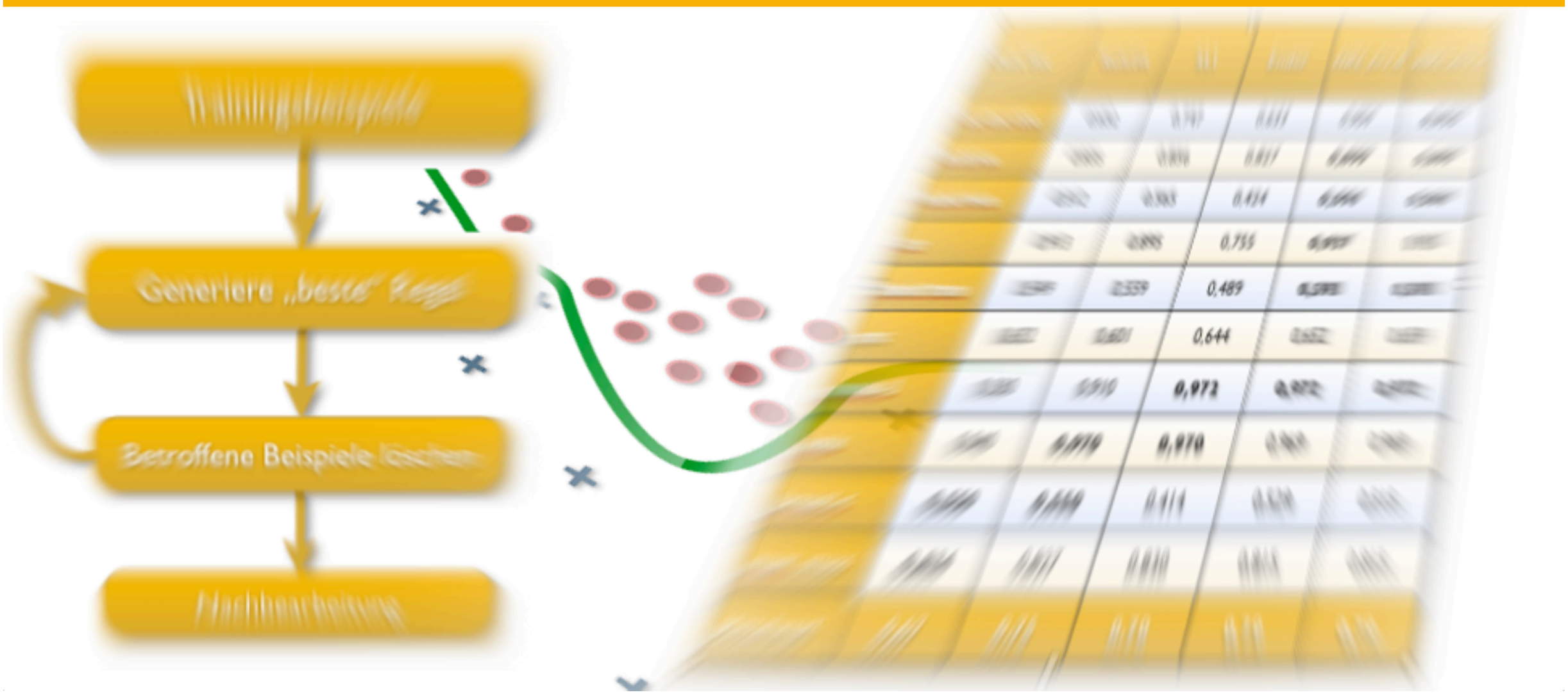
Statistical Methods for Rule Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

A Statistical Approach to Rule Learning
Efficient Rule Ensemble Learning using Hierarchical Kernels



Überblick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Einführung Regel-Lernen
  - Rule Ensemble Learning
  - Probleme
- A Statistical Approach to Rule Learning
  - Fehlerminimierung
  - Algorithmus
  - Ergebnisse
  - Fazit
- Efficient Rule Ensemble Learning using Hierarchical Kernels
  - Non-lineare Klassifikation
  - Ergebnisse
  - Probleme
- Fazit

Regel-Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Finde Regeln, die Vorraussagen für feste Variablen machen.
  - Gegeben: Trainingsbeispiele

Temperature	Outlook	Humidity	Windy	Play Golf?
mild	overcast	normal	false	true
mild	rain	high	true	false
hot	overcast	high	false	true
cool	rain	normal	false	false

- Gesucht: Regeln, die neue Beispiele klassifizieren
  - IF Outlook == rain => Play Golf == false
  - IF Outlook == overcast && Windy == false => Play Golf == true

Temperature	Outlook	Humidity	Windy	Play Golf?
hot	rain	high	true	false
mild	overcast	normal	false	true

Regel-Lernen
Separate-and-Conquer Strategie

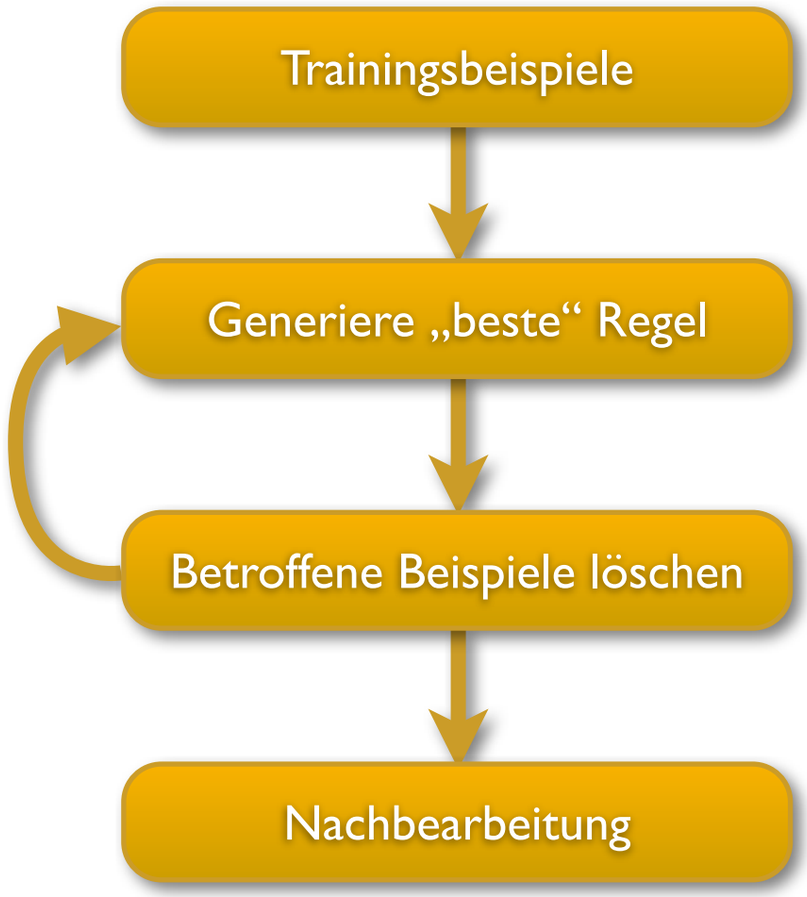
TECHNISCHE UNIVERSITÄT DARMSTADT

1. Einzelne Regel generieren, die Trainingsbeispiele am „besten“ erklärt

2. Von Regel abgedeckte Trainingsbeispiele entfernen

3. Keine Trainingsdaten mehr vorhanden? Fertig! Ansonsten zurück zu 1.

4. Nachbearbeitungsschritt (optional)



- Entscheidungsliste als Ergebnis
- Klassifizierung anhand einzelner Regel

Regel-Lernen
Rule Ensemble Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

- Erstelle Menge von gewichteten Regeln
  - Summiere Gewichte aller Regeln, deren Bedingungen zutreffen

  - IF Outlook == rain => -0.6
  - IF Temperature == mild && Windy == false => 0.3
  - IF Outlook == overcast && Humidity == high => -0.1

Temperature	Outlook	Humidity	Windy	Play Golf?
mild	rain	high	false	0.3-0.6 = -0.3 = false
mild	overcast	high	false	0.3-0.1 = 0.2 = true

Regel-Lernen
Probleme

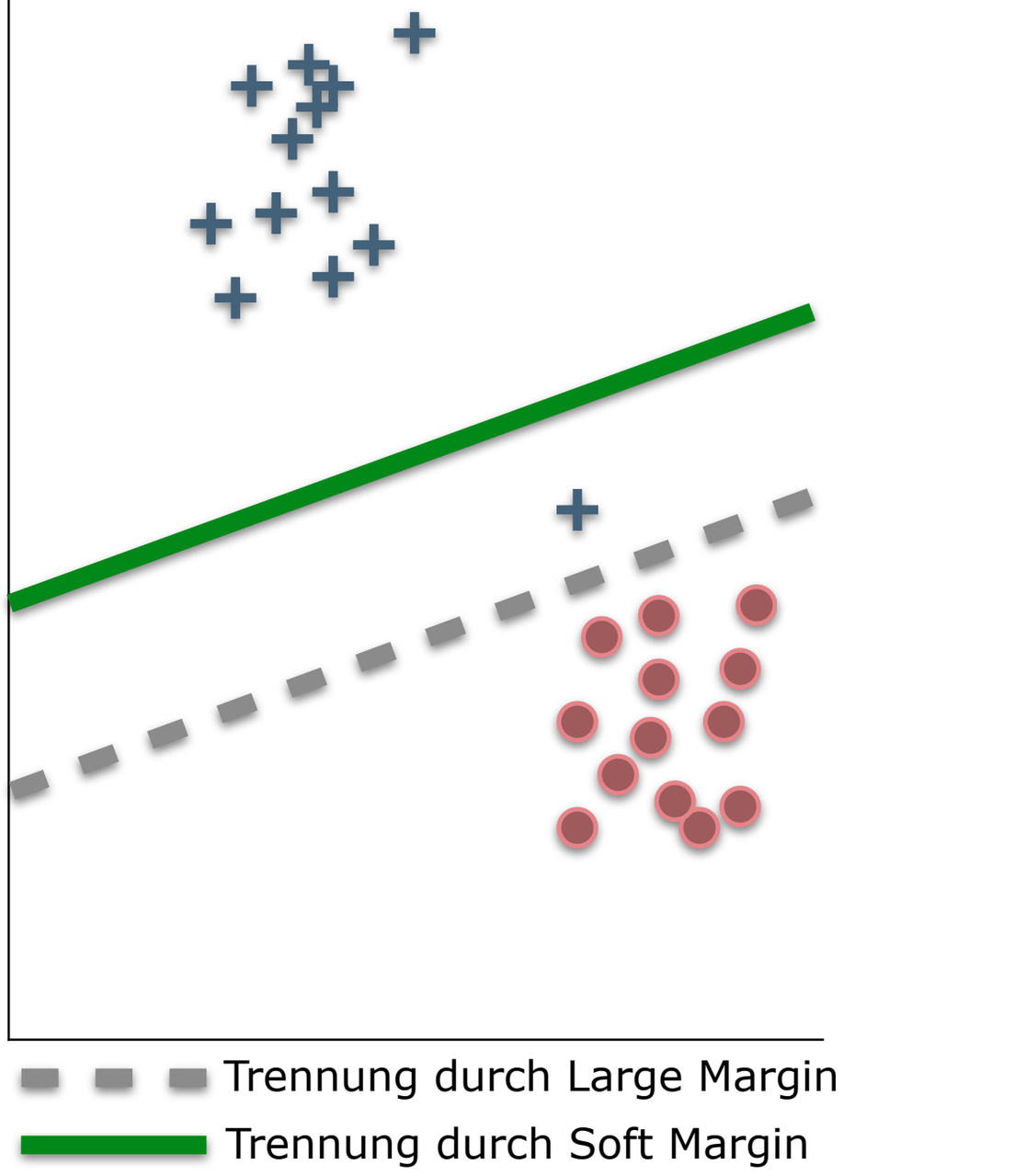
TECHNISCHE UNIVERSITÄT DARMSTADT

- Potenziell große Anzahl an möglichen Regeln
  - Steigt exponentiell mit Anzahl der Attribute
- Statistische Analyse traditioneller Regellerner schwierig
  - Verwendung von Heuristiken
- Overfitting
  - Überangepasste Regelmengen haben Probleme mit unbekannten Beispielen
  - Klassifikation abhängig von zufälligen Eigenschaften von Beispielen
  - Gelernte Regelmengen sollten möglichst
    - wenig Regeln beinhalten
    - aus Regeln mit wenigen Attributen bestehen
- Underfitting
  - Zu simple Regelmengen klassifizieren nicht korrekt
- Regelmengen müssen simpel, aber nicht zu simpel sein

A Statistical Approach to Rule Learning
Klassifizierung nach SVMs

TECHNISCHE UNIVERSITÄT DARMSTADT

- Lege Hyperebene als Trennfläche in den Raum
- Maximiere Abstand der nächsten Trainingsobjekte auf beiden Seiten (Large Margin)
- Falsche Klassifizierung von Ausreißern kann bessere Bestimmung von neuen Beispielen ermöglichen (Soft Margin)



A Statistical Approach to Rule Learning
Definitionen

TECHNISCHE UNIVERSITÄT DARMSTADT

$X := \{x_1, x_2, ..., x_m\}$ Menge an möglichen Beispielen

$Y := \{y_1, y_2, ..., y_m\}$ Menge an Labeln

$R = \{r_1, r_2, ...\}$ Menge an Regeln

i

$r_j : X \to \{-1, 1\}$ Regel weist Beispiel Label -1 oder 1 zu

$x_i(j)$ Anwendung der Regel j auf Beispiel x_i

$x_i := (x_i(1), x_i(2), ..., x_i(n))^T$ Darstellung x_i durch Vektor der Regelwerte

A Statistical Approach to Rule Learning
Definitionen

TECHNISCHE UNIVERSITÄT DARMSTADT

$$p \in [-1, 1]^n$$

Gewichtsvektor

$$\sum_{j=1}^{n} p_j x_i(j)$$

Summierung der Gewichte. Positive Klasse für Ergebnis >= 0, sonst negativ

$$\hat{\varepsilon} := \frac{1}{m} \sum_{i=1}^{m} l(p^T x_i, y_i)$$

Empirischer Fehler

Loss Funktion: Fehler für Klassifikation.

A Statistical Approach to Rule Learning
Fehlerminimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gewichtsvektor handlich, wenn Vielzahl Komponenten gleich null
- Gesucht: Gewichtsvektor p, der empirischen Fehler minimiert
  - NP-Schweres Kombinatorisches Problem
  - Minimiere empirischen Fehler nicht direkt, sondern über verwandte Größen

$$\hat{\varepsilon} := \frac{1}{m} \sum_{i=1}^{m} l(p^T x_i, y_i)$$

A Statistical Approach to Rule Learning
Fehlerminimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- 1. Ansatz: 1-norm support vector machine (soft margin)

$$\underset{p}{\text{minimize}} \sum_{i=1}^{m} \left[1 - p^T x_i\right]_+$$

$$||p||_1 = \sum_{j=1}^{n} |p_j| = s$$

- s = Frei wählbare Variable
  - Zur Anpassung des Algorithmus
- Soft margin erlaubt falsche Klassifikation von Ausreißern
- Betragssummennorm sorgt für viele Gewichte gleich null
  - Wenige, relevante Regeln ungleich null

A Statistical Approach to Rule Learning
Fehlerminimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- 2. Ansatz: Empirical Margin

$$\mu_p(x, y) := p^T x \cdot y$$

$$\underset{p}{\text{maximize}}\ \hat{\mu}_p := \frac{1}{m} \sum_{i=1}^{m} \mu_p(x_i, y_i)$$

$$||p||_2 = \sqrt{\sum_{j=1}^{n} |p_j|^2} = 1$$

$$\hat{\varepsilon}_p \leq 1 - \hat{\mu}_p$$

Margin/Abstand: Abstand zu allen Trainingsbeispielen. Positiv für korrekte Klassifikation

Empirical Margin

Euklidische Norm: Komponenten von p sind größtenteils ungleich null

Schlecht, aber Vorteile bei der Berechnung

Empirischer Fehler wird durch Empirical Margin beschränkt

A Statistical Approach to Rule Learning
Fehlerminimierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- 3. Ansatz: Margin minus variance (MMV)

$$\hat{\sigma}_p := \frac{1}{m-1} \sum_{i=1}^{m} (\mu_p(x_i, y_i) - \hat{\mu}_p)^2$$

Empirische Varianz

$$\hat{\gamma}_p := \underset{p}{\text{maximize}}\ \hat{\mu}_p - \hat{\sigma}_p$$

$$||p||_1 = \sum_{j=1}^{n} |p_j| = 1$$

Wähle p so, dass Empiricial Margin maximiert, aber Empirische Varianz minimal ist

Betragssummennorm sorgt für wenige, relevante Regeln mit Gewicht ungleich null

Effizient zu berechnen

A Statistical Approach to Rule Learning
Algorithmus

TECHNISCHE UNIVERSITÄT DARMSTADT

Algorithm 1 Learn(X)

$R^{(0)} \leftarrow \emptyset$
for $i = 1, 2, \ldots$ and while new rules available **do**
  $R^{(i)} \leftarrow$ add new rule to $R^{(i-1)}$
  $T^{(i)} \leftarrow$ apply rules in $R^{(i)}$ to instances in X
  $p^{(i)} \leftarrow \text{argmax}\ \hat{\gamma}_p(T^{(i)})$
  $\gamma^{(i)} \leftarrow$ bound calculated from $p^{(i)}$ and $|R^{(i)}|$
end for
return $(R^{(i)}, p^{(i)})$ with the maximal $\gamma^{(i)}$

$$\gamma_p \geq \hat{\gamma}_p - \sqrt{\frac{18}{m}(2\ln 2n + \ln\frac{1}{\sigma})}$$

- Start mit leerer Regelmenge
- Solange neue Regeln verfügbar sind
- Füge einzelne Regel hinzu
- Wende Regel an
- Bestimme p (1-norm SVM, Emp. Margin, MMV)
- Berechne Obergrenze für zugehörigen echten Wert
- Ergebnis: Regeln und Gewichte mit bester Obergrenze

A Statistical Approach to Rule Learning
Ergebnisse

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Set	PART	SLIPPER	1-norm SVM	Empirical Margin	MMV
australian	84,3	85,1	**85,5**	80,9	**85,5**
breastcancer	69,6	**71,4**	70,3	52,8	69,2
breastw	94,9	96,0	90,8	**96,6**	94,0
colic	84,4	**84,6**	81,5	80,7	82,9
diabetes	74,0	73,5	74,2	65,0	**75,8**
german	70,0	71,8	70,0	61,1	**74,4**
glassg2	80,0	**83,1**	61,3	68,1	80,4
heartc	78,5	79,3	75,9	79,5	**85,1**
hearth	80,5	79,2	78,6	78,9	**81,6**
heartstatlog	78,9	78,5	75,2	78,5	**83,3**
hepatitis	80,2	79,5	79,4	59,3	**84,5**
ionosphere	90,6	**93,2**	78,6	78,1	86,9
krvskp	**99,3**	59,4	68,3	77,7	91,2
labor	77,3	**92,3**	75,4	84,2	80,7
mushroom	**100,0**	**100,0**	88,7	97,3	97,8
sick	**98,6**	98,5	93,9	66,0	93,9
sonar	76,5	74,2	56,7	69,2	**77,4**
vote	**95,9**	94,1	95,6	89,2	95,6
Mittelwert	84,08	82,98	77,77	75,73	84,46

Korrekte Klassifikationen in Prozent

Data Set	PART	SLIPPER	MMV
australian	32	**9**	15
breastcancer	20	**2**	20
breastw	10	15	**8**
colic	9	**2**	15
diabetes	**13**	21	17
german	78	24	**17**
glassg2	**7**	30	17
heartc	25	**19**	22
hearth	10	**8**	10
heartstatlog	**24**	30	25
hepatitis	8	**3**	10
ionosphere	**10**	19	**10**
krvskp	23	35	**6**
labor	**3**	**3**	12
mushroom	13	10	**5**
sick	20	22	**1**
sonar	**8**	26	22
vote	7	4	**2**
Wenigste Regeln	6	6	7

Anzahl Regeln

- MMV ähnlich akkurat wie PART und SLIPPER
- MMV, PART und SLIPPER ähnlich oft am kompaktesten, aber MMV ohne Ausreißer

A Statistical Approach to Rule Learning
Fazit

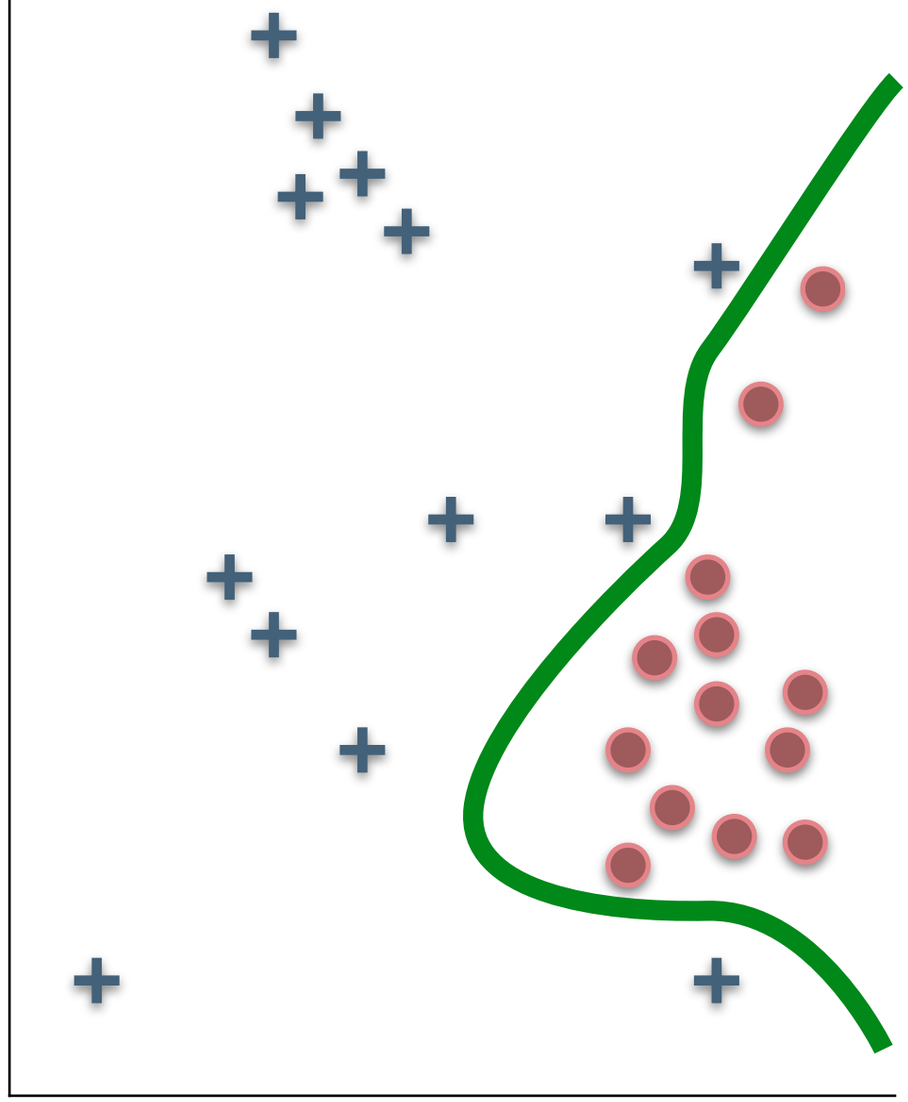
TECHNISCHE UNIVERSITÄT DARMSTADT

- Effiziente Lösung der MMV Optimierung
  - Zeit Komplexität linear gemäß Anzahl der Beispiele
- Erzeugt kompakte und aussagekräftige Regelmengen
- Autoren liefern keine Methode zum Finden von Regeln
  - Problem ist Datensatz spezifisch
- Wenig Aussagen zur Performance
  - Laut Autoren sind Probleme effizient zu lösen, genauere Informationen wären hilfreich
    - O-Notation
    - Vergleich Laufzeit/Speicherbedarf zwischen Beispielimplementation und PART/SLIPPER
- Ausgelassene Beweise und fehlerhafte sowie oberflächliche Auswertung

Efficient REL using Hierarchical Kernels
Kernel Methods

TECHNISCHE UNIVERSITÄT DARMSTADT

- Idee: Non-lineare Klassifizierung durch Anwendung des Kernel Tricks
  - Lege Beispieldaten in einen hochdimensionalen Raum
  - Jede Koordinate steht für eine Eigenschaft der Beispieldaten
  - Kernel Funktionen ermöglichen Operationen im Raum, ohne Koordinaten explizit berechnen zu müssen
  - Zurück im Ursprungsraum ist Trennung nicht mehr linear



Efficient REL using Hierarchical Kernels

Definitionen

TECHNISCHE UNIVERSITÄT DARMSTADT

$$\hat{\varepsilon} := \frac{1}{m} \sum_{i=1}^{m} l(p^T x_i, y_i)$$

Empirischer Fehler

$x_i(j)$

$$k_v := k_v(x_i, x_j) = \langle x_i(v), x_j(v)) \rangle$$

Regelanwendung muss nun keinen Boolean Wert mehr liefern sondern Vektor gegeben durch Kernel

Efficient REL using Hierarchical Kernels

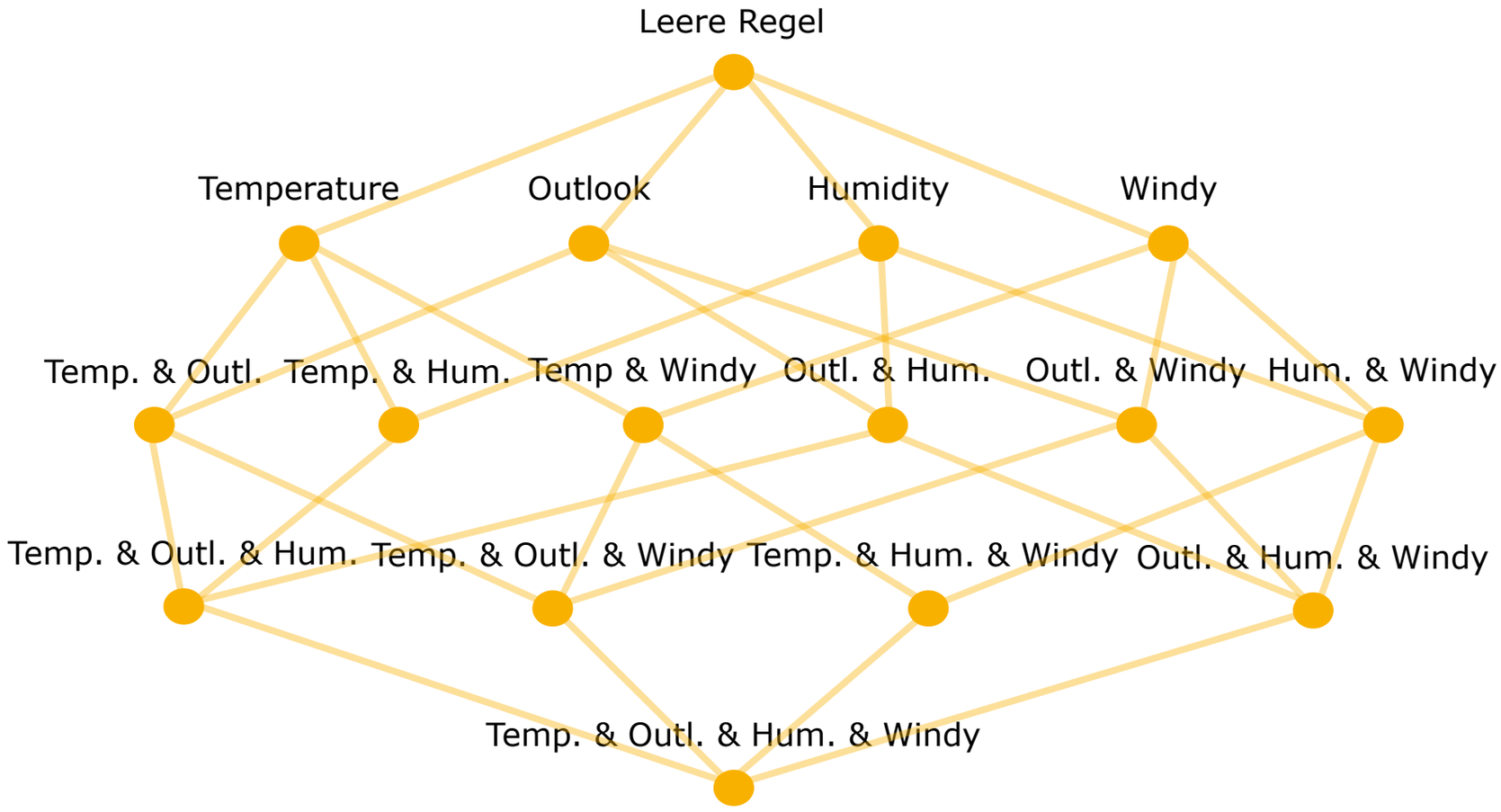
Hierarchical Kernels

TECHNISCHE UNIVERSITÄT DARMSTADT

$D(v)$ Menge der Nachfolger

$p_{D(v)}$ Gewichte der Nachfolgermenge

- Statt Betragssummennorm oder Euklidischer Norm jetzt z-norm (1, 2]



$$\underset{p,\xi}{\text{minimize}} \ \frac{1}{2}(\sum_{i=1}^{m} d_i ||p_{D(i)}||_z)^2 + C\sum_{j=1}^{m}\xi_j$$

$$y_j(\sum_{i=1}^{m}\langle p^T x_j\rangle) \geq 1-\xi_j, \xi_j \geq 0$$

- Normen zwischen 1 und 2 führen zu Gewichten mit vielen Nullen
- Paper liefert Algorithmus zum Lösen

Sehr Kompliziert!

Efficient REL using Hierarchical Kernels
Ergebnisse

TECHNISCHE UNIVERSITÄT DARMSTADT

Data Set	RuleFit	SLI	Ender	HKL z=1.5	HKL z=1.1
tic-tac-toe	0,652	0,747	0,633	0,904	**0,935**
balance	0,835	0,856	0,827	**0,899**	**0,899**
haberman	0,512	0,565	0,424	**0,594**	**0,594**
car	0,913	0,895	0,755	**0,937**	0,935
blood trans	0,549	0,559	0,489	**0,593**	**0,593**
cmc	0,632	0,601	0,644	0,652	0,659
monk-3	0,330	0,910	**0,972**	**0,972**	**0,972**
vote	0,940	**0,970**	**0,970**	0,969	0,969
breast-c	**0,550**	**0,550**	0,414	0,529	0,515
mam. mass	**0,834**	0,827	0,830	0,815	0,815
Mittelwert	0,67	0,75	0,70	0,79	0,79

RuleFit F-Score

Data Set	RuleFit	SLI	Ender	HKL z=1.5	HKL z=1.1
tic-tac-toe	**40**	59	111	111	79
balance	**17**	25	64	64	28
haberman	**6**	8	18	18	12
car	**34**	141	80	80	50
blood trans	18	**6**	58	58	7
cmc	39	**13**	74	74	43
monk-3	8	20	7	7	**2**
vote	12	**3**	8	8	**3**
breast-c	10	**7**	18	18	14
mam. mass	8	**4**	12	12	5
Min. Regeln	4	5	0	0	2

Anzahl Regeln

Data Set	RuleFit	SLI	Ender	HKL z=1.5	HKL z=1.1
Durchschnittliche Regellänge	2,31	2,12	2,06	1,66	**1,57**

Efficient REL using Hierarchical Kernels
Fazit

TECHNISCHE UNIVERSITÄT DARMSTADT

- HKL ermöglicht Berechnung in polynomieller Zeit
- Sehr akkurate Ergebnisse
- Einfache Regeln
  - Aber verhältnismäßig viele

Quellen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ulrich Rückert and Stefan Kramer. **A Statistical Approach to Rule Learning**, in Proc. 23th International Conference on Machine Learning, 2006
- Praktik Jawanpuria, J. Saketha Nath and Ganesh Ramakrishnan. **Efficient Rule Ensemble Learning using Hierarchical Kernels**, in Proc. 28th International Conference on Machine Learning, 2011

TECHNISCHE UNIVERSITÄT DARMSTADT

Danke für die Aufmerksamkeit!