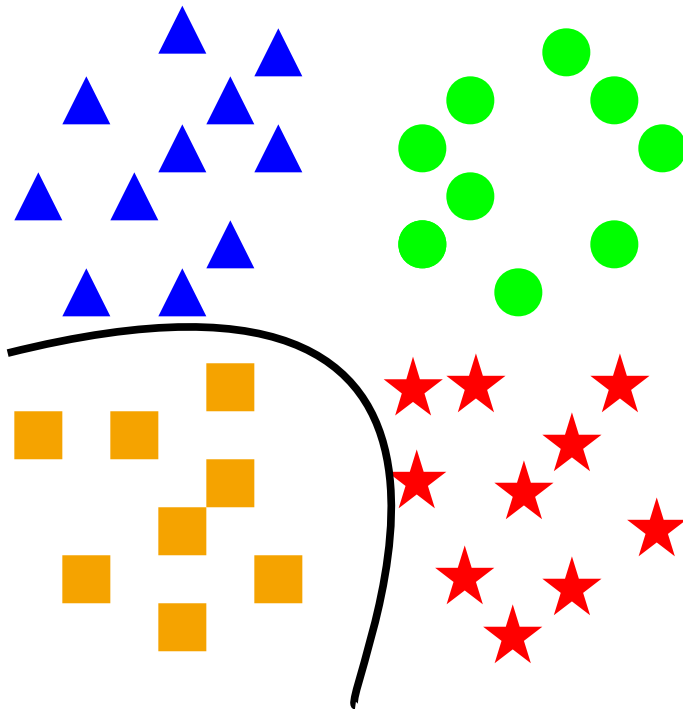
Wichtigste Lösungsansätze für Multilabel-Probleme:

- Anpassung bestehender Lern-Verfahren an Multilabel-Daten
 - nicht trivial und oft nicht möglich
- Zerlegung von Multilabel-Problemen in binäre Probleme
 - wohlbekanntes Szenario, klare Semantik
 - viele gute binäre Lerner vorhanden: SVMs, Regellerner

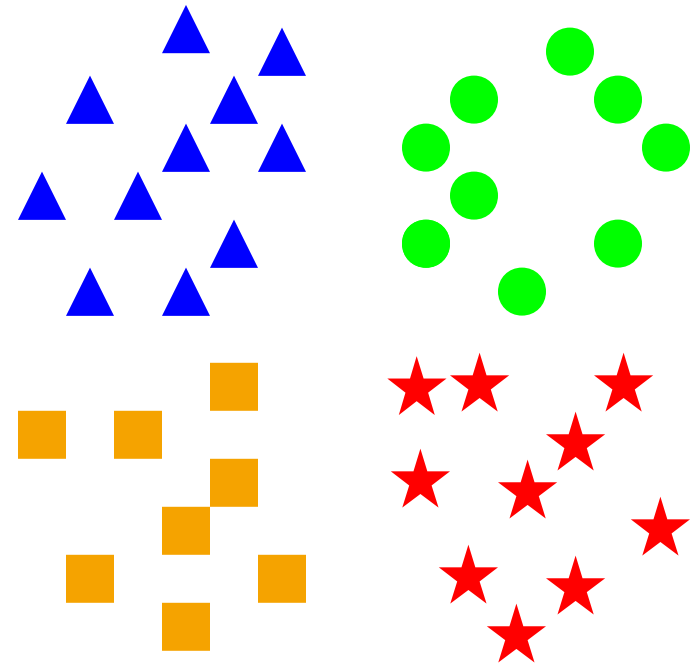
Zwei konkurrierende Dekompositions-Ansätze:

- Binary Relevance Zerlegung: lerne einen Klassifizierer für *jede Klasse*
 - alias One-against-all
- Paarweise Zerlegung: lerne einen Klassifizierer für *jedes Klassenpaar*
 - alias One-against-one, Round Robin

Dekomposition



binary relevance decomposition
(für alle Klassen)

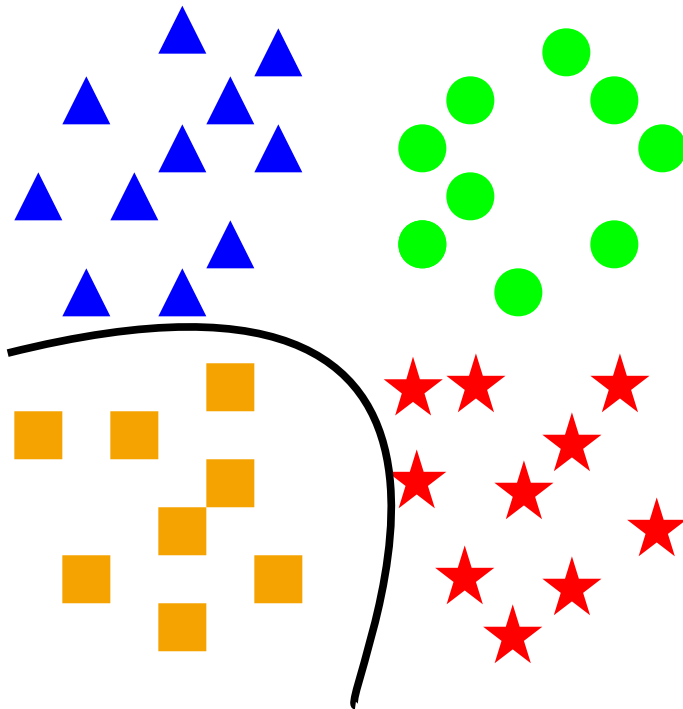


pairwise decomposition

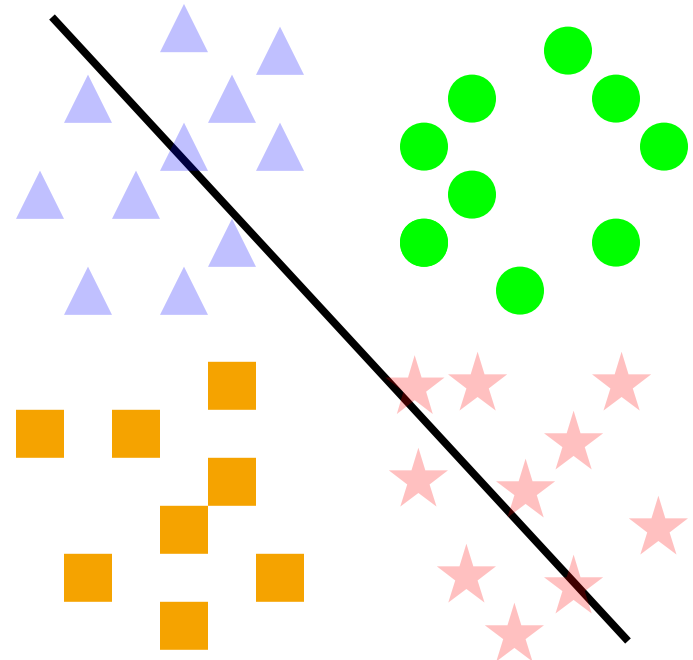
Dekomposition



TECHNISCHE
UNIVERSITÄT
DARMSTADT



binary relevance decomposition
(für alle Klassen)



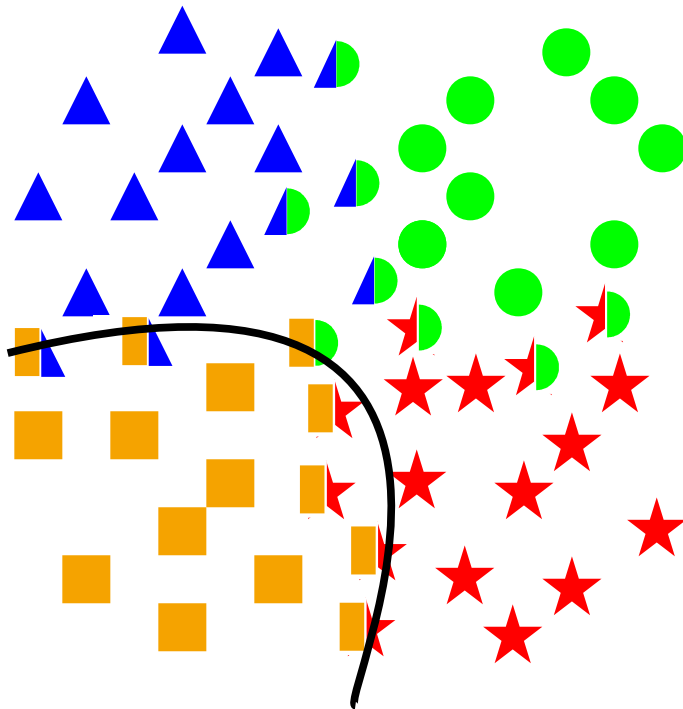
pairwise decomposition
(für alle Klassenpaare)

Dekomposition

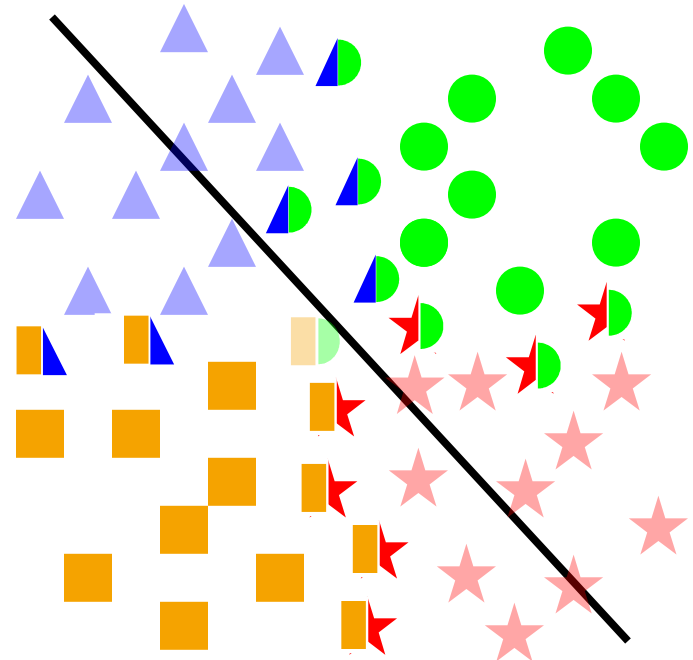
Multilabel



TECHNISCHE
UNIVERSITÄT
DARMSTADT



binary relevance decomposition
(für alle Klassen)



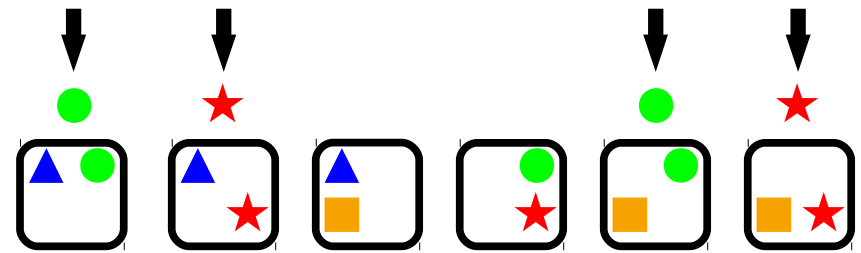
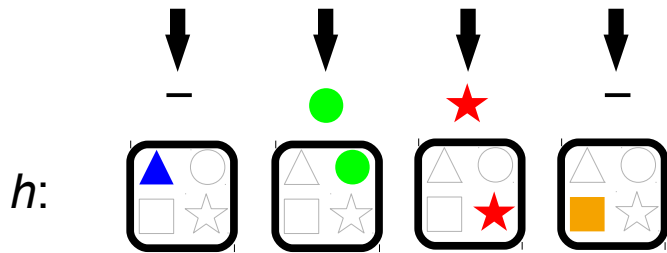
pairwise decomposition
(für alle Klassenpaare)

Dekomposition Training

in:

x_1 : 

x_1 : 




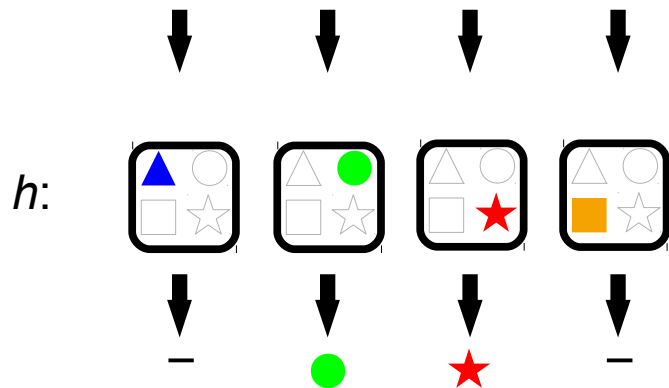
binary relevance decomposition

pairwise decomposition

Dekomposition Vorhersage

in:


x_1 : 

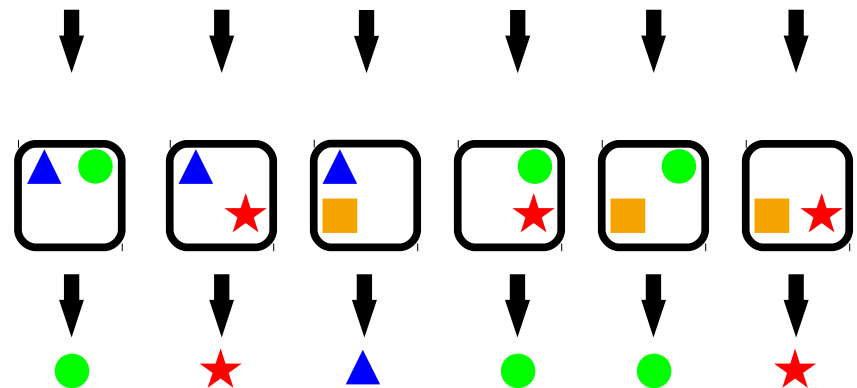


out:

binary relevance decomposition

x_1 : 



 :3  :2 |  :1  :0 Kalibrierung
(MLJ 2008)

pairwise decomposition

Vor- und Nachteile

Binary Relevance:

- gleichgroße Teilprobleme
 - schwer zu lernen
- + lineare Anzahl Teilprobleme
 - aber vergleichbare oder sogar höhere Trainingskosten
- Teilprobleme getrennt und unabhängig gelernt
 - Verlust von Informationen über Label-Abhängigkeiten
- + keine Parameter notwendig, funktioniert "out-of-the-box"

Paarweise Zerlegung:

- + kleine Teilprobleme
 - einfacher/schneller zu lernen
- quadratische Anzahl Teilprobleme
 - hohe Speichieranforderungen
 - hohe Kosten für Vorhersage
- + Beachtung von paarweisen Klassenabhängigkeiten
 - aber Verlust von Information in den Label-Schnittmengen
- + hoher Grad an Parallelisierung
- + Klassen-inkrementell

Herausforderungen der effizienten paarweisen Multilabel-Klassifizierung



TECHNISCHE
UNIVERSITÄT
DARMSTADT

▪ Dimensionalität des Klassenraums:

- Anzahl der Labels

Allgemeine Herausforderungen in Multilabel-Klassifizierung

- Größe und Umfang der Daten:
 - Anzahl Trainings- und Testbeispiele
- Verfügbarkeit von Daten:
 - Echtzeitverarbeitung
- Abhängigkeiten zwischen den Labels:
 - Ausnutzung von Label-Korrelationen

} in Dissertation

Ausgangssituation

Speicher: $O(n^2)$

- P_1 und P_2 auf y-Achse

Training: $O(d \cdot n)$

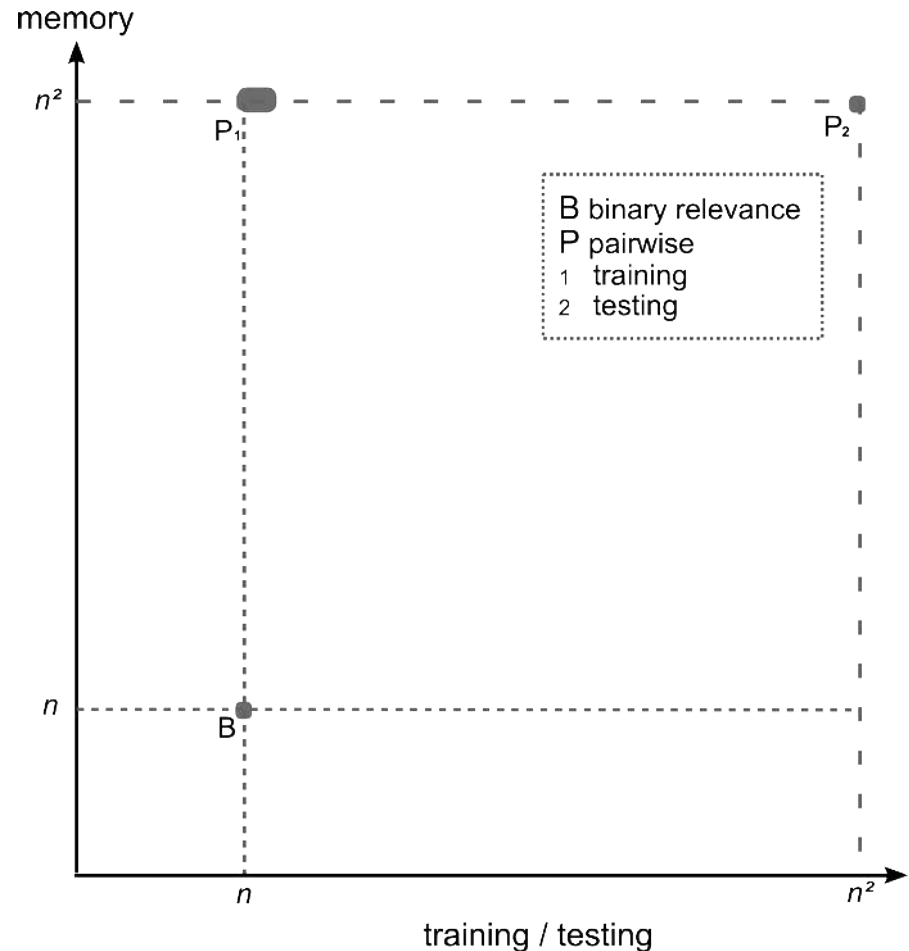
- P_1 auf x-Achse

Vorhersage: $O(n^2)$

- P_2 auf y-Achse

Ziel:

so nah wie möglich an
Binary Relevance (B)
kommen



Hohe Dimensionalität des Klassenraums



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Quick Voting (NC 2010)

Quick Voting

Hauptidee: viele Situationen, in denen ein bestimmtes Label nicht mehr gewinnen kann (Park 2007)

- Anzahl verlorene Vergleiche größer als Maximum, das ein relevantes Label haben kann
- ähnlich Situation in Sport, wenn der Gewinner schon vor Ende der Liga feststeht
 - weitere Spiele können ohne Einbußen ausgelassen werden

Beispiel aus der Fussball-WM 2006:

12 Jun	AUS : JPN	3:1
13 Jun	BRA : CRO	1:0
18 Jun	BRA : AUS	2:0
18 Jun	JPN : CRO	0:0
22 Jun	JPN : BRA	1:4
22 Jun	CRO : AUS	2:2

pos.	country	pts.
1.	Brazil	9
2.	Australia	4
3.	Croatia	2
4.	Japan	1

- Strategie: lasse bestes Team (Brasilien) zuerst alle Partien spielen
- Bester Fall: nur drei Spiele
- Schlimmster Fall: 6

Quick Voting

Hauptidee: viele Situationen, in denen ein bestimmtes Label nicht mehr gewinnen kann (Park 2007)

- Anzahl verlorene Vergleiche größer als Maximum, das ein relevantes Label haben kann
- ähnlich Situation in Sport, wenn der Gewinner schon vor Ende der Liga feststeht
 - weitere Spiele können ohne Einbußen ausgelassen werden

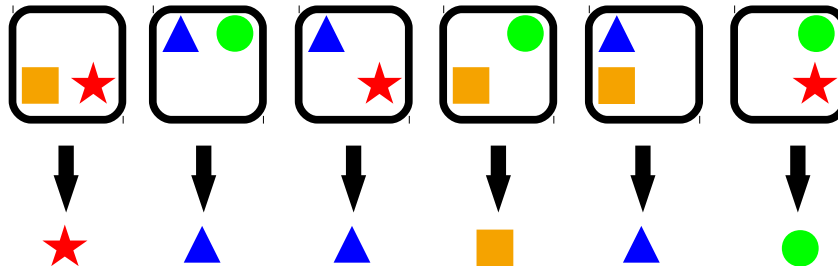
Beispiel aus der Fussball-WM 2006:

12 Jun	A★S : J■N	3:1
13 Jun	B▲A : C●O	1:0
18 Jun	B▲A : A★S	2:0
18 Jun	J■N : C●O	0:0
22 Jun	J■N : B▲A	1:4
22 Jun	C●O : A★S	2:2

pos.	country	pts.
1.	Brazil	9
2.	Australia	4
3.	Croatia	2
4.	Japan	1

- Strategie: lasse bestes Team (Brasilien) zuerst alle Partien spielen
- Bester Fall: nur drei Spiele
- Schlimmster Fall: 6

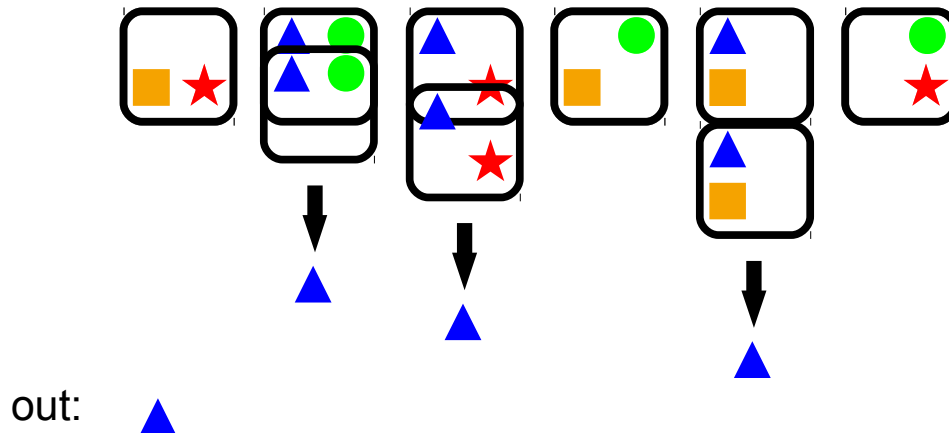
Vollständiges Voting



out: ▲

- Strategie: lasse bestes Team (Brasilien) zuerst alle Partien spielen
- Bester Fall: nur drei Spiele
- Schlimmster Fall: 6

Quick Voting



- Strategie: lasse bestes Team (Brasilien) zuerst alle Partien spielen
- Bester Fall: nur drei Spiele
- Schlimmster Fall: 6

Resultate

Vorhersage:

von $O(n^2)$ auf $\sim O(d n \log n)$
in der Praxis

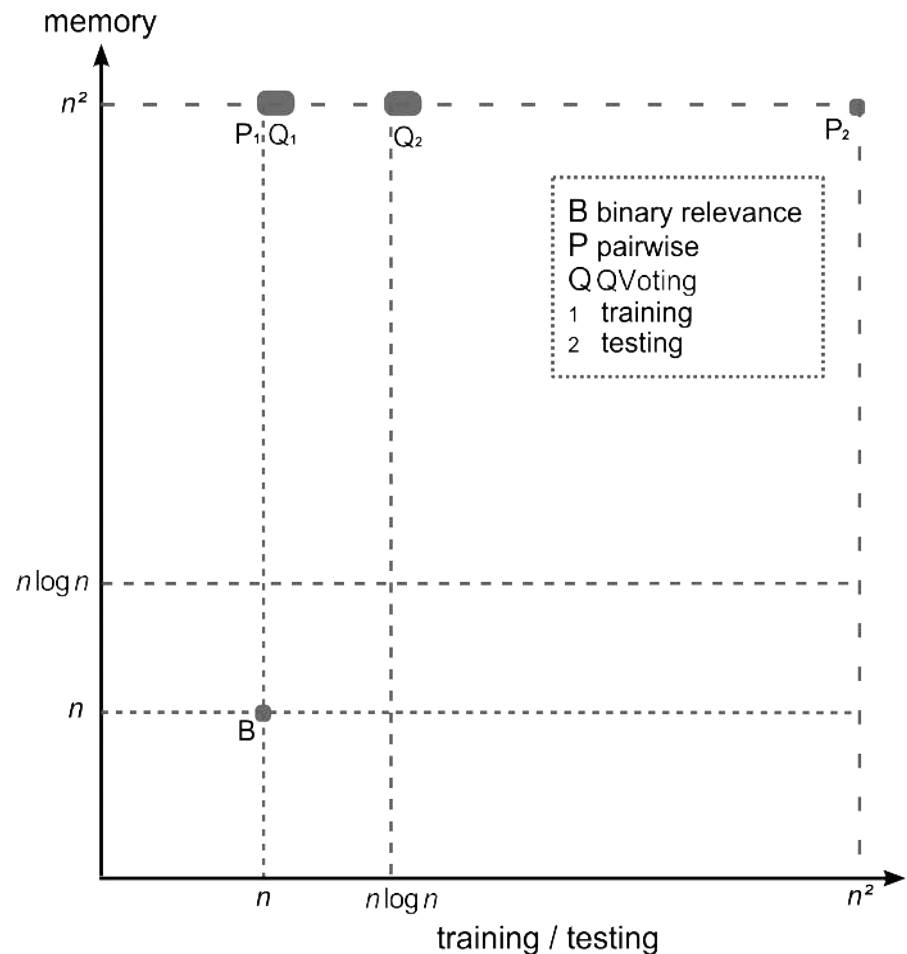
- ohne Änderung der Vorhersage

Speicher:

immer noch $O(n^2)$

Vorhersagequalität:

paarweiser Ansatz durchweg
besser als BR



Hohe Dimensionalität des Klassenraums



TECHNISCHE
UNIVERSITÄT
DARMSTADT

EUR-Lex Datensatz Dual MLPP

(ECML 2008)

- 19328 (frei verfügbare) Dokumente aus dem *Directory of Community legislation in force* der Europäischen Union
 - Texte in vielen europäischen Sprachen verfügbar
- mehrere Klassifikationen für jedes einzelne Dokument verfügbar

EUR-Lex repository



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Title and reference

Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs

Classifications

EUROVOC descriptor

- data-processing law
- computer piracy
- copyright
- software
- approximation of laws

Directory Code:

- Law relating to undertakings/IPR Law

Subject matter:

- Internal market
- Industrial and commercial property

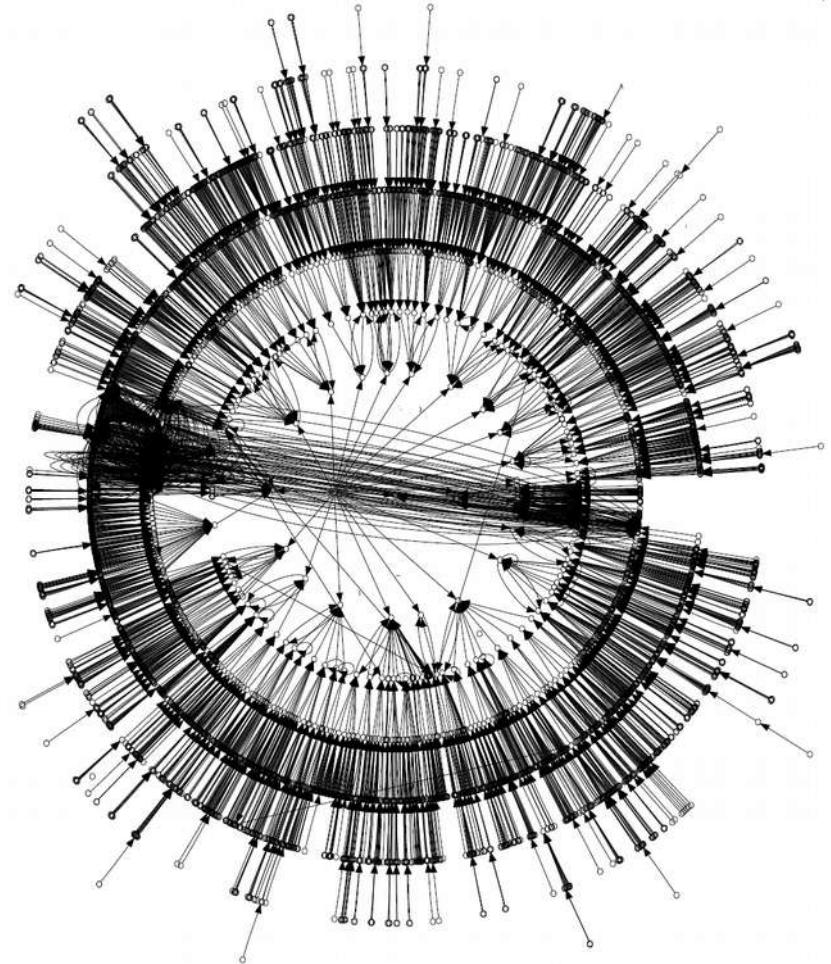
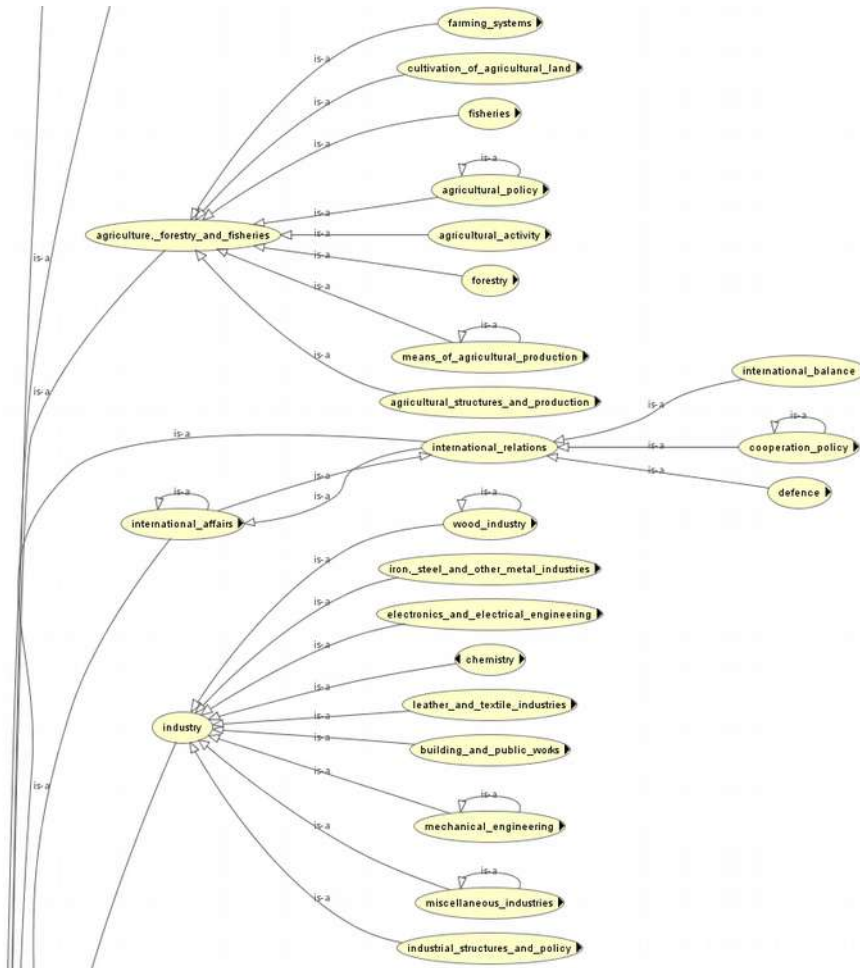
Text

COUNCIL DIRECTIVE of 14 May 1991 on the legal protection of computer programs (91/250/EEC)

THE COUNCIL OF THE EU,

Having regard to the Treaty establishing the European Economic Community and in particular Article 100a thereof,
Having regard to the proposal of the Commission (1), ...

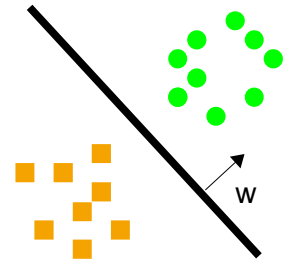
- 19328 (frei verfügbare) Dokumente aus dem *Directory of Community legislation in force* der Europäischen Union
 - Texte in vielen europäischen Sprachen verfügbar
- mehrere Klassifikationen für jedes einzelne Dokument verfügbar
- die anspruchsvollste:
Deskriptoren aus der **EUROVOC** Taxonomie
 - **3965** Deskriptoren, durchschnittlich 5,37 Labels pro Dokument
 - Deskriptoren sind in einer Hierarchie mit bis zu 7 Ebenen organisiert



- 19328 (frei verfügbare) Dokumente aus dem *Directory of Community legislation in force* der Europäischen Union
 - Texte in vielen europäischen Sprachen verfügbar
- mehrere Klassifikationen für jedes einzelne Dokument verfügbar
- die anspruchsvollste:
Deskriptoren aus der **EUROVOC** Taxonomie
 - **3965** Deskriptoren, durchschnittlich 5,37 Labels pro Dokument
 - Deskriptoren sind in einer Hierarchie mit bis zu 7 Ebenen organisiert
 - aber hier: Hierarchie wird ignoriert!
 - EUROVOC wird als *flaches* Multilabel-Problem betrachtet
 - wie in *Folksonomies* oder Keyword Indexing/Tagging

Ziel: EUROVOC-Problem durch paarweises Lernen lösen

- Perzeptron-Algorithmus:
 - lernt eine trennende lineare Hyperebene w zwischen zwei Punktmengen $h'(\mathbf{x}) := \mathbf{x} \cdot \mathbf{w}$
 - einfach und schnell
 - inkrementell lernbar, d.h. für Echtzeitszenarien geeignet
 - effiziente u. effektive Alternative zu SVMs für Textklassifizierung
- 8.000.000 Klassifizierer im Speicher!
 - 152 GB (statt 1,15 GB für BR)
- Training und Klassifizierung wären nicht so problematisch
 - Training nur $\sim 5x$, Klassifizierung $\sim 20x$ langsamer



→ Lösung: duale Form der paarweisen linearen Basisklassifizierer

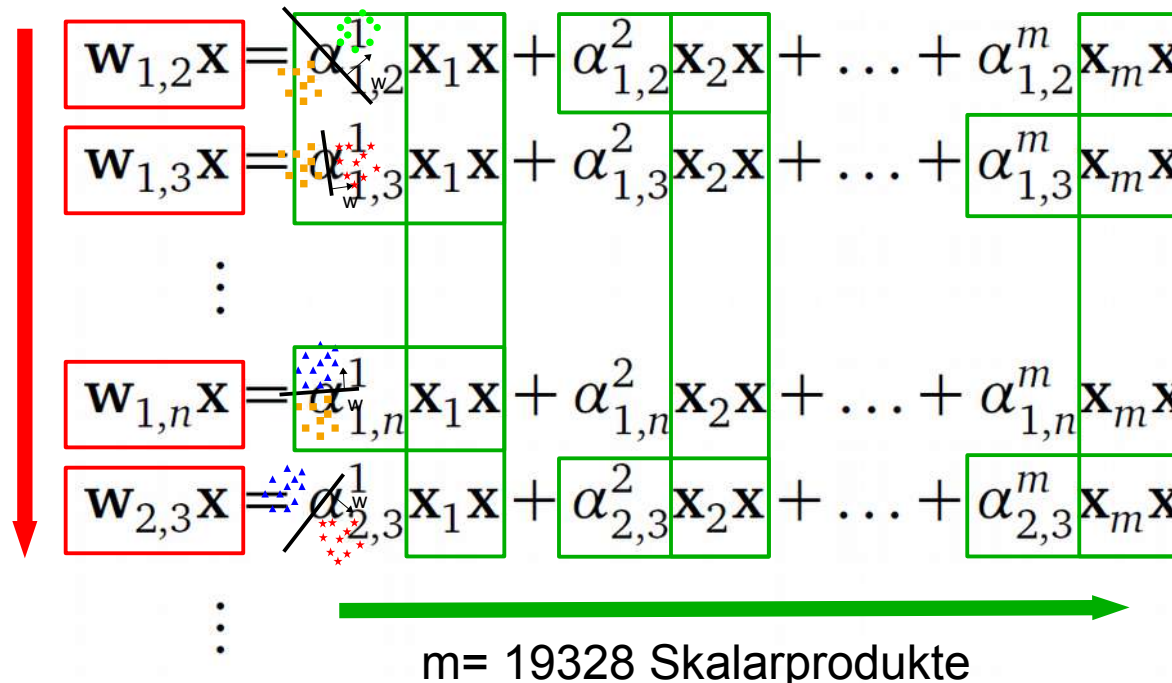
- Perzeptrons können als Linearkombination der Trainingbeispiele formuliert werden

$$\mathbf{w} = \sum_{i=1}^m \alpha_i \mathbf{x}_i$$

Dual Multilabel Pairwise Perceptrons

- zusätzliche Schleife über Trainingsbeispiele ist notwendig, aber
 - $x_i x$ können für alle n^2 Klassifizierer gleichzeitig berechnet werden
 - α 's sind dünn besetzt, nur $O(dn)$ pro Trainingsbeispiel relevant
 - davon nicht alle gebraucht, bei EUROVOC nur ~ 1.78 pro Spalte

$n^2 = 8$ Mio.
Skalar-
produkte



$$\begin{aligned}
 \mathbf{w}_{1,2}\mathbf{x} &= \alpha_{1,2}^1 \mathbf{x}_1 \mathbf{x} + \alpha_{1,2}^2 \mathbf{x}_2 \mathbf{x} + \dots + \alpha_{1,2}^m \mathbf{x}_m \mathbf{x} \\
 \mathbf{w}_{1,3}\mathbf{x} &= \alpha_{1,3}^1 \mathbf{x}_1 \mathbf{x} + \alpha_{1,3}^2 \mathbf{x}_2 \mathbf{x} + \dots + \alpha_{1,3}^m \mathbf{x}_m \mathbf{x} \\
 &\vdots \\
 \mathbf{w}_{1,n}\mathbf{x} &= \alpha_{1,n}^1 \mathbf{x}_1 \mathbf{x} + \alpha_{1,n}^2 \mathbf{x}_2 \mathbf{x} + \dots + \alpha_{1,n}^m \mathbf{x}_m \mathbf{x} \\
 \mathbf{w}_{2,3}\mathbf{x} &= \alpha_{2,3}^1 \mathbf{x}_1 \mathbf{x} + \alpha_{2,3}^2 \mathbf{x}_2 \mathbf{x} + \dots + \alpha_{2,3}^m \mathbf{x}_m \mathbf{x} \\
 &\vdots
 \end{aligned}$$

$m = 19328$ Skalarprodukte

Experimente auf EUROVOC

Speicherverbrauch nun vergleichbar ($\sim 1,4$ GB zu BR: $\sim 1,15$ GB)

- speichern der Trainingsbeispiele x_i und Alphas
- das war die Haupthürde

Vorhersagequalität: wesentliche Verbesserung gegenüber (optimierten Variante von) BR

- Average Precision von 53% im Vergleich zu 38%
- Break-even point von 48% im Vergleich zu 35%
- \sim Hälfte der Labels an der Spitze der Rangliste relevant

Vergleich der Laufzeiten weniger klar

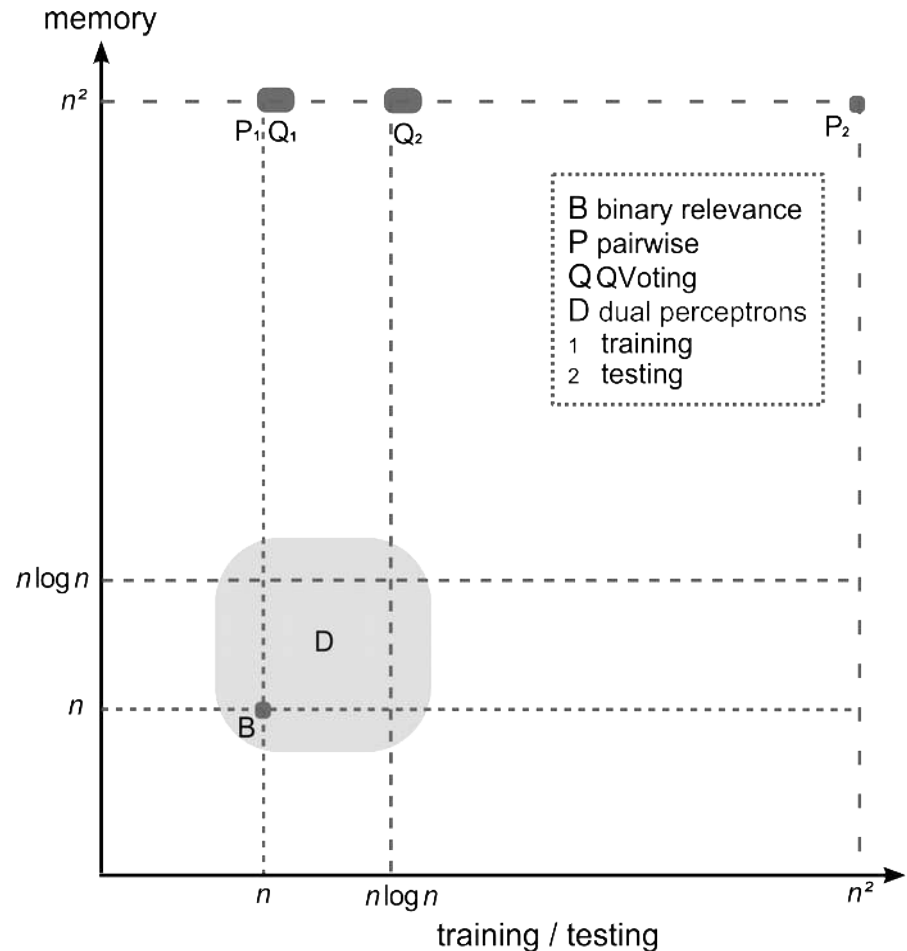
- Dual MLPP braucht weniger arithmetische Operationen für das Training
 - aber dennoch mehr CPU-Zeit
- BR ist klar schneller bei der Klassifizierung

Zusammenfassung

Dual-Form erlaubt
Verarbeitung von hoch-
dimensionalen Daten

- Flaschenhals ist nun hauptsächlich die Anzahl der Trainingsbeispiele
- Komplexität ist nun grob $O(m \cdot d \cdot n)$

EUROVOC trotzdem noch
sehr anspruchsvoll für
paarweises Lernen



Hohe Dimensionalität des Klassenraums



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Paarweises HOMER

(PL 2009)

HOMER: Hierarchy of Multilabel Classifiers

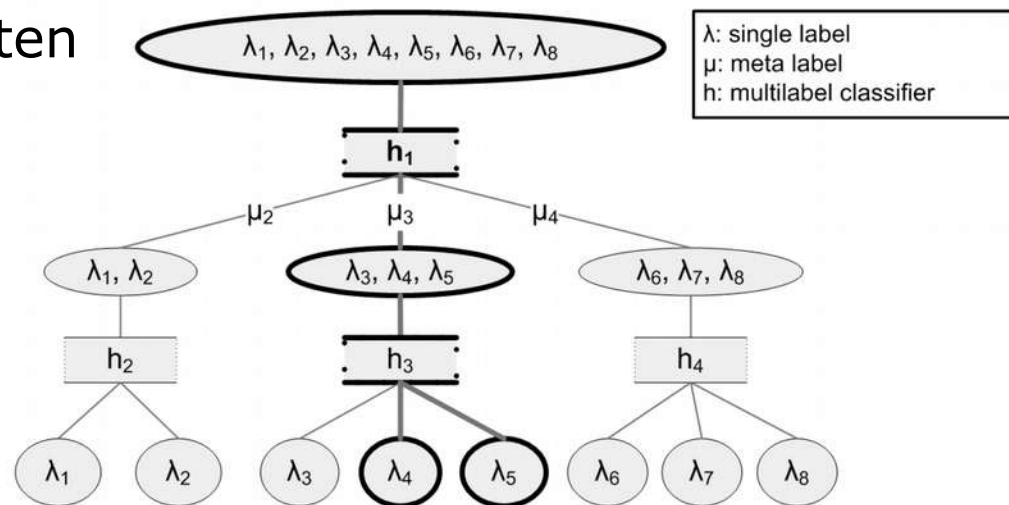


TECHNISCHE
UNIVERSITÄT
DARMSTADT

- bricht Originalproblem in Teilprobleme auf, die in einer Hierarchie organisiert sind
- k Labels werden zu einem Metalabel vereint, das wiederum ein mögliches Label im Elternproblem darstellt
- Labels werden durch balanced k -means Clustering vereint

Idee: Verwendung von paarweiser Zerlegung an den inneren Knoten

- weiter den Speicher-
verbrauch reduzieren
- reduziere auch Trainings-/
Testlaufzeiten
- aber hoffentlich beibehalten
der Vorhersagequalität



Komplexität

Verbesserung hängt von benutzerdefinierten Größe k der Teilprobleme ab

- normalerweise zwischen 5-10

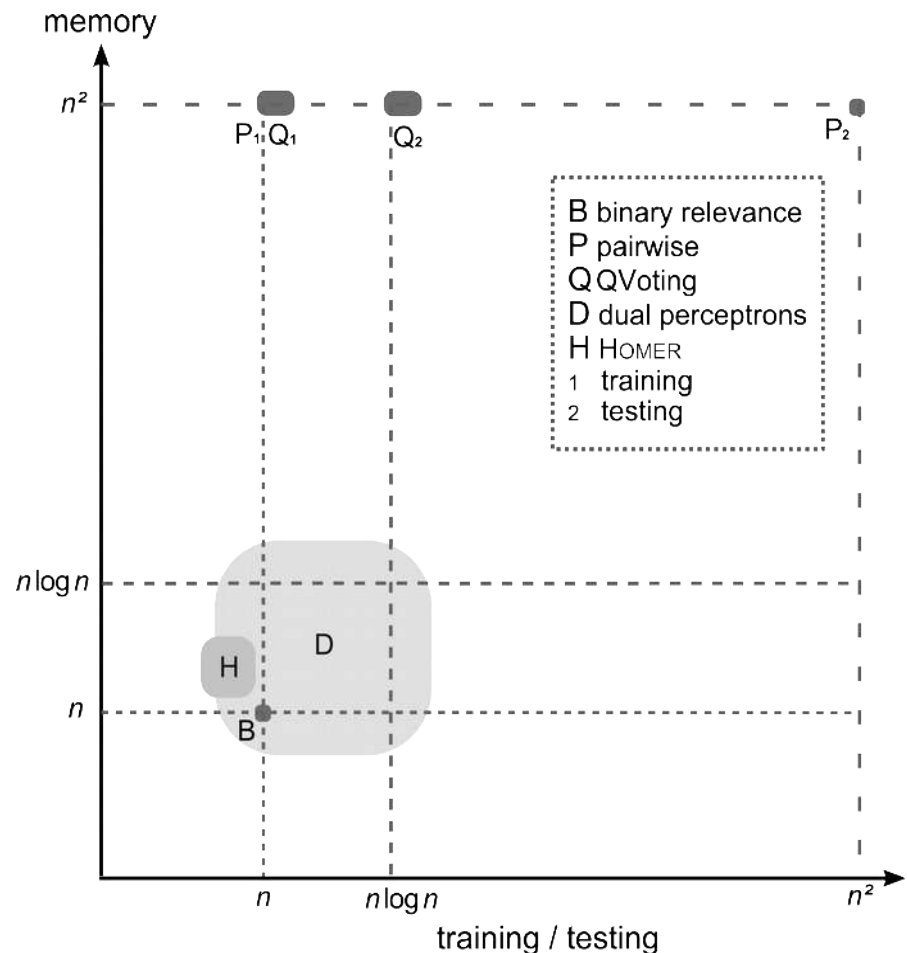
Speicher: von $O(n^2)$ auf $O(kn)$
Klassifizierer

Training: Reduktion um Faktor $O(1/n \log n)$ bei Anzahl der Trainingsbeispiele

- sub-linear
- sogar schneller als BR

Vorhersage: Reduktion um Faktor $O(k/n)$ bei Evaluationen von Basisklassifizierern

- weniger als BR für große n



Getestet auf vier Datensätzen mit 101 bis 632 Labels und 16000 bis 49000 Instanzen

- Formale Analyse der Laufzeiten bestätigt
- HOMER schneller als BR im Trainieren und fast beim Klassifizieren

HOMER mit paarweisem Lernen schlägt meistens alle anderen Kombinationen bezüglich Kombinationsmaß F1

- HOMER balanciert Recall und Precision sehr gut aus, insbesondere gegenüber konventionellem paarweisen Lernen

HOMER und paarweise Zerlegung harmonisieren sehr gut

- reduziert Laufzeit wesentlich
- behält Vorteile der paarweisen Zerlegung gegenüber BR bei
 - obwohl mehr Trainingsbeispiele in den Metalabels
→ schwierigere Teilprobleme in den inneren Knoten
 - obwohl viele paarweise Relationen nicht mehr berücksichtigt werden

HOMER erlaubt die paarweise Zerlegung für potentiell beliebig große Datensätze

- Abstand zu BR um einen benutzerdefinierten konstanten Faktor reduziert
- allerdings: Transformation ist nicht mehr äquivalent

Weitere Ansätze und Anwendungen

Multitask Learning bei *parallelen Datensätzen* (MLD 2010)

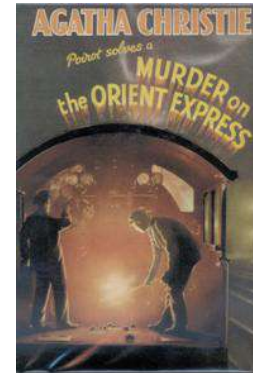
- Ausnutzung von Labelabhängigkeiten zwischen mehreren Klassendomänen
- betrachte alle parallelen Teilaufgaben als eine große Multilabel-Aufgabe

Syntaxanalyse von Texten (LWA 2010)

- ähnlich zu Multitask Learning
- betrachte alle Annotationen gleichzeitig statt unabhängig voneinander

Ausnutzung von Labelabhängigkeiten in Subgruppen von Datenpunkten (IDA 2012)

- Erweiterung der Fähigkeit der Ausnutzung von Labelabhängigkeiten



Genres:

Crime, Mystery, Thriller

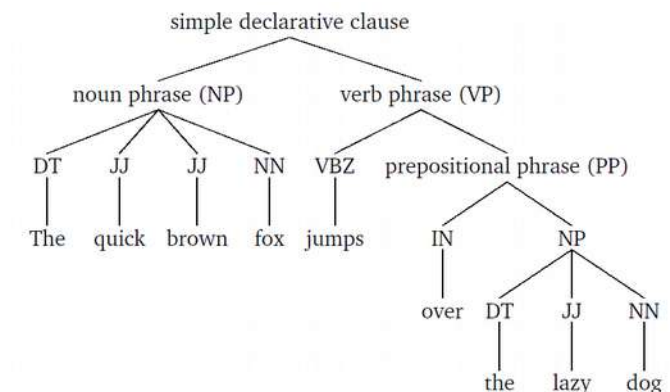
Subjects (LOC):

Private Investigators,
Orient Express, ...

Keywords:

mystery, fiction, crime,
murder, british, poiroot, ...

...





Zusammenfassung

Ausgangslage:

- paarweise Zerlegung gilt als akkurater als BR
- aber niedrige Effizienz und Skalierbarkeit

Vorgestellte Ansätze bewältigen wichtigste Hürde: die quadratische Abhängigkeit von der Anzahl der Labels

- Vorhersage (Quick Voting)
- Speicher (Dual MLPP)
- Training, Vorhersage, Speicher (Paarweises HOMER)

EUR-Lex: prototypisches Anwendungsszenario

- sehr anspruchsvoller Multilabel-Datensatz

Erforschung von Semantic Hashing Techniken

- Abbildung des Problems auf reduzierten Label-Raum
- vielversprechende Alternative zu HOMER
- Nächste Herausforderungen z.B.: ECML 2012 Discovery Challenge
 - 2,4 Millionen Wikipedia-Dokumente, 325.000 Kategorien

Erweiterung des paarweisen Frameworks

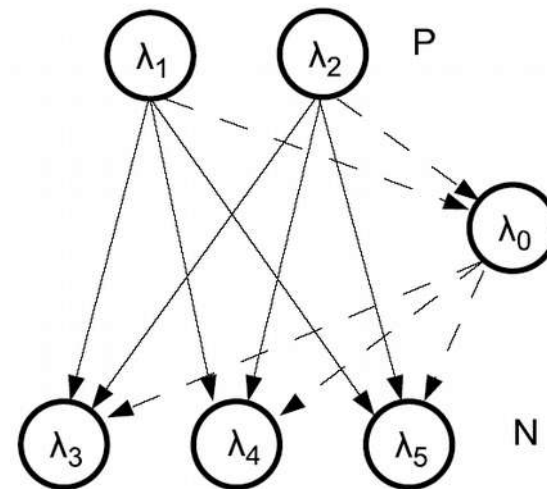
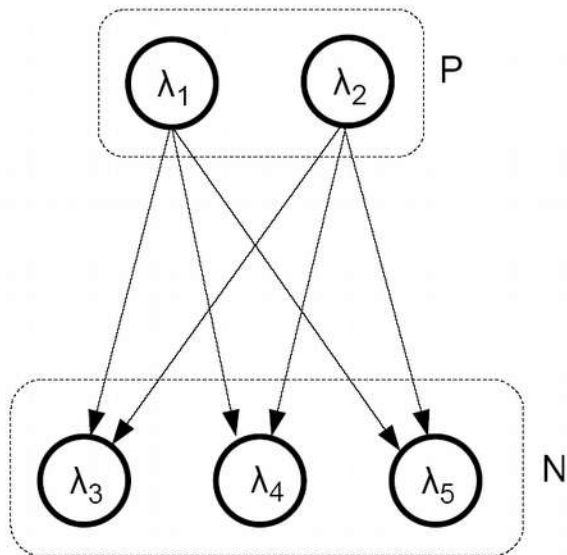
- Miteinbeziehung von Instanzen in Label-Schnittmengen
 - Berücksichtigung von hierarchischen Label-Strukturen
- Anpassung des Votings an unterschiedliche Anforderungen (Zielmaße)

Fragen?

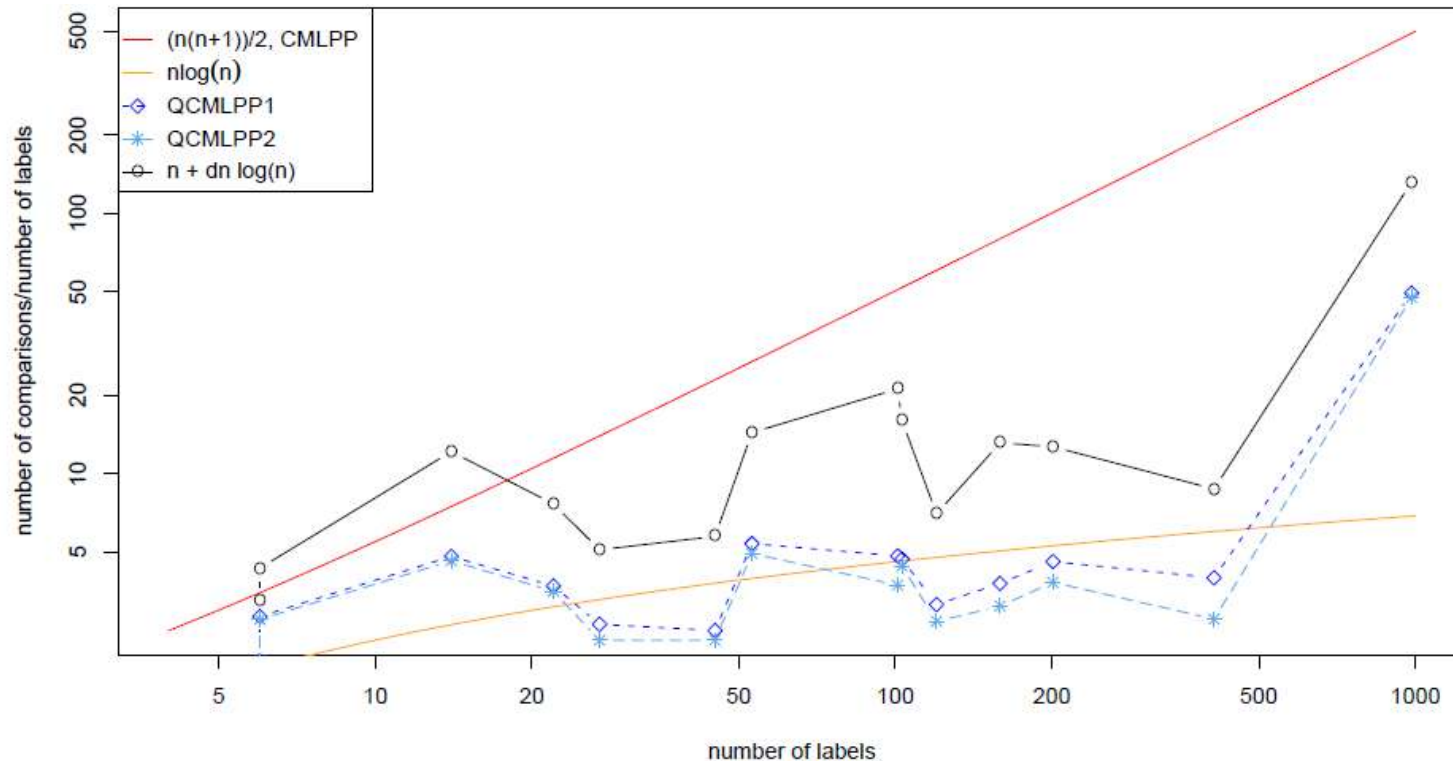
- (Park 2007) S.-H. Park, J. Fürnkranz: *Efficient Pairwise Classification*, Proceedings of the 18th European Conference on Machine Learning (ECML 2007)
- (MLJ 2008) J. Fürnkranz, E. Hüllermeier, E. Loza, K. Brinker: *Multilabel Classification via Calibrated Label Ranking*. Machine Learning, vol. 73 (2): pp. 133–153
- (IJCNN 2008) E. Loza, J. Fürnkranz: *Pairwise Learning of Multilabel Classifications with Perceptrons*, Proceedings of the 2008 IEEE International Joint Conference on Neural Networks
- (ECML 2008) E. Loza, J. Fürnkranz: *Efficient Pairwise Multilabel Classification for Large-Scale Problems in the Legal Domain*, Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD-2008)
- (Tsoumakas 2008) G. Tsoumakas, I. Katakis, I. Vlahavas: *Effective and Efficient Multilabel Classification in Domains with Large Number of Labels*. Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)
- (PL 2009) G. Tsoumakas, E. Loza, I. Katakis, S.-H. Park, J. Fürnkranz: *On the Combination of Two Decompositive Multi-Label Classification Methods*. Proceedings of the ECML PKDD 2009 Workshop on Preference Learning
- (NC 2010) E. Loza, S.-H. Park, J. Fürnkranz: *Efficient Voting Prediction for Pairwise Multilabel Classification*. In: Neurocomputing, vol. 73 (7-9): pp. 1164 - 1176
- (MLD 2010) E. Loza: *Multilabel Classification in Parallel Tasks*. Working Notes of the 2nd International Workshop on Learning from Multi-Label Data at ICML/COLT 2010
- (LWA 2010) E. Loza: *An Evaluation of Multilabel Classification for the Automatic Annotation of Texts*. Proceedings of the LWA 2010: Lernen, Wissen, Adaptivität, Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML 2010)
- (IDA 2012) W. Duivesteijn, E. Loza, J. Fürnkranz, A. Knobbe: *Multi-label LeGo – Enhancing Multi-label Classifiers with Local Patterns*. Eleventh International Symposium on Intelligent Data Analysis

Backup-Folien

Paarweiser Ansatz und Kalibrierung



Resultate QVoting: Effizienz



- reduziert Klassifizierung von quadratisch $O(n^2)$ zu log-linear $O(d n \log(n))$ in der Praxis

Resultate QVoting: Vorhersagequalität

- Perzeptron-Algorithmus als Basislerner
- paarweiser Ansatz durchweg besser als BR
- konkurrenzfähig zu SVMs
 - aber: SVMs oft nicht anwendbar

dataset	n	HAMLoss		PREC		REC		F1	
		BR	CMLPP	BR	CMLPP	BR	CMLPP	BR	CMLPP
<i>scene</i>	6	10.42	10.00	71.80	71.83	71.21	74.20	71.19	72.76
<i>emotions</i>	6	35.64	34.08	46.78	48.62	60.15	61.90	52.63	54.47
<i>yeast</i>	14	24.09	22.67	60.47	62.37	59.07	63.31	59.76	62.83
<i>tmc2007</i>	22	7.37	6.78	62.57	64.16	66.47	73.61	64.46	68.56
<i>genbase</i>	27	0.26	0.48	99.22	99.59	95.49	90.60	97.32	94.88
<i>medical</i>	45	1.51	1.51	71.72	76.02	75.84	66.75	73.72	71.08
<i>enron</i>	53	7.56	6.01	41.56	52.82	47.05	49.51	44.13	51.11
<i>mediamill</i>	101	4.52	4.16	42.28	56.66	10.05	19.70	16.24	29.23
<i>rcv1</i>	103	1.26	1.03	80.15	84.89	79.70	81.61	79.93	83.22
<i>r21578</i>	120	0.78	0.55	59.98	72.89	78.36	76.68	67.92	74.63
<i>bibtex</i>	159	1.57	1.35	46.53	57.97	36.30	34.84	40.78	43.53
<i>eurlex_sm</i>	201	0.76	0.54	63.39	77.88	74.11	71.57	68.32	74.59
<i>eurlex_dc</i>	410	0.26	0.17	56.26	79.21	70.54	61.98	62.58	69.54
<i>delicious</i>	983	5.58	3.48	11.88	19.77	29.59	26.51	16.95	22.65

Foundations: QVoting and perceptrons



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Multilabel Pairwise Perceptrons

(IJCNN 2008, MLJ 2008)

Multilabel Pairwise Perceptrons (MLPP)

perceptron algorithm: learns a separating hyperplane between positive and negative examples

- simple and fast:

classifying: compute scalar product of instance vector \bar{x} and hyperplane normal vector \bar{w} and predict class $\text{sgn}(\bar{x} \cdot \bar{w}) \in \{-1, 1\}$
training: add either \bar{x} or $-\bar{x}$ to normal vector only if training instance is misclassified

- good performance in text-classification (large and sparse feature space)
- on-line learning algorithm
 - efficient alternative to Support Vector Machines

Results: efficiency of perceptrons



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Reuters Corpus Volume 1 (rcv1)

- 535,987 training news articles
- 268,427 for testing
- 25,000 features
- 103 distinct labels
- ~3.24 labels per example

dataset	n	HAMLoss		PREC		REC		F1	
		BR	CMLPP	BR	CMLPP	BR	CMLPP	BR	CMLPP
<i>scene</i>	6	10.42	10.00	71.80	71.83	71.21	74.20	71.19	72.76
<i>emotions</i>	6	35.64	34.08	46.78	48.62	60.15	61.90	52.63	54.47
<i>yeast</i>	14	24.09	22.67	60.47	62.37	59.07	63.31	59.76	62.83
<i>tmc2007</i>	22	7.37	6.78	62.57	64.16	66.47	73.61	64.46	68.56
<i>genbase</i>	27	0.26	0.48	99.22	99.59	95.49	90.60	97.32	94.88
<i>medical</i>	45	1.51	1.51	71.72	76.02	75.84	66.75	73.72	71.08
<i>enron</i>	53	7.56	6.01	41.56	52.82	47.05	49.51	44.13	51.11
<i>mediamill</i>	101	4.52	4.16	42.28	56.66	10.05	19.70	16.24	29.23
<i>rcv1</i>	103	1.26	1.03	80.15	84.89	79.70	81.61	79.93	83.22
<i>r21578</i>	120	0.78	0.55	59.98	72.89	78.36	76.68	67.92	74.63
<i>bibtex</i>	159	1.57	1.35	46.53	57.97	36.30	34.84	40.78	43.53
<i>eurlex_sm</i>	201	0.76	0.54	63.39	77.88	74.11	71.57	68.32	74.59
<i>eurlex_dc</i>	410	0.26	0.17	56.26	79.21	70.54	61.98	62.58	69.54
<i>delicious</i>	983	5.58	3.48	11.88	19.77	29.59	26.51	16.95	22.65

BR throughput:

- 0.9 ms / training doc
- 1.3 ms / test doc

MLPP throughput:

- 1.3 ms / training doc
- 4.2 ms / test doc
- 13.5 ms without QVoting

- an efficient SVM only terminated on 8 of 14 datasets

Dual MLPP

- perceptron can be reformulated as linear combination of the (misclassified) training instances

classifying with \mathbf{w} : $h'(\mathbf{x}) := \mathbf{x} \cdot \mathbf{w}$

training rule for \mathbf{w} : $\alpha_i = (y_i - h_i(\mathbf{x}_i)) \quad \mathbf{w}_{i+1} = \mathbf{w}_i + \alpha_i \mathbf{x}_i$

dual form of \mathbf{w} : $\mathbf{w} = \sum_{i=1}^m \alpha_i \mathbf{x}_i$

dual classifying: $h'(\mathbf{x}) = \sum_{i=1}^m \alpha_i \cdot \mathbf{x}_i \cdot \mathbf{x}$

- we maintain factors α and training instances in memory instead of the \mathbf{w} 's

Dual MLPP: Komplexität

	training time	prediction time	memory requirement
MMP/ BR	$\mathcal{O}(na')$	$\mathcal{O}(na')$	$\mathcal{O}(na)$
MLPP	$\mathcal{O}(dna')$	$\mathcal{O}(n^2a')$	$\mathcal{O}(n^2a)$
QCMLPP	$\mathcal{O}(dna')$	$\sim na' + dn \log(n) a'$	$\mathcal{O}(n^2a)$
DMLPP	$\mathcal{O}(m(dn + a'))$	$\mathcal{O}(m(dn + a'))$	$\mathcal{O}(m(dn + a') + n^2)$

Dual MLPP

		1 epoch					2 epochs			5 epochs			10 epochs		
		FC	MLNB	BR	MMP	DCMLPP	BR	MMP	DCMLPP	BR	MMP	DCMLPP	BR	MMP	DCMLPP
subject matter	IsERR	99.58	99.47	65.99	55.70	51.38	58.78	51.96	44.07	53.42	42.77	38.23	50.19	40.22	36.34
	ONEERR	77.83	98.68	35.71	30.58	22.78	27.13	27.09	17.29	22.69	18.38	13.49	20.64	15.97	12.55
	RANKLOSS	12.89	8.885	17.38	2.303	1.064	13.89	2.520	0.911	11.58	2.091	0.796	9.752	1.85	0.762
	MARGIN	40.16	25.04	62.31	10.11	4.316	52.28	11.22	3.757	44.77	9.366	3.337	38.45	8.177	3.214
	AVGP	22.57	11.91	59.33	74.01	78.68	66.07	76.95	82.73	70.69	82.10	85.64	73.30	83.75	86.52
	F1 _p	19.61	1.88	54.79	64.61	70.07	60.79	68.54	74.56	65.64	74.33	78.18	68.45	76.15	79.34
directory code	IsERR	91.51	99.34	52.80	47.68	36.55	46.26	40.01	32.38	40.76	33.28	29.22	37.55	31.39	28.30
	ONEERR	90.13	99.04	44.40	40.85	28.22	37.38	32.99	24.42	31.48	25.79	21.41	28.1	23.9	20.65
	RANKLOSS	14.17	7.446	19.40	2.383	0.972	15.09	2.058	0.863	11.69	1.874	0.824	9.876	1.529	0.815
	MARGIN	68.33	34.44	96.43	14.18	5.626	77.32	12.18	5.045	61.48	10.95	4.831	52.94	8.947	4.785
	AVGP	18.98	6.714	57.10	68.70	77.89	63.68	74.90	80.87	68.75	79.84	82.87	71.61	81.30	83.38
	F1 _p	8.47	0.93	49.37	55.08	67.19	56.37	63.29	71.27	61.83	70.00	74.19	64.83	71.86	75.04
EUROVOC	IsERR	99.82	99.82	99.25	99.14	98.20	98.70	98.00	96.75	97.46	96.14		97.06	95.13	
	ONEERR	93.52	99.58	53.11	78.98	34.76	44.93	56.88	28.01	36.69	39.46		33.84	34.99	
	RANKLOSS	12.97	22.34	39.78	3.669	2.692	35.25	4.091	2.398	30.93	4.573		28.59	4.509	
	MARGIN	1357.10	1623.72	3218.12	562.81	426.28	3040.01	670.65	387.51	2846.47	757.01		2716.63	740.12	
	AVGP	5.504	1.060	25.55	27.04	46.79	30.71	38.42	52.72	35.95	47.65		38.31	50.71	
	F1 _p	5.646	0.427	28.67	24.76	43.07	33.64	35.16	48.04	38.61	44.24		40.81	47.07	

	1 epoch		2 epochs		10 epochs
	MMP	DCMLPP	MMP	DCMLPP	MMP
AVGP	27.04	46.79	38.42	52.72	50.71
F1 _p	24.76	43.07	35.16	48.04	47.07

Dual MLPP



TECHNISCHE
UNIVERSITÄT
DARMSTADT

	<i>subject matter</i>		<i>directory code</i>		<i>EUROVOC</i>	
	training	testing	training	testing	training	testing
BR	29.96 s 1,680 M op.	7.09 s 184 M op.	50.67 s 3,420 M op.	9.62 s 378 M op.	368.02 s 33,074 M op.	53.34 s 3,662 M op.
MMP	31.95 s 1,807 M op.	6.89 s 184 M op.	53.38 s 3,615 M op.	9.46 s 378 M op.	479.14 s 40,547 M op.	52.90 s 3,662 M op.
DMLPP	372.14 s 6,035 M op.	151.98 s 4,471 M op.	383.40 s 3,047 M op.	187.65 s 5,246 M op.	13,058.01 s 17,647 M op.	6,780.51 s 123,422 M op.
MLPP	69.50 s 3,886 M op.	164.04 s 18,427 M op.	120.70 s 4,735 M op.	643.34 s 77,629 M op.	– (175 G op.)	– (7 · 10 ¹² op.)
QCMLPP	86.83 s 5,566 M op.	35.21 761 M op.	159.39 s 8,155 M op.	78.3 s 1,053 M op.	– (209 G op.)	– (74 G op.)

dataset	BR/MMP	DMLPP	DCMLPP	MLPP
<i>subject matter</i>	153 MB	199 MB	210 MB	541 MB
<i>directory code</i>	167 MB	210 MB	229 MB	1,818 MB
<i>EUROVOC</i>	1,145 MB	1,242 MB	1,403 MB	(152 GB)

HOMER: Komplexität



TECHNISCHE
UNIVERSITÄT
DARMSTADT

(a) total number of binary base classifiers

		H+BR	H+CLR
		$\mathcal{O}(n)$	$\mathcal{O}(kn)$
BR	$\mathcal{O}(n)$	$\mathcal{O}(1)$	$\mathcal{O}(k)$
CLR	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$	$\mathcal{O}(k/n)$

(b) preferences learned per training example

		H+BR	H+CLR
		$\mathcal{O}(kd \log_k n)$	$\mathcal{O}(kd \log_k n)$
BR	$\mathcal{O}(n)$	$\mathcal{O}(\frac{d \log n}{n})$	$\mathcal{O}(\frac{d \log n}{n})$
CLR	$\mathcal{O}(dn)$	$\mathcal{O}(\frac{\log n}{n})$	$\mathcal{O}(\frac{\log n}{n})$

(c) binary base classifier evaluations for one prediction

		H+BR	H+QCLR
		$\mathcal{O}(kd \log_k n)$	$\mathcal{O}(kd \log n)$
BR	$\mathcal{O}(n)$	$\mathcal{O}(\frac{d \log n}{n})$	$\mathcal{O}(\frac{kd \log n}{n})$
QCLR	$\mathcal{O}(dn \log n)$	$\mathcal{O}(\frac{k}{n \log k})$	$\mathcal{O}(\frac{k}{n})$

HOMER: Experimente

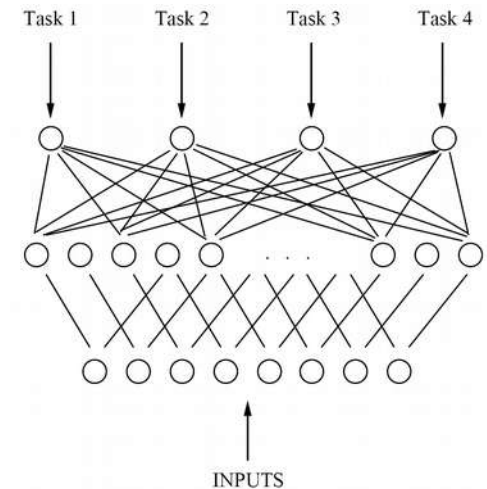


TECHNISCHE
UNIVERSITÄT
DARMSTADT

Method	<i>mediamill</i>	<i>jmlr2003</i>	<i>eccv2002</i>	<i>hifind</i>	
BR	58.00 %	32.27 %	36.58 %	59.43 %	micro Prec
QCLR	73.89 %	56.18 %	58.11 %	–	
H+BR	56.98 %	26.48 %	28.91 %	55.31 %	
H+QCLR	58.35 %	31.93 %	28.43 %	55.26 %	
BR	44.79 %	9.85 %	7.42 %	45.73 %	micro Rec
QCLR	43.86 %	4.57 %	3.84 %	–	
H+BR	44.91 %	10.81 %	13.21 %	48.64 %	
H+QCLR	48.77 %	10.28 %	15.07 %	54.06 %	
BR	50.55 %	15.09 %	12.34 %	51.65 %	micro F1
QCLR	55.04 %	8.45 %	7.21 %	–	
H+BR	50.23 %	15.36 %	18.14 %	51.76 %	
H+QCLR	53.13 %	15.55 %	19.70 %	54.65 %	
BR	2413.40	2801.17	2701.32	4179.66	training time in s
CLR	7423.19	6542.51	7460.14	–	
H+BR	1065.21	1101.61	1144.47	2345.39	
H+CLR	1667.29	1871.00	1836.34	3801.53	
BR	3.84	6.67	5.47	50.47	testing time in s
QCLR	103.59	119.28	154.65	–	
H+BR	4.35	7.70	4.48	48.77	
H+QCLR	4.90	9.26	5.62	60.02	

Motivation: Multi-task learning

- setting: multiple related tasks with shared feature representation
 - learn task simultaneously: find shared internal representation
 - use training signals of related tasks as an inductive bias to improve generalization
 - goal: to improve accuracy
 - especially when training data is rare
- it has been shown that this approach outperforms tackling tasks separately
- **parallel tasks:** multi-task, but with *same* input space and *same* instances in all tasks
- other links: multi-variate regression, multi-output learning etc.

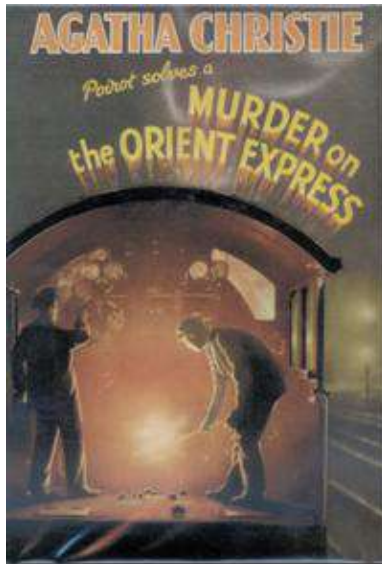


trained multi-task neural network
by Caruana, 1997

Parallel Tasks



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Summary: Returning from an important case in Syria, Hercule Poirot boards the Orient Express in Istanbul. The train is unusually crowded for the time of year. Poirot secures a berth only with ...

Text: It was five o'clock on a winter's morning in Syria. ... "Then," said Poirot, "having placed my solution before you, I have the honour to retire from the case."

Author:

Agatha Christie

Genres:

Crime, Mystery, Thriller

Subjects (LOC):

Private Investigators, Orient Express, ...

Keywords:

mystery, fiction, crime, murder, british, poirot, ...

Rate:

4 of 5 stars

Epoch:

1930ies

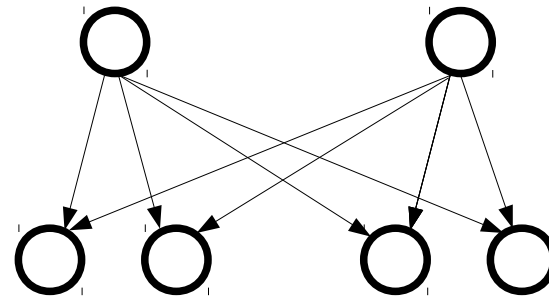
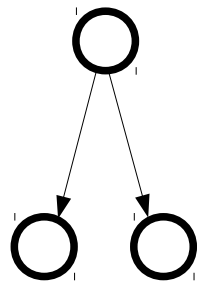
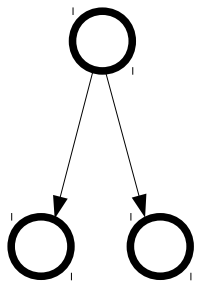
Country:

UK

...

Parallel Tasks

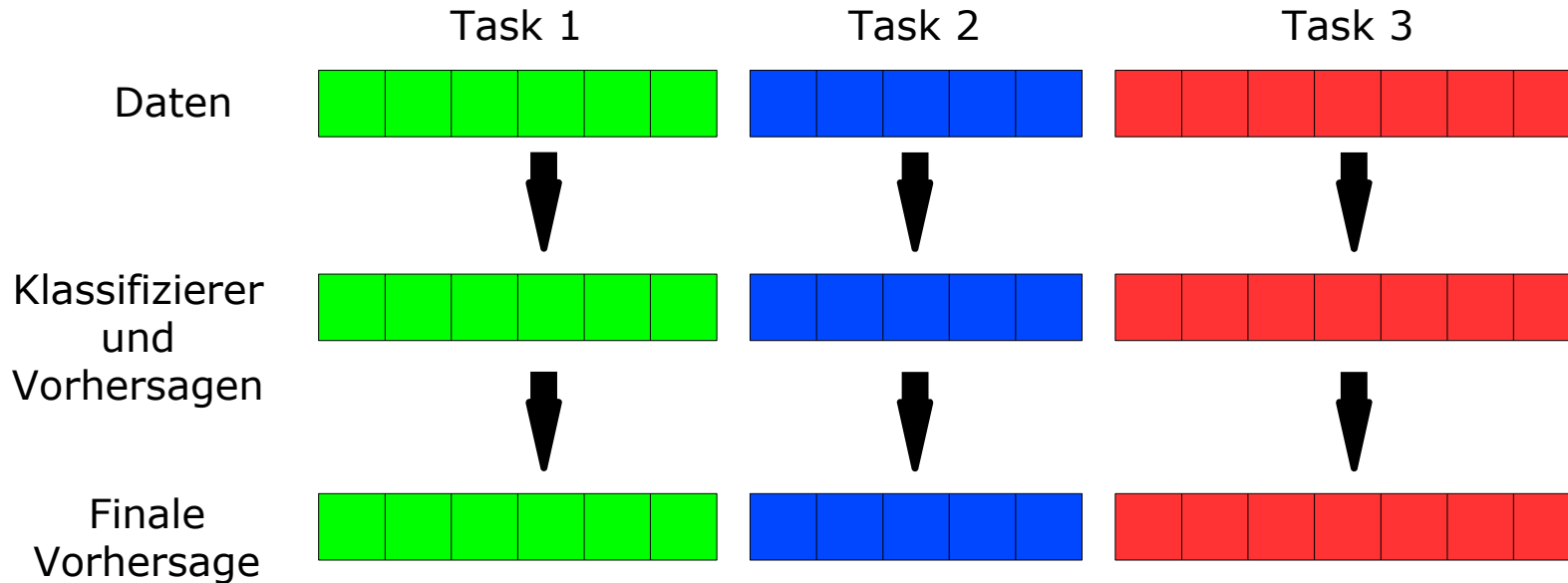
- Szenario: Mehrere Tasks teilen sich die selben Instanzen aber mit unterschiedlichen Labelzuordnungen
- Idee: statt sie unabhängig voneinander zu lernen, vereine Labels und lerne sie gemeinsam (paarweise)
 - profitiere von zusätzlichen Abhängigkeiten zwischen Klassen aus den unterschiedlichen Tasks



Parallel tasks learning: Unabhängige Tasks Ansatz



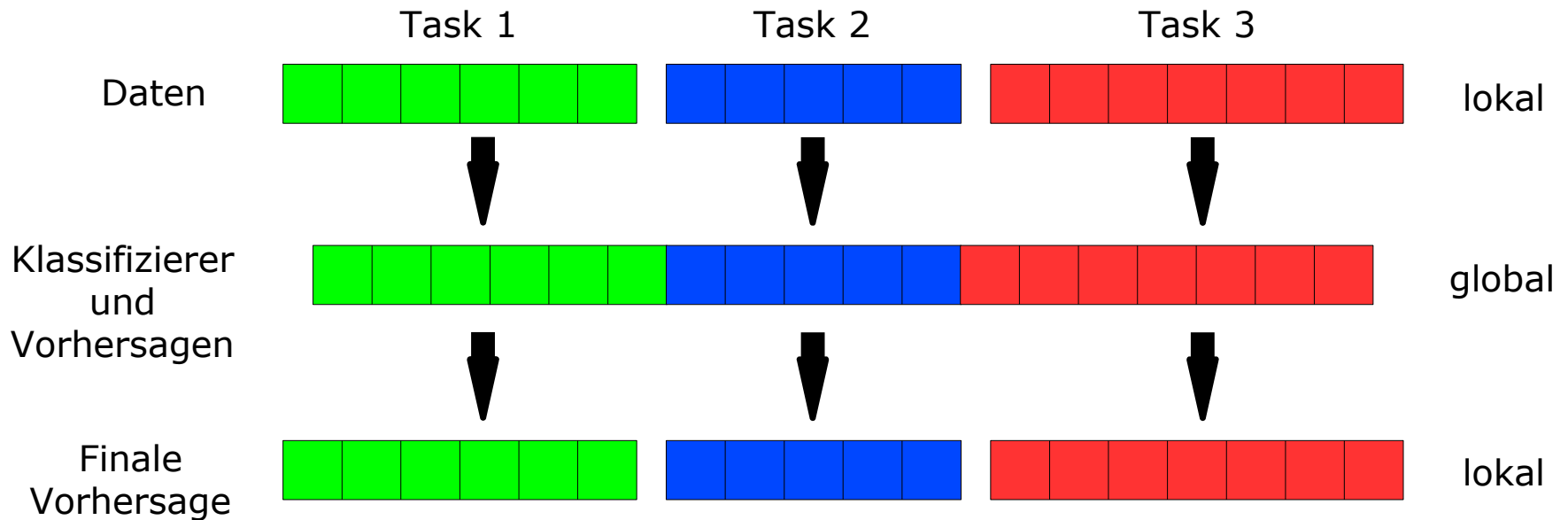
TECHNISCHE
UNIVERSITÄT
DARMSTADT



Vereinte Parallel Tasks Ansatz

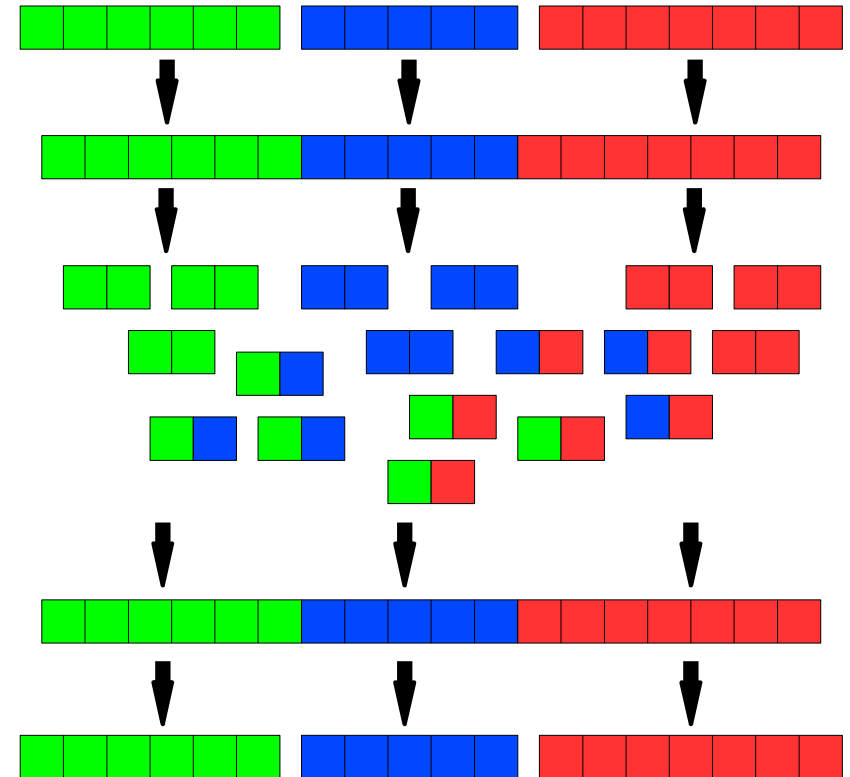
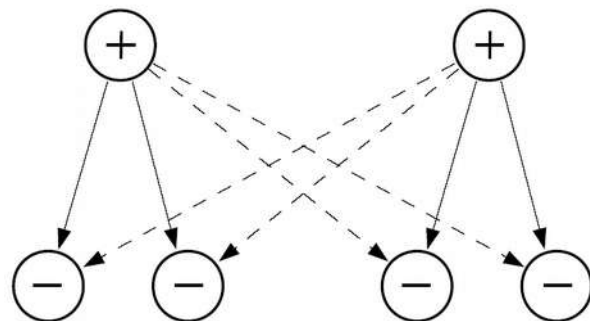
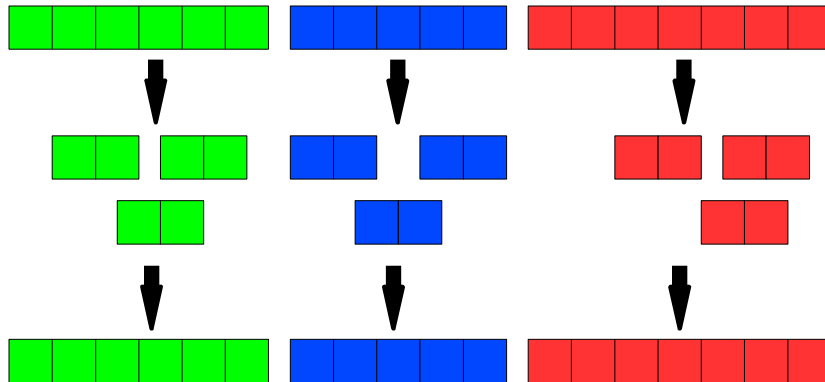


TECHNISCHE
UNIVERSITÄT
DARMSTADT



- Training: Aggregation von lokal zu global
- Vorhersage: Reduktion von global zurück zu lokal

Global Ansatz, paarweise



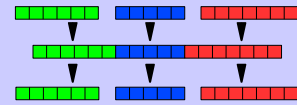
Experiments:

subset
accuracy / 0-1 loss

the lower
the better

F1 as if true number
of labels was returned

	REC	PREC	ACC	RANK	AVGP	F1(y)
<i>EURlex</i>	36.64	76.48	0.284	1.683	63.20	58.34
<i>sm</i>	64.50	75.48	33.29	0.874	83.26	75.25
PT	57.34	85.36	36.42	0.851	84.45	76.96
<i>dc</i>	54.23	77.11	45.95	0.844	81.05	71.42
PT	48.09	83.94	44.98	0.840	82.20	73.25
<i>ev</i>	25.48	66.63	0.636	2.325	53.35	48.59
PT	25.22	67.10	0.610	2.307	53.47	48.71
<i>mean</i>	48.07	73.07	26.63	1.348	72.55	65.09
PT	43.55	78.80	27.34	1.333	73.37	66.31
wins	0	3	1	3	3	3



(!) mean over
domains

- Ranking-Maße: globaler Ansatz besser (stat. signifikant) aber manchmal nur kleine Verbesserung
- kann als Test verwendet werden, um herauszufinden, inwiefern Label-Abhängigkeiten ausgenutzt werden