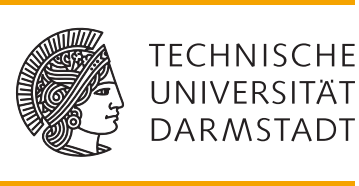
Vergleich von SVM und Regel- und Entscheidungsbaum-Lernern

Chahine Abid

Bachelor Arbeit



Betreuer: Prof. Johannes Fürnkranz

Frederik Janssen

Inhaltsübersicht

TECHNISCHE UNIVERSITÄT DARMSTADT

Einführung
Maschinelles Lernen

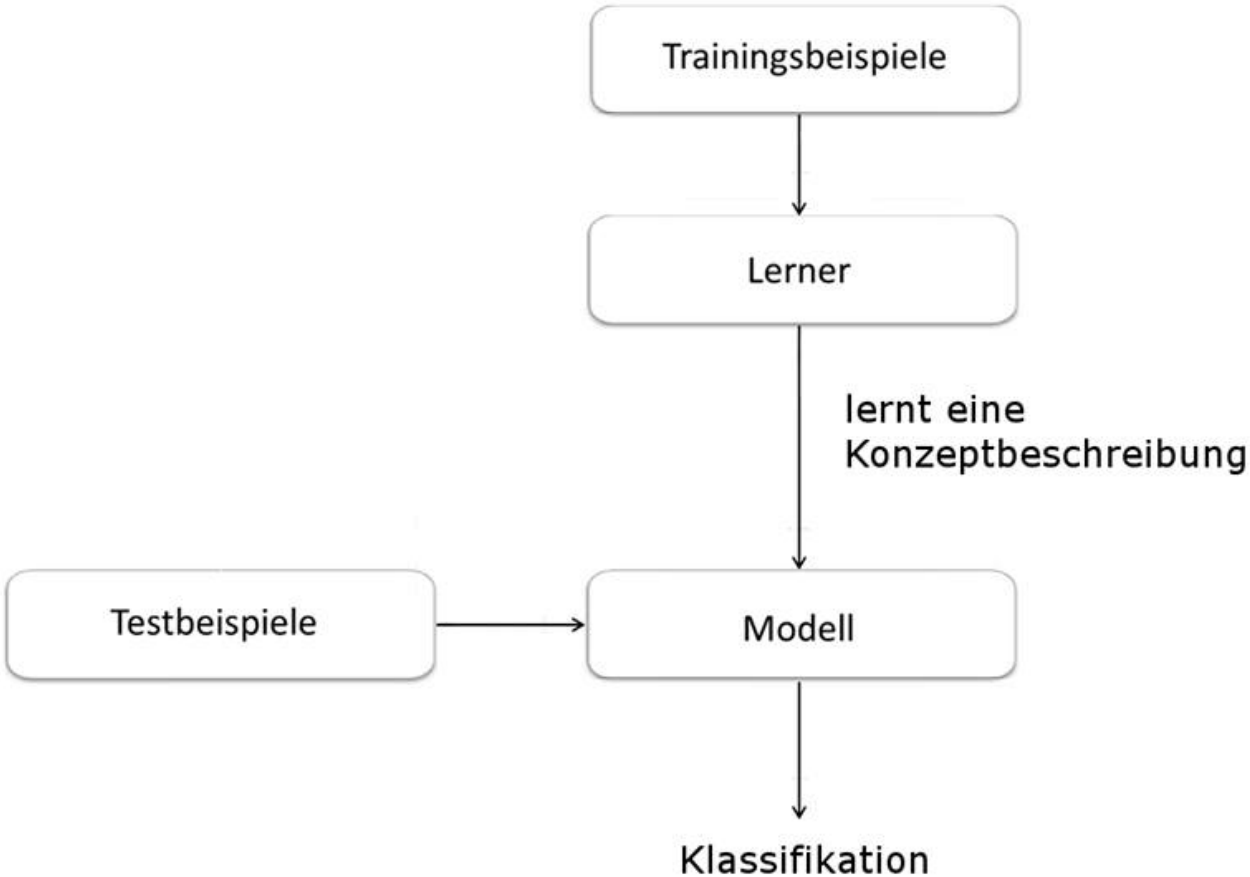
TECHNISCHE UNIVERSITÄT DARMSTADT

- Maschinelles Lernen:
  - aus Erfahrung lernen
  - die Leistung wird durch die Erfahrung verbessert
- Anwendungsgebiete:
  - Texterkennung
  - Klassifikation von DNA-Sequenzen
  - Bioinformatik . . .
- Teilgebiete:
  - Klassifikation
  - Regression
  - Clustering . . .

Einführung
Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Klassifikation



Einführung
Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Instanz : ein Beispiel aus dem Datensatz
- Attribute : die Eigenschaften der Instanz
  - nominales Attribut: besteht aus einer endlichen Menge von Werten
    Beispiel: die Farbe (rot, grün, blau, gelb)
  - numerisches Attribut: besteht aus Zahlen
    Beispiel: die Temperatur (-5, 0, 15, 27, . . .)
- Klasse: ordnet dem Beispiel die entsprechende Kategorie zu

Wetter	Temperatur	Feuchtigkeit	Wind	Spielen ?
sonnig	25	90	nein	ja
regnerisch	6	80	ja	nein

Einführung
Maschinelles Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- 10-Fache Kreuzvalidierung:
  - 9 Teilmengen: für das Training
  - 1 Teilmenge: für das Testen

$$Genauigkeit = \frac{Anzahl\ der\ korrekten\ klassifizierten\ Beispiele}{Anzahl\ der\ gesamten\ klassifizierten\ Beispiele}$$

- Überanpassung (Overfitting):
  - das Modell ist sehr an die Trainingsdaten überangepasst
  - schlechte Verallgemeinerung für die Testdaten
- Pruning: Methode zur Vermeidung von Überanpassung
  - prepruning vs postpruning

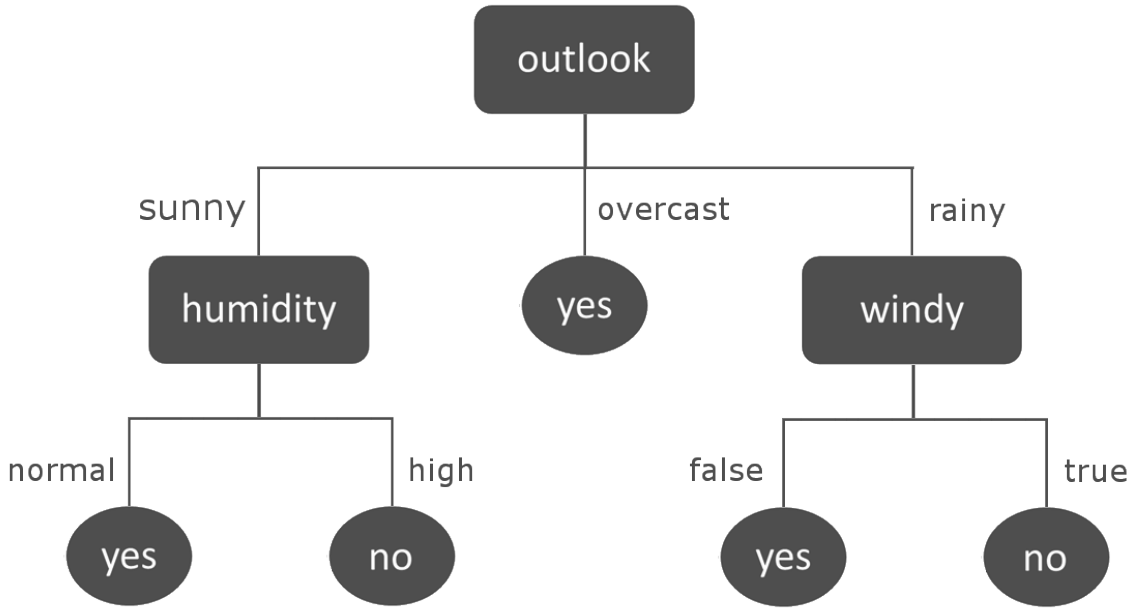
Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Support-Vector Maschinen:
  - die Parameter werden oft an die Daten angepasst
  - gute Performanz
- Entscheidungsbaum-Lerner und Regel-Lerner:
  - oft in ihrer Standard-Konfiguration verwendet
  - schlechtere Performanz als die Support-Vector Maschinen
- Ziele:
  - die Parameter für beide Systeme gründlich optimieren
  - die Ergebnisse vergleichen und analysieren

Maschinelle Lernverfahren (1)
Entscheidungsbaum-Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT



- Tests in den inneren Knoten
- die Klassen werden in den Blättern dargestellt

Maschinelle Lernverfahren (1)
Entscheidungsbaum-Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- ID3 Algorithmus:
  - einfache Entscheidungsbäume
  - möglichst reine Konten erzeugen
  - Information-Gain für jedes Attribut berechnen
- C4.5 Algorithmus:
  - Erweiterung von ID3
  - Umgang mit Daten der realen Welt
  - Pruning

Maschinelle Lernverfahren (2)
Regel-Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

Wetter	Temperatur	Feuchtigkeit	Wind	Spielen ?
sonnig	25	90	ja	ja
regnerisch	6	80	ja	nein
sonnig	21	90	nein	nein
sonnig	27	80	nein	ja

- die positive Beispiele werden mit der folgenden Regel abgedeckt:
  if (Wetter = sonnig) and (Temperatur $\geq$ 25) then Spielen = ja

Maschinelle Lernverfahren (2)
Regel-Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- RIPPER (Repeated incremental pruning to produce error reduction):
  - Modifikation des IREP-Algorithmus
  - Trainingsmenge aufteilen: Growingmenge und Pruningmenge
  - Prepruning und Postpruning verwenden: Regeln lernen und gleich prunen
  - Abbruchkriterium: basierend auf dem MDL-Prinzip
  - Optimierungsphase

Maschinelle Lernverfahren (3)
Ensemble Lernen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Kombination mehrerer Modelle

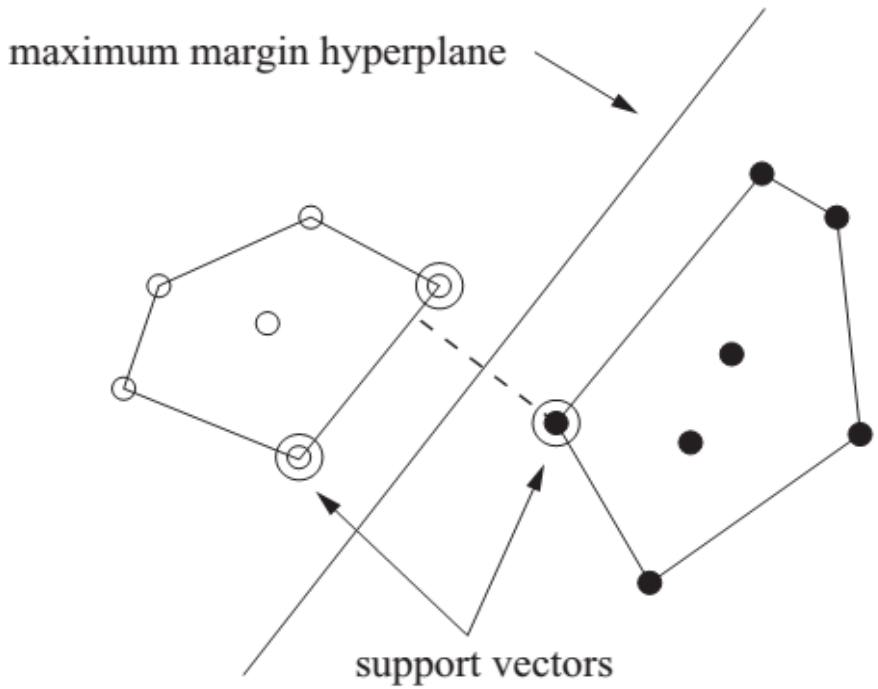
Beispiel: Random Forest

- Bagging:
  - die Trainingsmenge in mehreren Teilmengen aufteilen
  - eine Menge von Entscheidungsbäumen erzeugen (ein Baum für jede Teilmenge)
  - Kombination der Vorhersagen dieser Modelle
- zufällige Featureauswahl: Nur ein Teil der Attribute wird berücksichtigt

Maschinelle Lernverfahren (4)
Support-Vector Maschinen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Beispiele aus verschiedenen Klassen mit einer Hyperebene trennen
- den Abstand derjenigen Beispiele, die der Hyperebene am nächsten liegen, maximieren (garantiert eine gute Generalisierbarkeit)



Quelle: Data Mining: Practical Machine Learning Tools and Techniques

Maschinelle Lernverfahren (4)
Support-Vector Maschinen

TECHNISCHE UNIVERSITÄT DARMSTADT

- die Daten in einem Raum höherer Dimension abbilden
- Berechnung im hochdimensionalen Raum mit Hilfe von Kernel Funktionen:
  - RBF Kernel: $K(x, y) = e^{-\gamma \|x-y\|^2}$
  - Polynomiale Kernel: $K(x, y) = (\langle x \cdot y \rangle + c)^{d}$
  - Sigmoid Kernel: $K(x, y) = tanh\,(\gamma \langle x \cdot y \rangle + c)$
- die Daten in verschiedenen Kategorien klassifizieren

Experimente

TECHNISCHE UNIVERSITÄT DARMSTADT

- WEKA (Experiment Environment)
- 21 Datensätze
- 5 Algorithmen:
  - 1 Entscheidungsbaum-Lerner (J48)
  - 1 Regel-Lerner (JRip)
  - 1 Ensemble (Random Forest)
  - 2 Support-Vector Maschinen (SMO und LibSVM)

Experimente
21 Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- Klassen:
  - 6 Datensätze mit binären Klassen
  - 15 Datensätze mit Multi-Klassen
- Instanzen:
  - die Größe der Datensätze variiert zwischen 24 and 2310
- Attribute:
  - nominale und numerische Attribute
  - die Anzahl der Attribute variiert zwischen 5 and 230

Experimente: 5 Algorithmen
1- J48

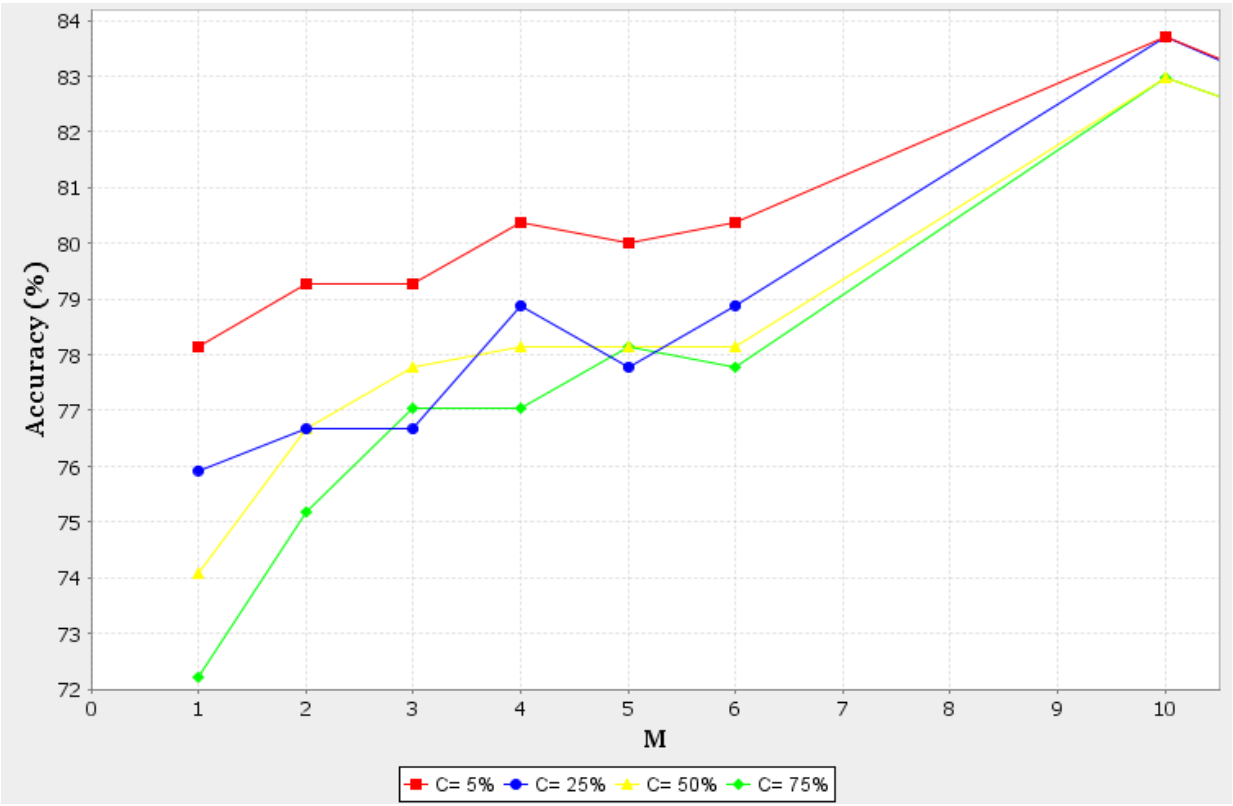
Parametereinstellungen und maximale Genauigkeit

TECHNISCHE UNIVERSITÄT DARMSTADT

- **C:** der Konfidenzfaktor (12 Werte):
  - 0.005, 0.01, 0.05, 0.1, 0.15, 0.2 und 0.25: 11 Datensätze
    2 Datensätze für C = 0.25
  - 0.3, 0.4
  - 0.5, 0.75 und 0.99 : 6 Datensätze: 2 Datensätze für C = 0.5
    4 Datensätze für C = 0.75
- **M:** die minimale Anzahl von Instanzen pro Blatt (11 Werte)
  - 1 : 8 Datensätze (sehr spezifische Bäume)
  - 2, 3, 4, 5
  - 6, 10, 20, 30, 60 und 100 : 7 Datensätze
- **B:** die binäre Aufteilung für nominale Attribute verwenden
  (Standardwert: false)
- **A:** die Laplace-Glättung verwenden (Standardwert: false)

Experimente: 5 Algorithmen
1- J48

TECHNISCHE UNIVERSITÄT DARMSTADT



Veränderung der Genauigkeit in Abhängigkeit von C und M (Datensatz: heart-statlog)

Experimente: 5 Algorithmen
2- Random Forest

Parametereinstellungen und maximale Genauigkeit

TECHNISCHE UNIVERSITÄT DARMSTADT

- **I:** die Anzahl der Bäume (11 Werte):
  - 5, 10, 20, 30, 50 und 75
  - 100, 250, 350, 500 und 700: 12 Datensätze (mehr als die Hälfte unseres Datensatzes)
- **K:** die Anzahl der zufällig zu berücksichtigenden Attribute
  - 12 Werte: 0, 1, 2, 3, 4, 5, 7, 10, 15, 20, 50 und 100
- **Depth:** die maximale Tiefe der Bäume
  - 10 Werte: 0 (unbegrenzte Tiefe), 2, 3, 5, 10, 20 , 30, 50, 75 und 100

Experimente: 5 Algorithmen
3- JRip

Parametereinstellungen und maximale Genauigkeit

TECHNISCHE UNIVERSITÄT DARMSTADT

- **O:** die Anzahl der Optimierungen (12 Werte):
  - 1, 2, 3 und 5: 11 Datensätze (nach einer Optimierung, für das Drittel unseres Datensatzes: 7 Datensätze)
  - 10, 20, 50, 80, 100, 200, 400 und 600
- **N:** die minimale Gewicht der Instanzen in einer Regel (9 Werte)
  - 1, 2, 3, 4 und 5: 18 Datensätze (wenn N=1 für 10 Datensätze (ungefähr die Hälfte unseres Datensatzes))
  - 6, 10, 20 und 30

Experimente: 5 Algorithmen
3- JRip

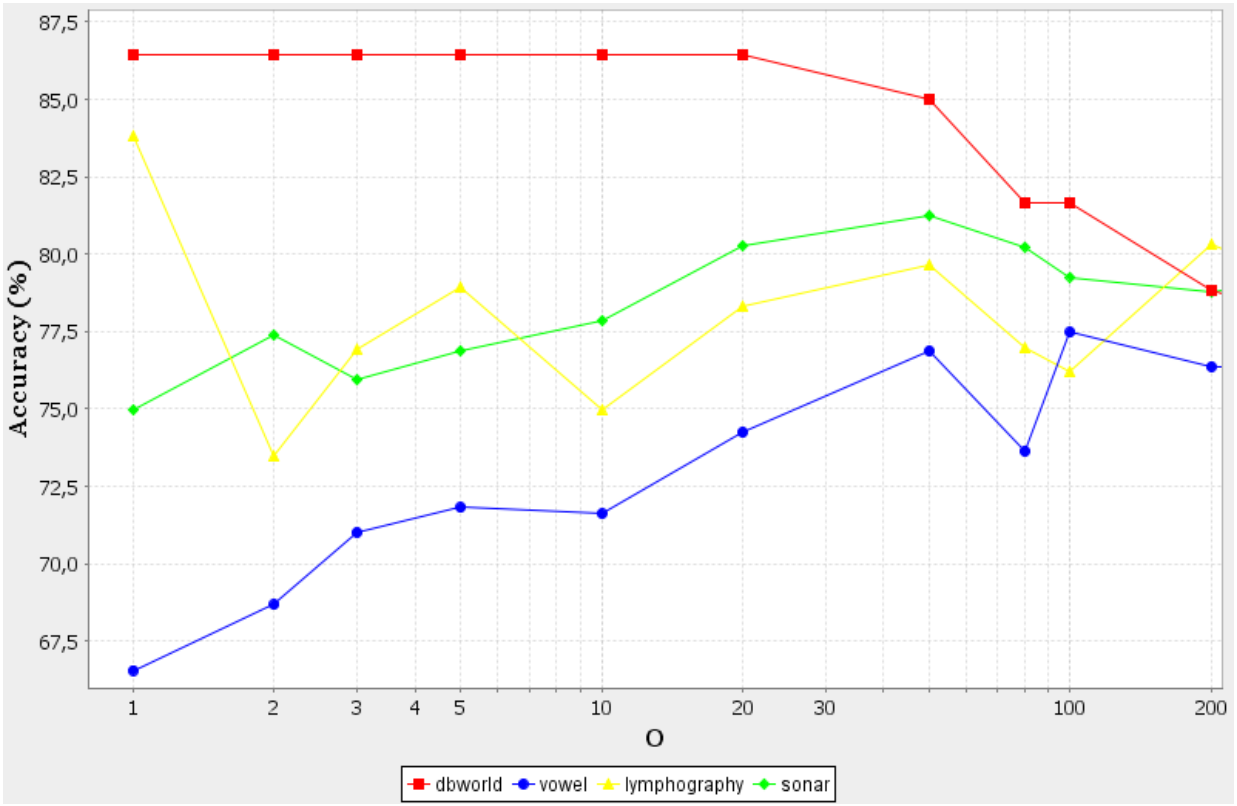
Parametereinstellungen und maximale Genauigkeit

TECHNISCHE UNIVERSITÄT DARMSTADT

- **F:** die Verteilung der Datenmenge (Growingmenge und Pruningmenge)
  - 4 Werte: 2, 3, 5 und 10
- **P:** ob das Pruning verwendet wird (Standardwert: true)
- **E:** ob die 50% Fehlerrate im Abbruchkriterium enthalten ist (Standardwert: true)

Experimente: 5 Algorithmen
3- JRip

TECHNISCHE UNIVERSITÄT DARMSTADT



Veränderung der Genauigkeit in Abhängigkeit von der Anzahl der Optimierungen für verschiedene Datensätze

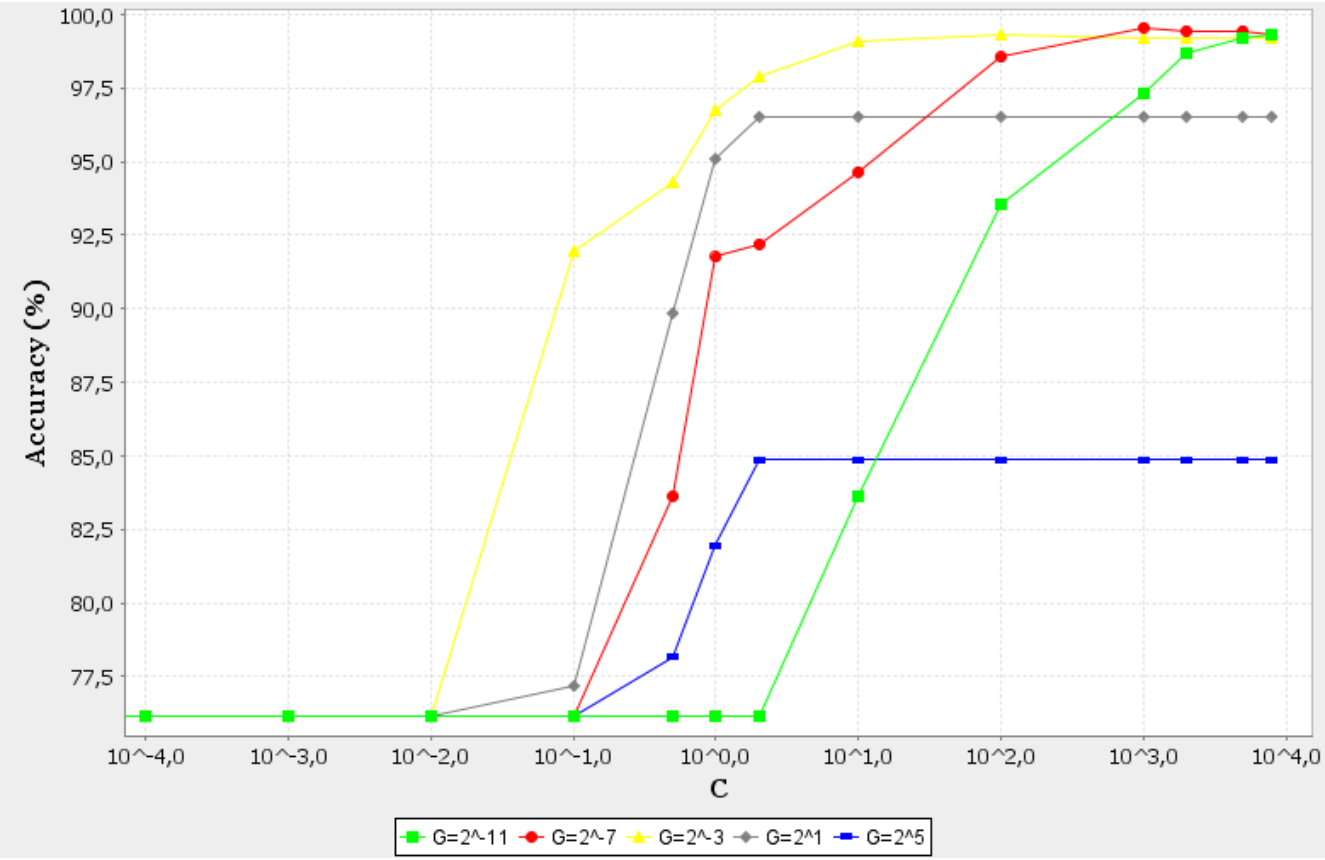
Experimente: 5 Algorithmen
4- SMO

TECHNISCHE UNIVERSITÄT DARMSTADT

- **C:** der Kostenparameter.
  - 14 Werte: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1, 2, 10, 100, 1000, 2000, 5000 und 8000
- **N:** Wie werden die Daten umgewandelt?
  - 0 (Normalisierung der Daten)
  - 1 (Standardisierung der Daten)
  - 2 (keine Umwandlung der Daten)
- γ**:** der gamma Parameter (Breite des Kernels)
  - 20 Werte: $0, 2^{-13}, 2^{-12}, ..., 2^4, 2^5$
- **E:** der Exponent des Kernels
  - 5 Werte: 1, 2, 3, 4 und 5

Experimente: 5 Algorithmen
4- SMO

TECHNISCHE UNIVERSITÄT DARMSTADT



Veränderung der Genauigkeit in Abhängigkeit von C und G (RBF Kernel: N=0 Datensatz: anneal)

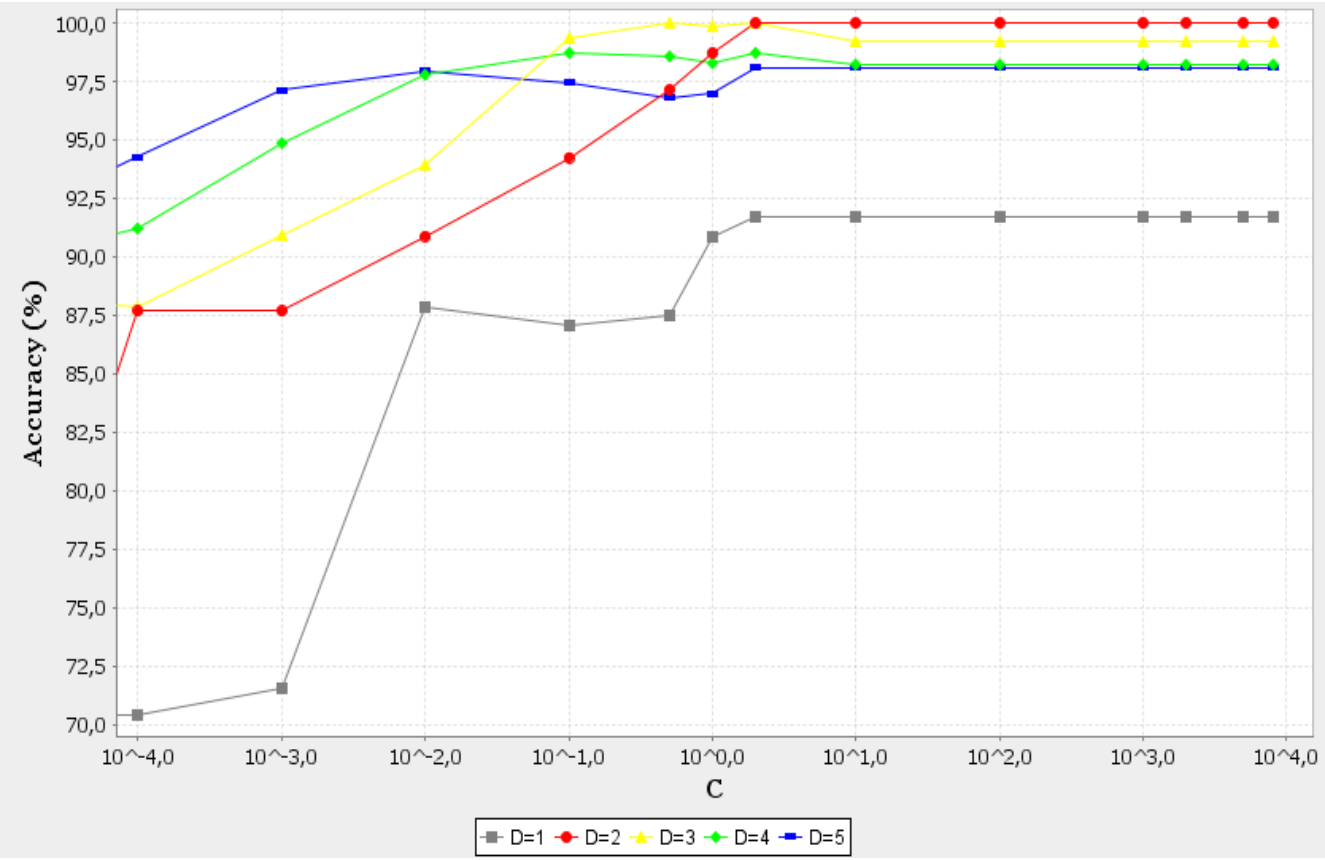
Experimente: 5 Algorithmen
5- LibSVM

TECHNISCHE UNIVERSITÄT DARMSTADT

- **C:** Kostenparameter.
  - 14 Werte: 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5, 1, 2, 10, 100, 1000, 2000, 5000 und 8000
- **Z:** ob die Daten normalisiert werden (Standardwert: false)
- γ**:** der gamma Parameter (Breite des Kernels)
  - 20 Werte: $0, 2^{-13}, 2^{-12}, ..., 2^4, 2^5$
- **D:** der Exponent des Kernels
  - 5 Werte: 1, 2, 3, 4 und 5
- **R:** Konstante
  - 3 Werte: 0, -1000, 1000

Experimente: 5 Algorithmen
5- LibSVM

TECHNISCHE UNIVERSITÄT DARMSTADT



Veränderung der Genauigkeit in Abhängigkeit von C und D (Polykernel: N=0 Datensatz: balance-scale)

Ergebnisse

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	# Instanzen	# Klassen	# Nominale Attribute	# Numerische Attribute
balance-scale	625	3	1	4
german_credit	1000	2	14	7
ionosphere	351	2	1	34
solar-flare-c	1712	6	11	0
spectrometer2	531	48	3	100
zoo	101	7	17	1

Ergebnisse
1- J48

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	C	M	B / A	Genauigkeit (%)
balance-scale	0.5	1		80.002
german_credit	0.2	20	B	73.2
ionosphere	0.25	2		91.46
solar-flare-c	0.3	3	B	85.573
spectrometer2	0.1	6		49.531
zoo	0.1	1		94.181

C: der Konfidenzfaktor
M: die minimale Anzahl von Instanzen pro Blatt
B: die binäre Aufteilung für nominale Attribute
A: die Laplace-Glättung

Ergebnisse
2- Random Forest

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	# Bäume	# Features	Tiefe	Genauigkeit (%)
balance-scale	100	1	5	88.476
german_credit	250	10	10	77.5
ionosphere	20	5	5	95.452
solar-flare-c	350	0	5	85.981
spectrometer2	250	0	20	58.567
zoo	20	2	10	97.09

Ergebnisse
3- JRip

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	F	N	O	Fehlerrate / Pruning	Genauigkeit (%)
balance-scale	2	2	1	ja	82.096
german_credit	2	20	20	ja	75
ionosphere	3	4	1	ja	91.476
solar-flare-c	2	1	1	ja	85.573
spectrometer2	5	1	400	ja	43.871
zoo	2	1	3	ja	91.181

F: die Verteilung der Datenmenge (Growingmenge und Pruningmenge)
N: die minimale Gewicht der Instanzen in einer Regel
O: die Anzahl der Optimierungen

Ergebnisse
4- SMO

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	C	Kernel	G / E	N	Genauigkeit (%)
balance-scale	2000	RBF	2^{-3}	0	100
german_credit	1000	RBF	2^{-11}	1	78.3
ionosphere	0.5	Normalized	8	2	96.293
solar-flare-c	10^{-4}	Poly	5	0	86.098
spectrometer2	1000	Normalized	2	2	71.942
zoo	10	RBF	2^{-3}	2	98

Ergebnisse
5- LibSVM

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	C	Kernel	G / D	Normalize	Genauigkeit (%)
balance-scale	10	Polykernel	2		100
german_credit	5000	Polykernel	1	ja	77
ionosphere	100	RBF	2^{-1}		95.174
solar-flare-c	1	RBF	2^{2}		85.922
spectrometer2	10^{-4}	Polykernel	1		71.757
zoo	0.5	linear	-		96.09

Vergleich der Algorithmen (1)

Über einzelne Datensätze: Maximale und minimale Genauigkeit - Mittelwert - Standardabweichung

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	J48	Random Forest	JRip	SMO	LibSVM
balance-scale	80.002 64.493 73.325 4.486	88.476 72.956 81.38 2.019	82.096 71.208 78.376 1.998	100 19.859 80.537 16.529	100 4.966 64.22 21.577
zoo	94.181 40.636 76.154 19.685	97.09 42.636 73.077 23.01	91.181 40.636 73.911 19.871	98 40.636 61.139 25.041	96.09 23.727 46.43 17.532
spectrometer2	49.531 13.19 42.659 7.577	58.567 10.359 43.345 13.851	43.871 10.359 31.027 8.538	71.942 10.359 21.079 17.245	71.757 5.461 16.75 17.56

Vergleich der Algorithmen (1)

Über einzelne Datensätze: Maximale und minimale Genauigkeit - Mittelwert - Standardabweichung

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	J48	Random Forest	JRip	SMO	LibSVM
ionosphere	91.46 74.087 88.32 4.684	95.452 84.333 92.609 1.254	91.476 80.373 89.342 1.761	96.293 64.103 78.517 13.38	95.174 39.587 69.534 12.154
german_credit	73.2 65.7 70.423 1.413	77.5 69.8 74.47 1.83	75 68.4 72.415 0.995	78.3 61.2 70.857 2.624	77 36.5 68.374 5.934
solar-flare-c	85.573 84.055 85.076 0.268	85.981 83.47 84.55 0.496	85.573 83.703 85.098 0.352	86.098 83.877 85.193 0.273	85.922 30.804 83.941 5.5

Vergleich der Algorithmen (3)
Über den gesamten Datensatz

TECHNISCHE UNIVERSITÄT DARMSTADT

	J48	Random Forest	JRip	SMO	LibSVM
P	528	1320	1296	1204	2688
N	11088	27720	27216	25284	56448
Mittelwert (in %)	75.853	80.223	75.796	68.774	58.3
Varianz	245.078	222.619	264.655	505.177	634.991
Standardabweichung	15.654	14.92	16.268	22.476	25.199

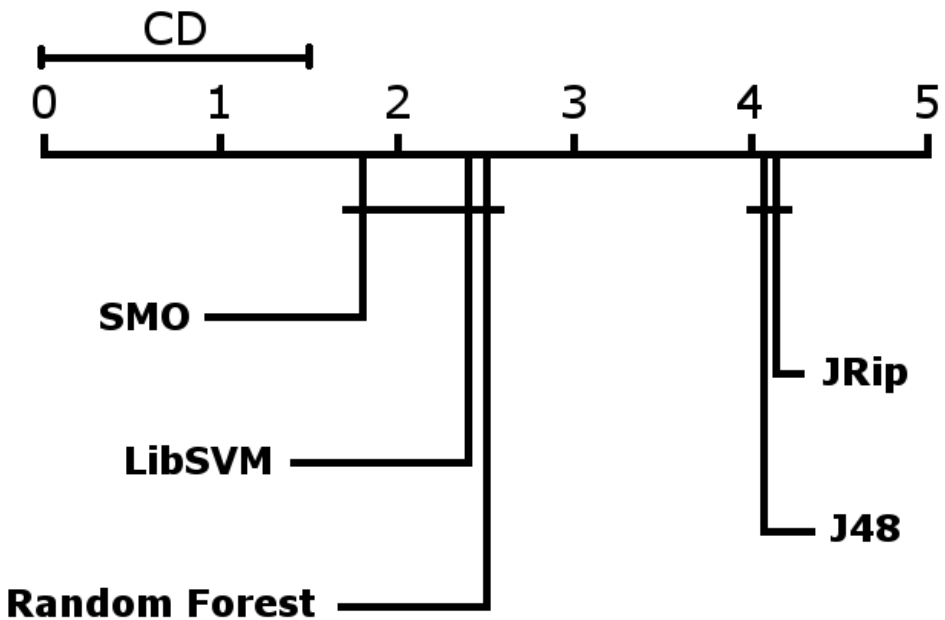
P: Anzahl der Parametereinstellungen

N: Anzahl der Parametereinstellungen * Anzahl der Datensätze

Vergleich der Algorithmen (4)
Signifikanztest

TECHNISCHE UNIVERSITÄT DARMSTADT

- den durchschnittlichen Rang für jeden Algorithmus bestimmen
- Signifikanzniveau α = 0.01
- die kritische Differenz berechnen (CD: 1.588)



SMO	1.88
LibSVM	2.404
Random Forest	2.5
J48	4.095
JRip	4.119

Zusammenfassung und Konklusion

TECHNISCHE UNIVERSITÄT DARMSTADT

- Verbesserung der Performanz nach der Optimierung der Parameter
- die symbolische Ansätze sind mehr stabil (niedrige Varianz)
- signifikanter Unterschied zwischen zwei Gruppen:
  - SVMs und Random Forest
  - J48 und JRip
- Random Forest:
  - gute Genauigkeitswerte (ähnlich wie bei den SVMs)
  - niedrige Varianz (ähnlich wie bei den symbolischen Ansätzen)
- Wie verhalten sich andere Ensemble Lerner im Vergleich mit den SVMs ?

Vielen Dank für die Aufmerksamkeit !

Vergleich der Algorithmen (2)

Über einzelne Datensätze:

optimierte Parameter vs Standardeinstellungen

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensätze	J48	Random Forest	JRip	SMO	LibSVM
balance-scale	80.002 76.653	88.476 82.237	82.096 79.05	100 87.677	100 89.754
german_credit	73.2 70.5	77.5 74.3	75 71.7	78.3 75.1	77 70
ionosphere	91.46 91.46	95.452 94.595	91.476 89.753	96.293 88.603	95.174 93.46
solar-flare-c	85.573 85.105	85.981 84.346	85.573 85.397	86.098 85.163	85.922 85.163
spectrometer2	49.531 47.449	58.567 49.346	43.871 31.656	71.942 51.418	71.757 10.359
zoo	94.181 92.181	97.09 91.09	91.181 87.272	98 96.181	96.09 49.636