Maschinelles Lernen Symbolische Ansätze: Projekt Aufgaben 7-8

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgabe 7 - Ensemble-Lernen
Benutzte Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- labor
- yeast
- car

Aufgabe 7 - Ensemble-Lernen
Datensatz labor

TECHNISCHE UNIVERSITÄT DARMSTADT

Anzahl Iterationen	1	2	3	4	5
J48	73.7%	/	/	/	/
AdaBoost	73.7%	77.2%	82.5%	80.7%	82.5%
Bagging	71.9%	78.9%	77.2%	80.7%	82.5%
Random Forests	84.2%	80.7%	84.2%	86.0%	87.7%

Anzahl Iterationen	6	7	8	9	10	1000
J48	/	/	/	/	/	/
AdaBoost	86.0%	84.2%	87.7%	87.7%	89.5%	87.7%
Bagging	82.5%	86.0%	86.0%	84.2%	86.0%	86.0%
Random Forests	87.7%	87.7%	87.7%	87.7%	87.7%	89.5%

Aufgabe 7 - Ensemble-Lernen
Datensatz yeast

TECHNISCHE UNIVERSITÄT DARMSTADT

Anzahl Iterationen	1	2	3	4	5
J48	56.0%	/	/	/	/
AdaBoost	56.0%	46.6%	53.8%	54.6%	56.3%
Bagging	50.3%	51.2%	55.3%	57.1%	56.9%
Random Forests	47.8%	50.1%	53.4%	54.0%	56.4%

Anzahl Iterationen	6	7	8	9	10	1000
J48	/	/	/	/	/	/
AdaBoost	54.3%	55.9%	55.4%	56.9%	56.4%	60.0%
Bagging	58.0%	59.0%	59.4%	59.6%	59.1%	62.1%
Random Forests	56.7%	57.7%	57.5%	58.8%	58.8%	62.1%

Aufgabe 7 - Ensemble-Lernen
Datensatz car

TECHNISCHE UNIVERSITÄT DARMSTADT

Anzahl Iterationen	1	2	3	4	5
J48	92.4%	/	/	/	/
AdaBoost	92.4%	92.4%	94.7%	93.1%	95.5%
Bagging	90.9%	91.6%	92.1%	92.5%	92.8%
Random Forests	83.6%	87.5%	90.0%	91.4%	91.4%

Anzahl Iterationen	6	7	8	9	10	1000
J48	/	/	/	/	/	/
AdaBoost	95.1%	96.1%	95.5%	96.2%	96.1%	97.2%
Bagging	92.4%	92.5%	92.9%	93.2%	93.1%	93.7%
Random Forests	91.9%	92.4%	92.6%	93.2%	93.6%	95.0%

Aufgabe 7 - Ensemble-Lernen
Vergleich und Interpretation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Für alle Methoden gilt allgemein, dass mehr Iterationen eine höhere Genauigkeit bedeuten
- Bei geringen Iterationszahlen kann es bei AdaBoost und Bagging zu großen Schwankungen kommen
- Random Forests wird mit mehr Bäumen immer genauer, aber nie schlechter
- Random Forests ist immer am schnellsten, Bagging am zweitschnellsten und AdaBoost am langsamsten
  - Random Forests ist zwischen 2x und 6x so schnell, je nach Datensatz
- Für hohe Iterationszahlen sind alle Methoden besser als J48
- Die erzielte Accuracy und Verarbeitungsgeschwindigkeit scheinen stark datenabhängig zu sein
  - Die Anzahl der Daten spielt dabei anscheinend eine untergeordnete Rolle
- Für schnelle Ergebnisse ist Random Forests am geeignetsten

Aufgabe 8 - Pre-Processing
Benutzte Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- autos
- iris
- sonar

Aufgabe 8 - Pre-Processing
Erzielte Genauigkeiten

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensatz	J48 Ursprünglich	J48 Diskretisiert
autos	Acc. 82%, Size 69	Acc. 84%, Size 103
iris	Acc. 96%, Size 9	Acc. 94%, Size 4
sonar	Acc. 71%, Size 35	Acc. 80%, Size 31

Aufgabe 8 - Pre-Processing
Vergleich und Interpretation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Genauigkeit
  - J48 auf den ursprünglichen Daten ist im Schnitt schlechter als J48 auf den diskretisierten Daten. Eine mögliche Erklärung wäre, dass die Daten nach dem Pre-Processing einfacher und bereits gruppiert sind, und dadurch leichter ein besseres, generalisiertes Modell gelernt werden kann.

Abschlussüberblick

TECHNISCHE UNIVERSITÄT DARMSTADT

FRAGEN?

Gruppenmitglieder

TECHNISCHE UNIVERSITÄT DARMSTADT

Joachim F. Brehmer-Moltmann, 1766932

Jeannine Endreß, 1669152