Projekt Maschinelles Lernen WS 06/07

1. Auswahl der Daten
2. Evaluierung
3. Noise und Pruning
4. Regel-Lernen
5. ROC-Kurven
6. Pre-Processing
7. Entdecken von Assoziationsregeln
8. Ensemble-Lernen
9. Wettbewerb

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!

Variieren der Cross-Validation

Datensätzen		*Contact Lenses*	*Vehicle*	*Labor*	*Autos*	*Zoo*
Instanzen		**24**	**57**	**846**	**205**	**101**
All-training-data		*91.7%*	*87.2%*	*96.9%*	*95.1%*	*99.0%*
Cross-Validation	2-fold	*66.7%*	*68.9%*	*68.9%*	*68.8%*	*92.1%*
	5-fold	*83.3%*	*72.1%*	*72.1%*	*79.5%*	*92.1%*
	10-fold	*83.3%*	*72.5%*	*72.5%*	*82.0%*	*92.1%*
	20-fold	*83.3%*	*73.5%*	*73.5%*	*82.0%*	*92.1%*
	10 mal 10-fold	*83.8%*	*79.8%*	*72.5%*	*81.4%*	*92.2%*
	Leave-one-out	*83.3%*	*75.3%*	*75.3%*	*84.9%*	*92.1%*

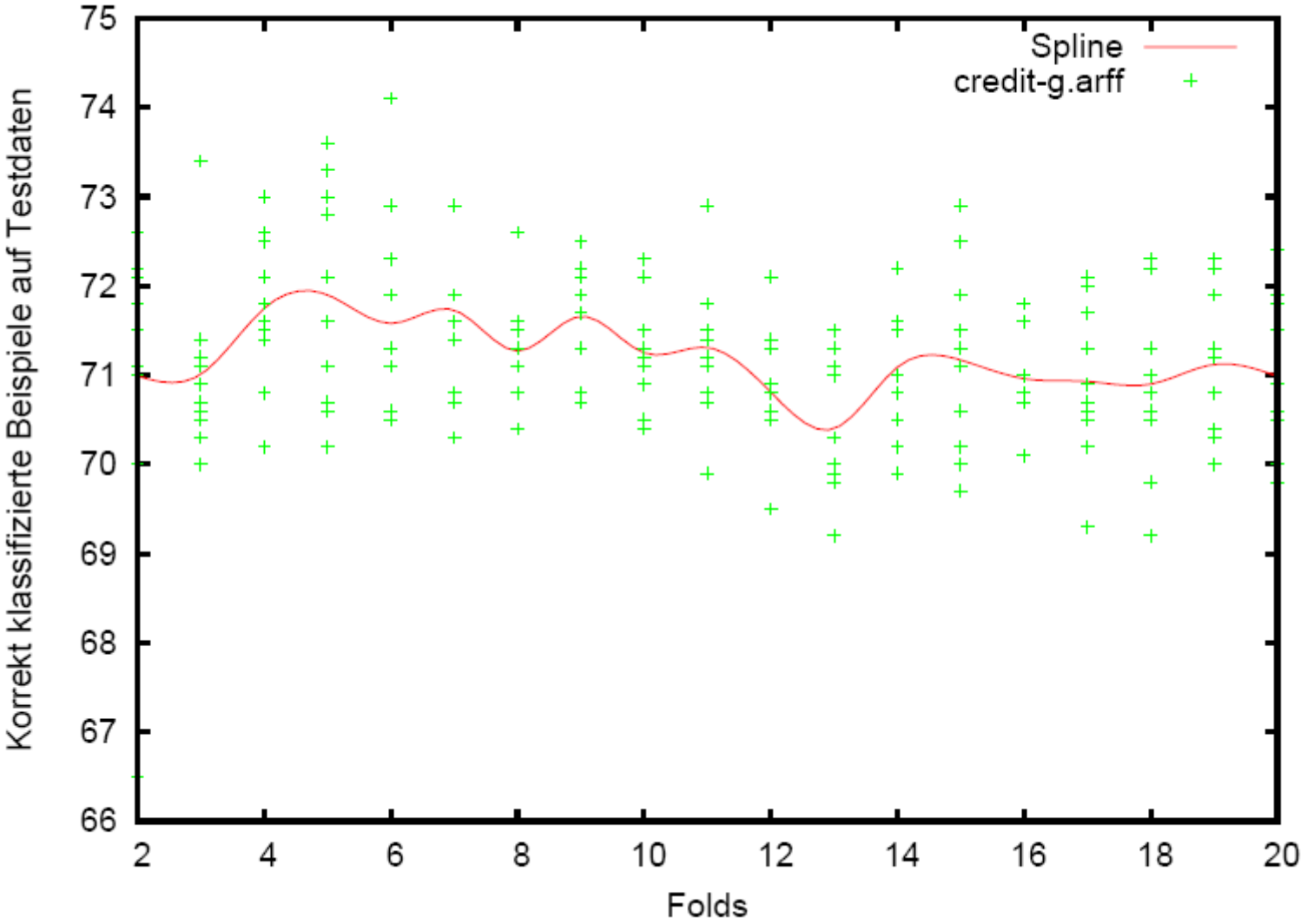
Verschiedene Cross-Validierungen

- 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 ursprünglichen Daten ähnlicher
  - Andererseits: größere Abhängigkeiten zwischen den Sets und größerer Aufwand beim Evaluieren

- "... ist ein Trend zu beobachten, dass bei steigender Anzahl an Folds die gemessene Genauigkeit steigt. Dies hängt vermutlich damit zusammen, dass hier mehr Beispiele zum Trainieren verwendet werden konnten"

Verschiedene Seeds für CV



Verschiedene Seeds für CV

- Auch hier gibt es eine hohe Varianz
  → Vorsicht bei Interpretation von Genauigkeits-unterschieden zwischen Algorithmen!
    - Unterschiede z.T. auf Varianz zurückzuführen
- Oft wird n-fache m-fold Cross-Validierung angewandt, um Varianz zu senken
  - Falsche Interpretation:
    - 10-fache 10-fold Cross -validation ist genauer/ungenauer als 10-fold Cross-validation.
  - Richtig:
    - Reduziert die Varianz des Schätzers um den Erwartungswert

Häufige Fehler

- ”Wie deutlich zu sehen ist, funktioniert der Algorithmus mit 10- und 20-fold Crossvaliation am besten”
  - Zu sehen ist nur, daß sich die Genauigkeits-abschätzungen unterscheiden, nicht welche besser bzw. genauer ist!
- “Höhere Anzahl von Cross-Validations führen zu größerer Genauigkeit”
- “Durch gute Wahl der Seed kann man die Genauigkeit verbessern”
- “10-fold ist gut, weil 20-fold langsamer ist und nicht signifikant besser.”
- Größere Datensets (mehr Beispiele) haben eine höhere Genauigkeit

Noise and Pruning

- **mit Default-Parametern (-C 0.25 und –M 2) und mit x% Noise**

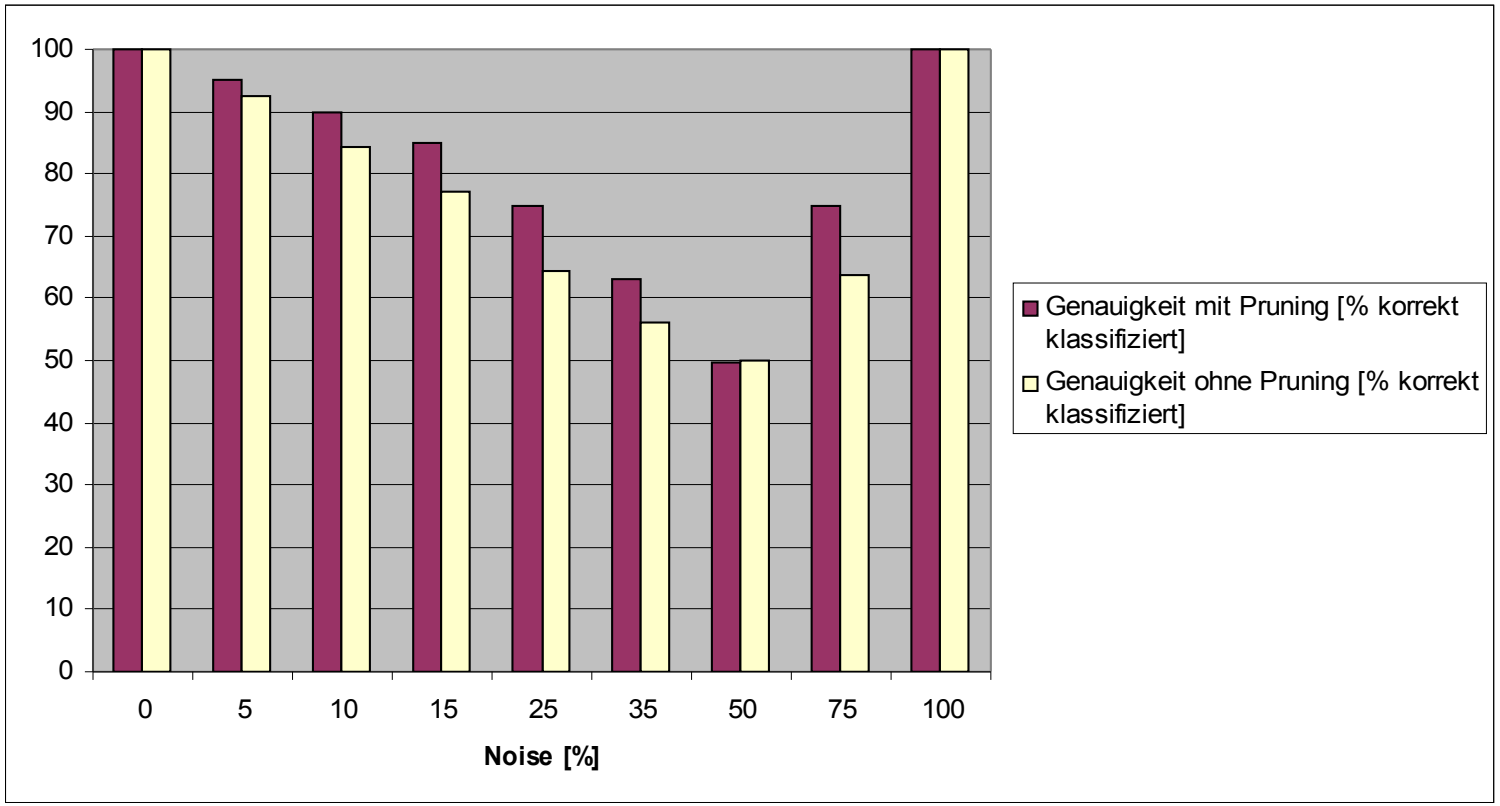
	ohne Noise	5%	10%	25%	50%	75%	100%
Genauigkeit	98.807	93.4517	86.2142	69.3531	52.0944	48.7275	48.8335
Baumgröße	61	46	80	57	333	357	357

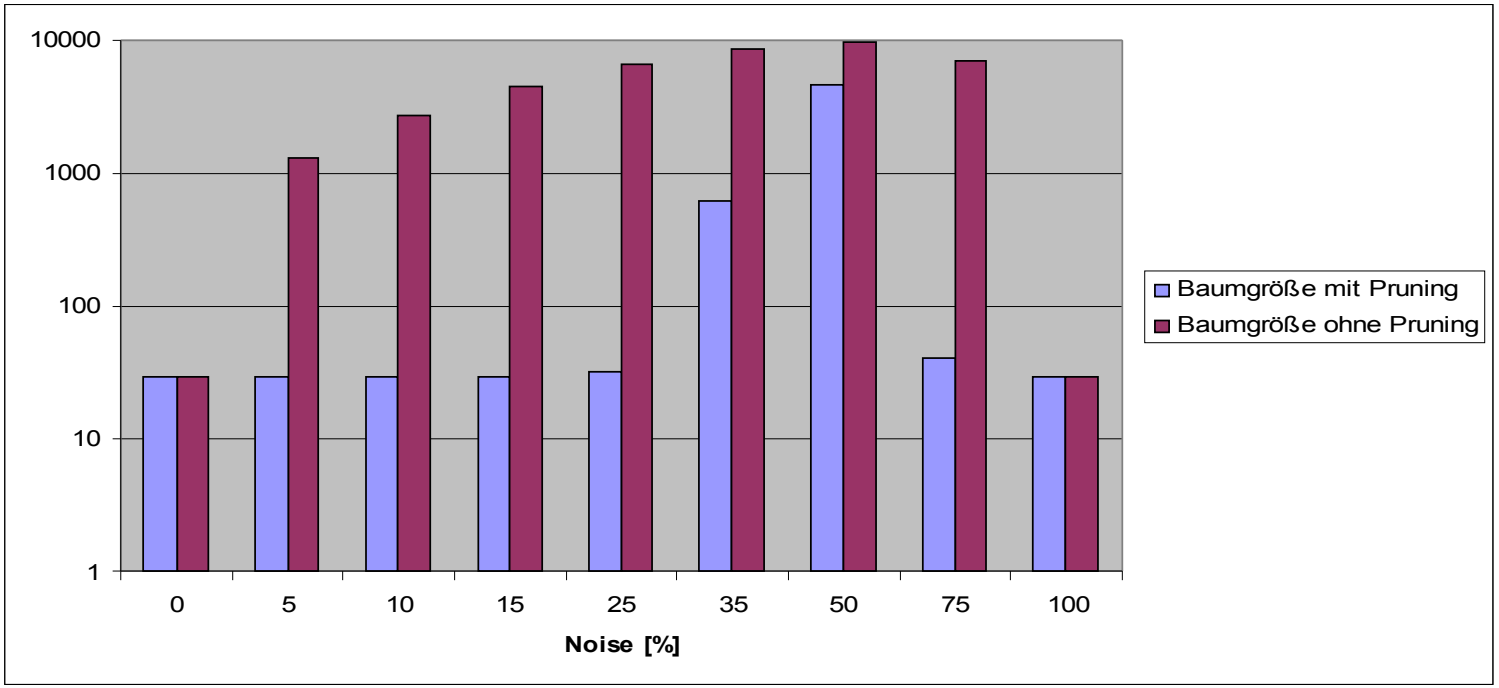
- **ohne Pruning (-U und –M 1) und mit x% Noise**

	ohne Noise	5%	10%	25%	50%	75%	100%
Genauigkeit	98.807	83.2715	85.9756	64.3425	52.4125	50.1591	50.0265
Baumgröße	95	710	512	944	907	824	824

- Pruning ist wichtig!
  - ansonsten große Bäume und geringe Genauigkeit!
  - Sogar bei völlig zufälligen Daten (100% Noise) wird noch etwas gelernt
    - ist das sinnvoll?

Noise and Pruning





Mushroom

- auch bei relativ großem Noise wird noch eine korrekte Theorie gelernt (Zwei-Klassen-Problem)
- die geschätzte Genauigkeit entspricht dem Noise Level
  - warum?

Noise	Genauigkeit	Anz. Bed.	Größe	Blätter
0%	100%	5	30	25
5%	95,0025%	5	30	25
10%	89,9926%	5	30	25
25%	74,8646%	5	33	28
50%	49,7046%	633	4704	4071
75%	74,9138%	12	41	33
90%	89,9680%	5	30	25
95%	95,0025%	5	27	22
100%	100%	5	30	25

Verwandtes Problem

- Für m = 10 und m = 25 entsteht auf meinen Daten derselbe Baum, es werden allerdings verschiedene Genauigkeiten ermittelt.
  - Warum?
  - die Bäume, die entstehen, haben mindestens 25 Beispiele in den Blättern
  - da aber mit CV evaluiert wird, werden auf den einzelnen Folds natürlich unterschiedliche Bäume gelernt (bei m=10 zB welche die min. 10 Beispiele i.d. Blättern haben)
  - daraus resultieren die unterschiedlichen Genauigkeiten

Variieren des Pruning-Levels

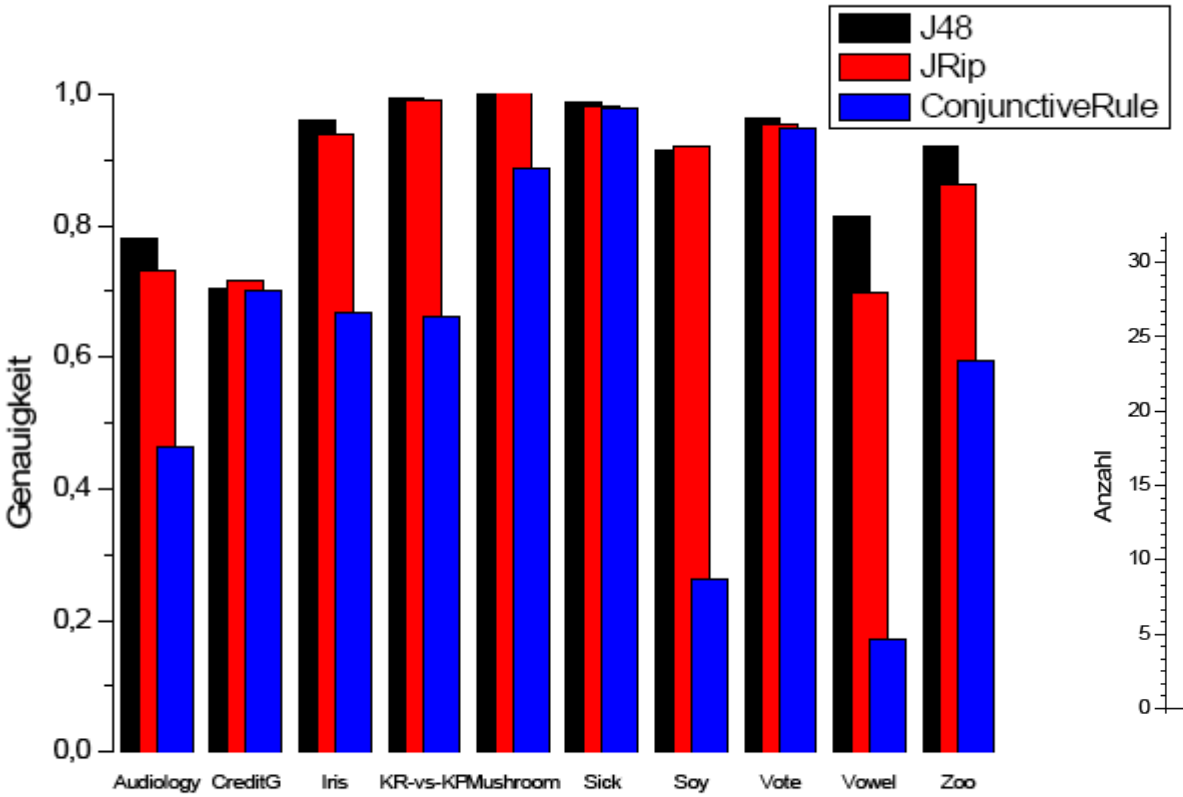
- -m ist eine Anzahl
  - kann auch große Werte annehmen
- -C ist ein Konfidenzmaß
  - Bereich [0,1]
- ab einem gewissen Punkt dominiert -m (-C ist egal)

	C=0,1	**C=0,2**	**C=0,3**	**C=0,4**	**C=0,5**
M=1	98.7805	98.7275	98.754	98.807	**98.913**
M=2	98.701	98.7805	**98.913**	**98.913**	**98.913**
M=5	98.6744	98.7805	98.7805	98.8335	98.807
M=10	98.5154	98.5419	98.5949	98.5419	98.6479
M=20	98.1442	98.1442	98.1442	98.1442	98.1442
M=50	97.9056	97.9056	97.9056	97.9056	97.9056
M=100	97.5345	97.5345	97.5345	97.5345	97.5345

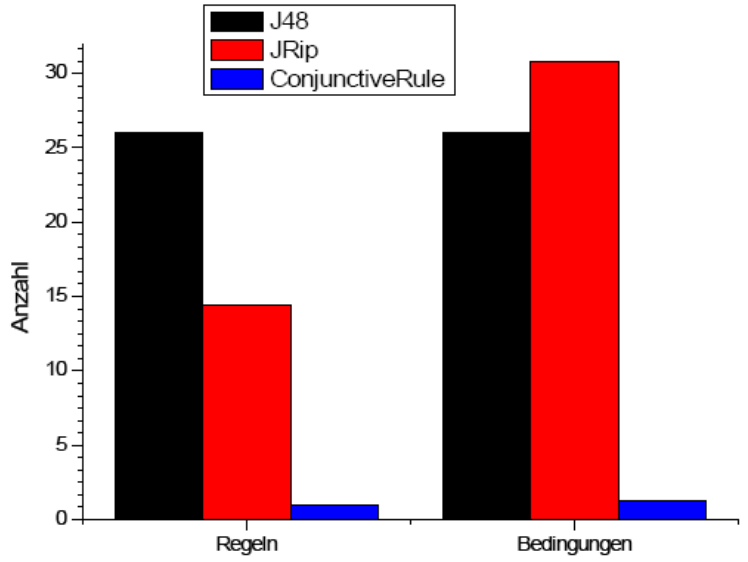
Vergleich Regellerner

Datensatz	Genauigkeit in %			Größe		
	ConjunctiveRule	JRip	J48	ConjunctiveRule (Regelanzahl)	JRip (Regelanzahl)	J48 (Blattanzahl | Baumgröße)
zoo	59.4059	86.1386	92.0792	1	6	9 | 17
splice	62.3824	93.6991	94.0752	1	14	184 | 229
labor	77.193	77.193	73.6842	1	4	3 | 5
colic	81.5217	84.2391	85.3261	1	4	4 | 6
anneal	76.7261	98.3296	98.441	1	7	35 | 47
vowel	17.0707	69.697	81.5152	1	48	106 | 198
soybean	26.2079	91.9473	91.5081	1	26	61 | 93
iris	66.6667	94	96	1	4	5 | 9
glass	44.3925	68.6916	66.8224	1	8	30 | 59
diabetes	68.75	76.0417	73.8281	1	4	20 | 39

Vergleich Regel-Lerner



Genauigkeit der unterschiedlichen Klassifizierer auf verschiedenen Datensätzen



Durchschnittliche Anzahl der Regeln und Bedingungen

Vergleich Regellerner

- ConjunctiveRule natürlich schlechter
  - aber oft nicht viel
- JRip vs. J48: kaum Unterschiede in der Genauigkeit
  - JRip funktioniert möglicherweise bei Mehr-Klassen-Problemen schlechter
- Aber große Unterschiede in der Größe der Bäume
  - JRip pruned aggressiver
  - ist auch nicht daran gebunden, nicht überlappende Regeln zu lernen

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

Diskretisierung

	J48 (ursprüngliche Daten)	J48 (diskretisierte Daten)	FilteredClassifier (ursprüngliche Daten)
Genauigkeit	81.9512 %	83.9024 %	73.1707 %
Größe des Baumes	69	103	103
Anzahl der Blätter	49	90	90

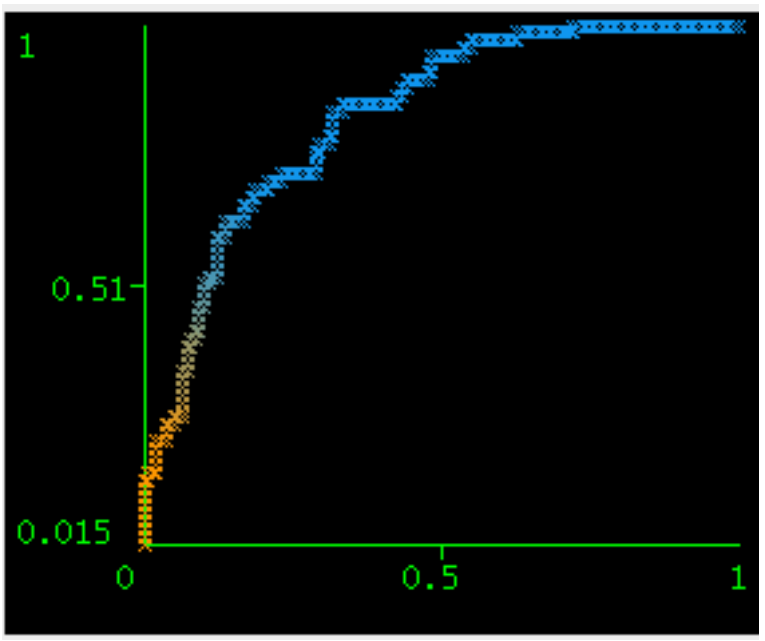
- Durch die Diskretisierung des gesamten Datensets fließt Information über das Test Set in die Evaluierung
- daher kann es zu viel zu optimistischen Abschätzungen der Genauigkeit kommen
  - muß aber nicht, z.B. ionosphere
  - Im Praxis-Fall werden die Beispiele, auf denen der Klassifizierer angewendet 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)

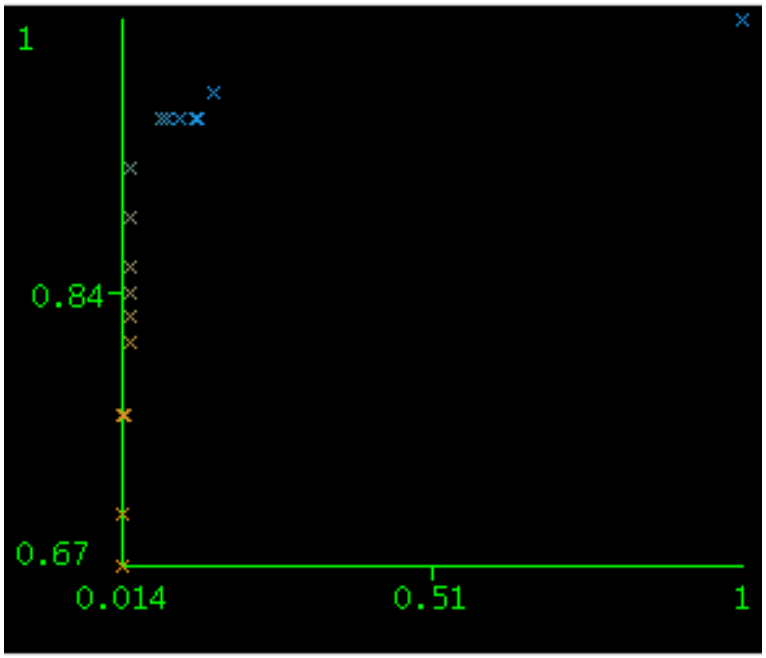
Typische Fehler

- Der Grund hierfür ist, dass zum Ermitteln der Genauigkeit beim FilteredClassifier
  - nicht die diskretisierten Daten verwendet werden,
  - sondern die ursprünglichen Daten,
  - der erzeugte Baum allerdings aus den diskretisierten Daten gelernt wurde!
- Der Unterschied zwischen Trainings- und Testdaten führt somit dazu,
  - dass der FilteredClassifier schlechter abschneidet, als J48 auf diskretisierten Daten, wo sowohl Trainings- als auch Testdaten diskretisiert sind, was zu einer höheren Genauigkeit führt.

→ stimmt natürlich nicht! Im ersten Fall werden alle Daten diskretisiert (im Vorhinein) und im zweiten Fall wird erst der Trainings- und dann der Testfold diskretisiert

ROC Kurven

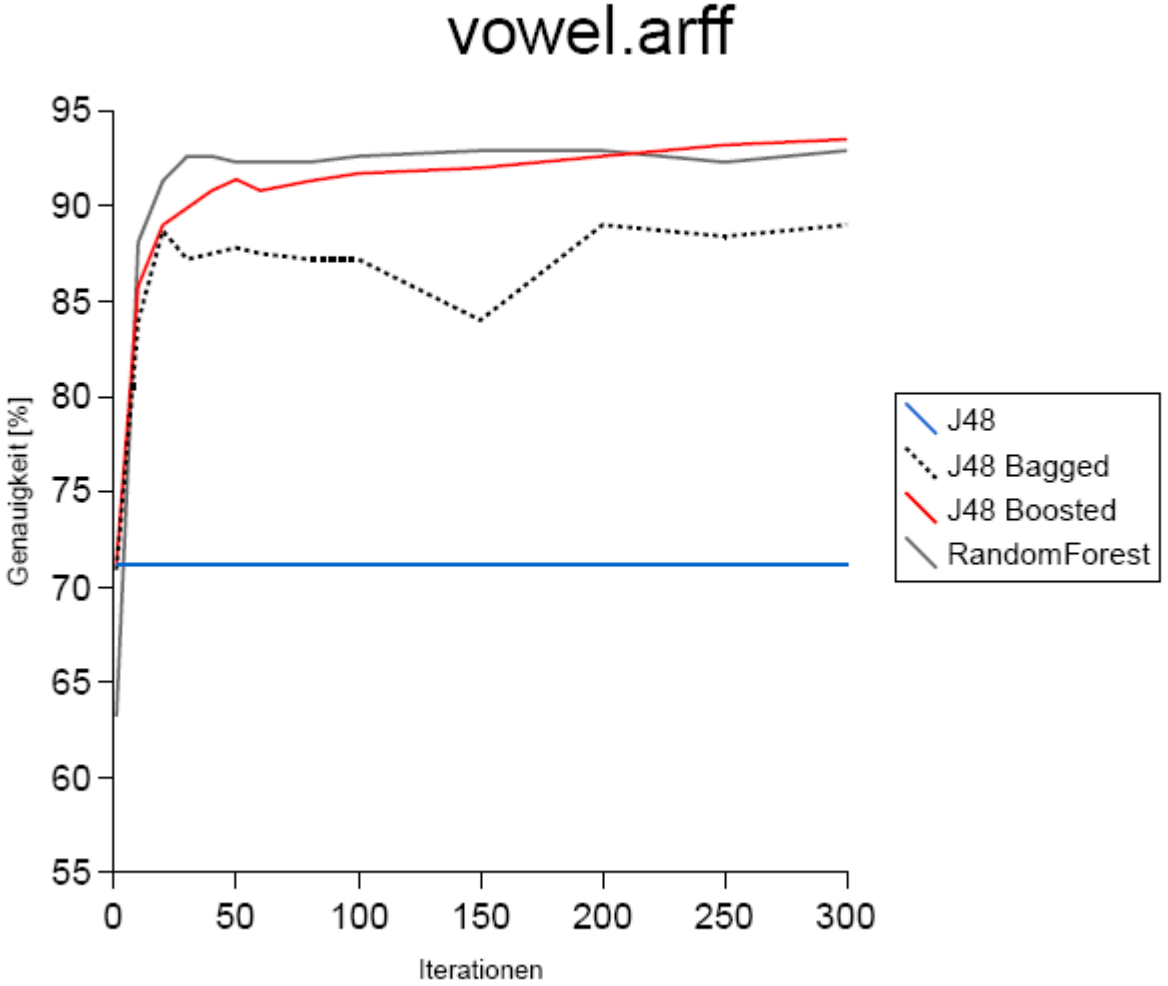




- Das unterschiedliche Aussehen ergibt sich aus der Tatsache, daß bei Entscheidungsbäumen viele Beispiele mit der gleichen Wahrscheinlichkeit bewertet werden.

Ensemble Lernen

- Erhöhung der Anzahl der Iterationen bringt nur anfangs einen Fortschritt, irgendwann wird Sättigung erreicht



- mit größerer Anzahl an Iterationen kann aber auch eine Verschlechterung eintreten (vote, credit-a mit Bagging)

Assoziationsregeln

- Haupterkenntnis:
  - Es is nicht einfach, ohne entsprechende Vorverarbeitung vernünftige Regeln zu lernen
- Viele gefundene Regeln sind selbstverständlich
  - Personen unter 20 verdienen wenig
  - verheiratet UND männlich ==> Ehemann
  - Ehemann ==> männlich UND verheiratet
  - Ehemann ==> verheiratet
- Einige interessantere Regeln:
  - Reiche Amerikaner haben weisse Hautfarbe
  - Alle „craft-repair“ sind männlich

Wettbewerb

- höchste Genauigkeit auf dem Testset: 97,941 %
  - Algorithmus: 1-Nearest Neighbor
- zweitbester Algorithmus:
  - Bagging (10 Iterationen) mit SMO (Support Vector Machine die mit Sequential Minimal Optimization trainiert wird) – 97,7184 % Genauigkeit