Auswahl der Daten

- Datensets haben verschiedenste Charakteristiken

Datensatz	Anzahl Beispiele	Numerische Attribute	Nominale Attribute	Anzahl Klassen...	... im Zielattribut
zoo	101	1	15	7	type
auto	205	15	10	7	symbolic
soybean	683	19	16	19	class
sick	3772	7	22	2	class
letter	20000	16	0	26	letter

Unterschiede zwischen den Datensets

- Die Genauigkeit zwischen den einzelnen Datensets ist sehr verschieden.
- Man kann aber nicht sagen, daß eine Genauigkeit von 95% besser ist als eine Genauigkeit von 35%!
- Beispiel:
  - Datenset A:
    - 2 Klassen
    - 99% der Beispiele Klasse +, 1% Klasse -
    - Algorithmus erreicht 95%
      - schlechter als immer die Klasse + raten!
  - Datenset B:
    - 5 Klassen
    - alle 5 gleich groß (ca. 20% der Beispiele)
    - Algorithmus erreicht 35%
      - immerhin besser als zufällig raten!

Abbildung 2: Falsch klassifiziertes Beispiel beim Klassifizieren im Datensatz zoo

Verschiedene Cross-Validierungen

Data	Accuracy (%)				
	2 folds	5 folds	10 folds	20 folds	Accuracy (%)
autos.arff	68.7805	79.5122	81.9512	81.9512	84.8780
hepatitis.arff	81.2903	80.6452	83.8710	81.9355	80.0000
labor.arff	71.9298	77.1930	73.6842	75.4386	77.1930
soybean.arff	87.8477	90.7760	91.5081	92.3865	92.6794
zoo.arff	92.0792	92.0792	92.0792	92.0792	92.0792

- Varianz in den verschiedenen Genauigkeitsabschätzungen mitunter sehr groß!

  → Vorsicht bei deren Interpretation
  - Genauigkeitsabschätzung ist nur eine Abschätzung!
- Unterschiede resultieren z.T. auch aus unterschiedlichen Größen der Trainings-Sets
  - Größere Training-Sets sind den realen Bedingungen ähnlicher
  - Andererseits: größere Abhängigkeiten zwischen den Sets und größerer Aufwand beim Evaluieren

Verschiedene Seeds für Cross-Validation

- Auch hier gibt es eine hohe Varianz
  → Vorsicht bei Interpretation von Genauigkeitsunterschieden zwischen Algorithmen!
  - Unterschiede z.T. auf Varianz zurückzuführen
  - Oft wird n-fache m-fold Cross-Validierung angewandt, um Varianz zu senken

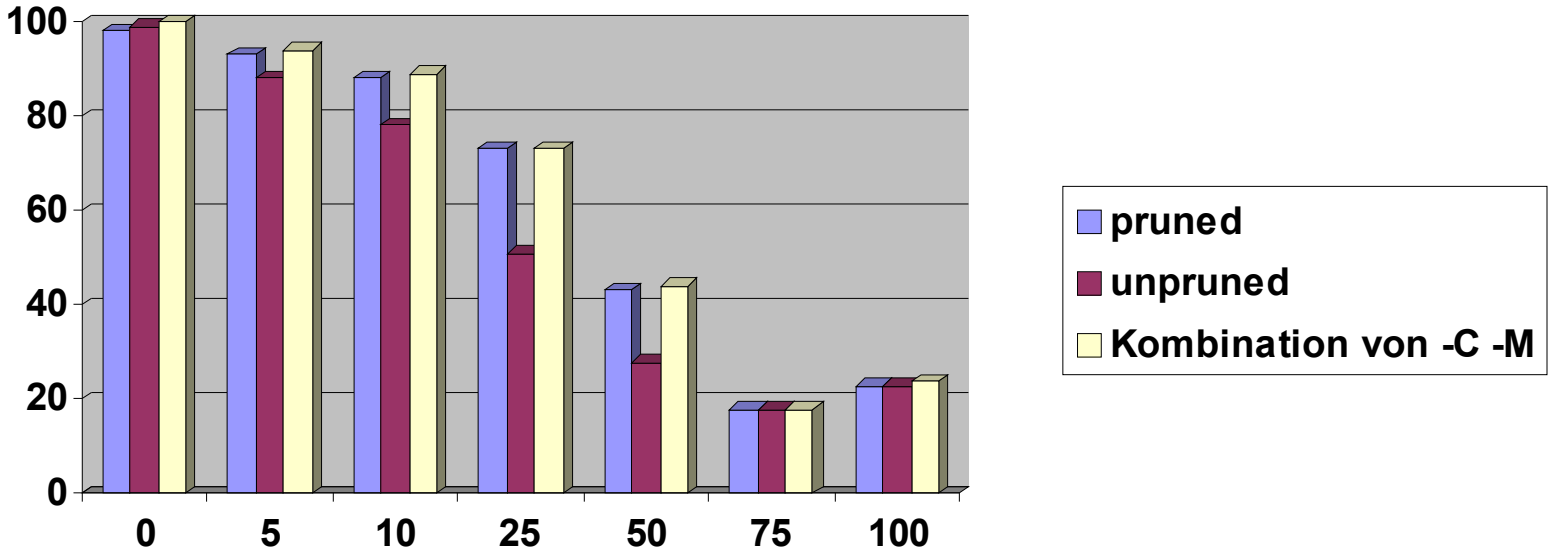
- man kann nicht sagen:
  - "Höhere Anzahl von Cross-Validations führen zu größerer Genauigkeit"
  - "Durch gute Wahl der Seed kann man die Genauigkeit verbessern"
    → analog zu (Scherz-)Artikel über Data Set Selection im "Journal of Machine Learning Gossip"
  - "10-fold ist gut, weil 20-fold langsamer ist und nicht signifikant besser."

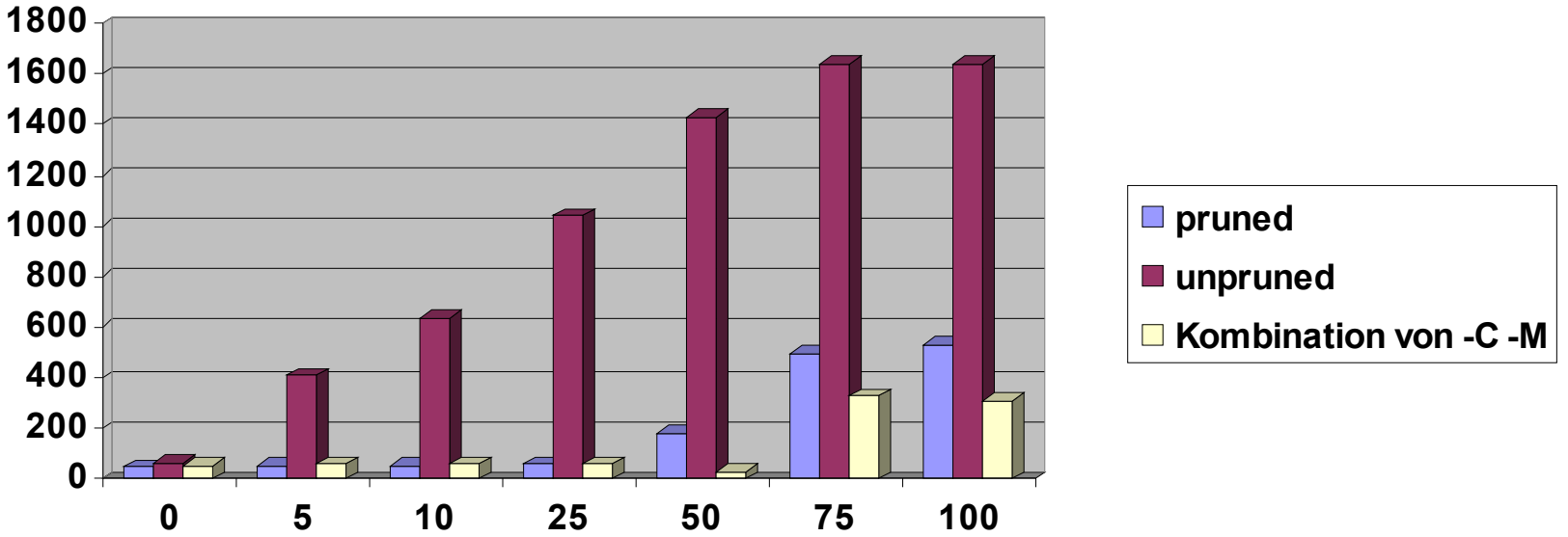
Vergleich von Algorithmen

Dataset	(1) trees.J4	(2) funct	(3) rules	(4) bayes	(5) trees
kr-vs-kp	(100) 99.44	95.79 *	99.21	87.79 *	99.44
soybean	(100) 91.78	93.10	91.85	92.94	90.69
labor-neg-data	(100) 78.60	92.97 v	83.70	93.57 v	79.13
iris	(100) 94.73	96.27	93.93	95.53	94.80
contact-lenses	(100) 83.50	72.50	80.67	76.17	75.67
weather.symbolic	(100) 47.50	65.00	72.00	57.50	64.00
	(v/ /*)	(1/4/1)	(0/6/0)	(1/4/1)	(0/6/0)

- Typischerweise ist kein Algorithmus immer besser als alle anderen und kein Algorithmus immer schlechter als alle anderen
  - Die Wahl des Algorithmus hängt letztendlich auch von der Problemstellung ab
  - Generelle Richtlinien, welcher Algorithmus auf welches Problem paßt, gibt es nur wenige

Noise and Pruning





Noise and Pruning

	Rauschen	unpruned	0,25/2	0,25/1	0,25/3	0,25/7	0,15/2	0,05/2	0,15/1
ER	0%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
ER	15%	28,30%	15,01%	15,02%	15,01%	14,99%	15,03%	15,03%	15,03%
ER	40%	48,54%	43,11%	43,22%	43,02%	42,10%	41,01%	40,55%	41,04%
Size	0%	30,00	30,00	30,00	30,00	30,00	30,00	30,00	30,00
Size	15%	13801,00	32,60	35,40	31,40	30,00	29,60	29,60	29,60
Size	40%	21966,10	2140,90	2464,60	1625,60	700,80	177,90	15,90	199,80

- Pruning ist wichtig!
  - ansonsten große Bäume und geringe Genauigkeit!
- Anmerkungen:
  - Bei -m Parameter kann man ohne weiteres größere Werte testen!
  - In einigen Fällen war 100% Noise gleich gut wie 0% Noise. Warum?

Vergleich JRip vs. J48: Genauigkeit

- Üblicherweise kaum Unterschiede in der Genauigkeit
  - JRip funktioniert möglicherweise bei Mehr-Klassen-Problemen schlechter

Dataset	(1) rules.JR	(2) trees
balance-scale	80.30	77.82
hypothyroid	99.42	99.54
lymph	76.31	75.84
sonar	73.40	73.61
soybean	91.85	91.78

Vergleich JRip vs. J48: Größe

- Aber große Unterschiede in der Größe der Bäume
  - JRip pruned aggressiver
  - ist auch nicht daran gebunden, nicht überlappende Regeln zu lernen

Dataset	(1) rules.JR	(2) trees
balance-scale	11.33	41.60
hypothyroid	5.04	14.44
lymph	6.21	17.30
sonar	4.69	14.45
soybean	27.09	61.12

Diskretisierung

		numerische (ursprüngliche) Daten	nominelle Daten (supervised)	nominelle Daten (unsupervised)
Normale J48	Anzahl Blätter	3	5	11
	Gesamtgröße	5	8	13
	Genauigkeit	73,68%	80,70%	57,42%
Discretized J48 Kombination	Anzahl Blätter	5	-	-
	Gesamtgröße	8	-	-
	Genauigkeit	71,93%	-	-

- Durch die Diskretisierung des gesamten Datensets fließt Information über das Test Set in die Evaluierung
- daher kann es zu viel zu optimistischen Abschätzungen der Genauigkeit kommen.
  - Im Praxis-Fall werden die Beispiele, auf denen der Klassifizierer angewendet wird, ja auch nicht beim Trainieren berücksichtigt!
- Größe der Bäume wächst oft ebenfalls mit Diskretisierung
- Genauigkeit im Vergleich zu Original-Daten kann aber auch steigen! (z.B. sonar)

Wettbewerb

- Generell zeigt sich
  - Je mehr Geduld, desto bessere Resultate :-)
  - einfaches J48 hat <90% Genauigkeit

- Datenset:
  - Optical DigitRecognition
    - file ./optdigits-orig.names
  - Es existiert ein separates Test Set
    - file ./task-test.arff
  - original publication:
    - Accuracy on test set < 98%
    - file ./optdigits.ps

Hall of Fame

98.93%	**98.887%**	removed constant attributs. SMO -E 3	S. Droste
98.93%	**98.887%**	SMO -E 3	Stefan Pohl
98,8752%	**98.8314%**	IBk -K 4 -W 0 -I -s 7	Anh-Minh Trinh
98.69%	**98.7757%**	removed constant & nearly constant attributes	
		IBk -K 3 -I -W 0	Jakob Andersen, Kimmo Palander
98.4567%	**98.8314%**	IB1	Thomas Wolf
98.3521%	**98.8314%**	IB1	Dominic Hajek, Heidrich, Schnitzspan
98,274%	**98.2749%**	SMO -V 15	David Becker, Wladimir Awerbuch, C. Stöhr
98,1951%	**98.7757%**	IBk -K 3 -x 3	Holger Koetting
98.169%	**98.1636%**	AdaBoost -I 100 -W J48	A. Beckhaus, J. Gönsch, T. Krause, C. Steinhardt
98.1428%	**98.2749%**	SMO	Matteo Ceruti
98,1428%	**98.2749%**	SMO	Benjamin Rank
98.0643%	**96.8837%**	AdaBoost -I 20 -W J48	Nicola Gutberlet
98.06%	**96.8837 %**	AdaBoost -I 20 -W J48	Jacqueline Vogel, Arne Pottharst
98,04%	**97.9967%**	SMO -N 1	Dietrich Bubenheim, Peter Eiselt
97.2535%	**96.1603%**	AdaBoost -I 10 -W J48	Jördis Hensen
92.8067%	**91.5415%**	BayesNet	Chr.Seufert, J-N. Sulzmann, T. Reinhard
89 %	**95.6038%**	FSS | InfoGain 25 Atts. AdaBoost -W J48	Nikolay Jetchev, Boiana Ivanova, M. Ruskov

recht gute Korrelation der Genauigkeitsabschätzungen (erste Spalte)
mit den auf dem unabhängigen Test-Set gemessenen (zweite Spalte)