Maschinelles Lernen Symbolische Ansätze: Projekt Aufgaben 7-9

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgabe 7 - Ensemble-Lernen
- Benutzte Datensätze
- Regulärer J48, Bagging und AdaBoost
- Vergleich und Interpretation

Aufgabe 8 - Entdecken von Assoziationsregeln
- Apriori-Algorithmus
- Interessante Zusammenhänge

Aufgabe 9 - Pre-Processing
- Benutzte Datensätze
- Erzielte Genauigkeiten und Baumgrößen
- Vergleich und Interpretation

Aufgabe 7 - Ensemble-Lernen
Benutzte Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- Balance Scale Weight & Distance Database
- Breast Cancer Data
- Database for Fitting Contact Lenses
- Sonar, Mines vs. Rocks
- Zoo database

Aufgabe 7 - Ensemble-Lernen
Regulärer J48, Bagging und AdaBoost

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensatz	balance	breast	lenses	sonar	zoo
Regulärer J48	76.6%	75.5%	83.3%	71.2%	92.1%
Bagging J48	82.2%	73.4%	79.2%	74.5%	93.1%
AdaBoost J48	78.9%	69.6%	70.8%	77.9%	95.0%
Bagging RandomForest	82.4%	69.2%	70.8%	86.5%	93.1%
AdaBoost RandomForest	78.4%	66.4%	79.2%	82.2%	90.1%

Aufgabe 7 - Ensemble-Lernen
Vergleich und Interpretation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Als Ensemble-Methode liefert Bagging insgesamt die besten Ergebnisse
- Dabei ist keine klare Struktur erkennbar, ob J48 oder RandomForest der bessere Lernalgorithmus für Bagging ist
- Bei AdaBoost lässt sich ebenfalls nicht definitiv festellen, ob J48 oder RandomForest besser geeignet wäre
- Auch wenn der reguläre J48 nicht überall schlechter ist, scheint insgesamt die Benutzung einer Ensemble-Methode sinnvoll zu sein, um bessere Genauigkeiten zu erzielen
- Alles in allem sieht es aber so aus, dass die erzielte Accuracy der einzelnen Algorithmen stark datenabhängig ist

Aufgabe 8 - Entdecken von Assoziationsregeln Apriori-Algorithmus

TECHNISCHE UNIVERSITÄT DARMSTADT

- Zunächst numerische Attribute “fnlwgt” und “education-num” entfernen
- Algorithmus Optionen: “car=true” (nur Regeln für die Klassenvariable lernen)

$\rightarrow$ Mehrere Durchläufe des Apriori Regellerners jeweils mit unterschiedlichen Attributmengen und anschließende Analyse der Ergebnisse

Aufgabe 8 - Entdecken von Assoziationsregeln Interessante Zusammenhänge

TECHNISCHE UNIVERSITÄT DARMSTADT

Wichtigste gefundene Regel basierend auf allen symbolischen Attributen:

- marital-status=Never-married ==> class=<=50K conf:(0.95)

Da alle Regeln auf die Condition “marital-status” basieren, wird diese entfernt:

- relationship=Not-in-family capitalgain=0 ==> class=<=50K conf:(0.92)
- sex=Female capitalgain=0 capitalloss=0 ==> class=<=50K conf:(0.92)
- workclass=Private capitalloss=0 ==> class=<=50K conf:(0.91)

Als nächstes “capitalgain” und “capitalloss” aufgrund von stark ungleicher Verteilung entfernen:

- relationship=Own-child ==> class=<=50K conf:(0.99)
- age=0 ==> class=<=50K conf:(0.98)
- native-country=United-States ==> class=<=50K conf:(0.98)

Aufgabe 8 - Entdecken von Assoziationsregeln
Interessante Zusammenhänge

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgrund von schlechter Verteilung Attribut “native-country” entfernen:

- relationship=Own-child race=White ==> class=<=50K conf:(0.98)
- age=0 race=White ==> class=<=50K conf:(0.98)
- age=0 sex=Male ==> class=<=50K conf:(0.98)

Zum Schluss noch “age” und “relationship” entfernen, da die letzten Regeln alle darauf basieren:

- workclass=Private sex=Female ==> class=<=50K conf:(0.91)
- hoursperweek=1 ==> class=<=50K conf:(0.9)

$\rightarrow$ Es lassen sich durchaus “intuitive” bzw. erwartete Zusammenhänge zwischen Geringverdiener und bestimmten gesellschaftlichen Schichten wie Frauen, Junge, Eltern oder Privatangestellte feststellen, die z.T. auch so in der realen Welt exisitieren können.

Aufgabe 9 - Pre-Processing
Benutzte Datensätze

TECHNISCHE UNIVERSITÄT DARMSTADT

- 1985 Auto Imports Database
- Iris Plants Database
- Sonar, Mines vs. Rocks

Aufgabe 9 - Pre-Processing
Erzielte Genauigkeiten und Baumgrößen

TECHNISCHE UNIVERSITÄT DARMSTADT

Datensatz	J48 Ursprünglich	J48 Diskretisiert	Filtered Classifier
autos	Acc. 82%, Size 69	Acc. 84%, Size 103	Acc. 73%, Size 103
iris	Acc. 96%, Size 9	Acc. 94%, Size 4	Acc. 93%, Size 4
sonar	Acc. 71%, Size 35	Acc. 80%, Size 31	Acc. 75%, Size 31

Aufgabe 9 - Pre-Processing
Vergleich und Interpretation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Genauigkeit
  - J48 auf diskretisierten Daten ist im Schnitt besser als J48 auf den ursprünglichen Daten. Eine mögliche Erklärung wäre, dass die Daten nach dem Pre-Processing bereits einfacher und gruppiert sind, und dadurch leichter ein besseres generalisiertes Modell gelernt werden kann.
  - Dagegen hat FilteredClassifier fast immer eine schlechtere Accuracy. Dies könnte daran liegen, dass die Kombination von Pre-Processing und Lern-Algorithmus zu einem einzigen Klassifizierer dazu führt, dass in einem gemeinsamen Schritt sowohl diskretisiert als auch klassifiziert werden muss und dabei die beiden Aufgaben nur mit gewissen Abstrichen zusammen kombiniert werden können.
- Baumgröße
  - J48 auf diskretisierten Daten und der FilteredClassifier liefern jeweils immer einen gleich großen Baum.
  - Ansonsten ist kein eindeutiges Ergebnis im Bezug zu J48 auf den ursprünglichen Daten erkennbar (resultierender Baum mal größer und mal kleiner).

Abschlussüberblick

TECHNISCHE UNIVERSITÄT DARMSTADT

Aufgabe 7 - Ensemble-Lernen
- Benutzte Datensätze
- Regulärer J48, Bagging und AdaBoost
- Vergleich und Interpretation

Aufgabe 8 - Entdecken von Assoziationsregeln
- Apriori-Algorithmus
- Interessante Zusammenhänge

Aufgabe 9 - Pre-Processing
- Benutzte Datensätze
- Erzielte Genauigkeiten und Baumgrößen
- Vergleich und Interpretation

FRAGEN?

Gruppenmitglieder

TECHNISCHE UNIVERSITÄT DARMSTADT

Matthias Krebs, 1620340

Thomas Pignede, 1626386

Svenja Stark, 1658147