Adatbányászati technikák
1. kisHF

(kiadás: 2012. március 8.)

Adatbányászati technikák
I. kisHF

- Töltse le a housing_labeled.arff és housing_unlabeled.arff fájlokat innen: http://www.cs.bme.hu/~buza/edu/dm_techn/kisHF1.html
- Ezek a fájlok különböző, ingatlanokra vonatkozó adatokat tartalmaznak (pl. az adott városrészben érvényes ingatlanadó nagyságát, bűnözés mértékét, autópálya-hálózat elérhetőségét, diák-tanár arányt... , lásd részletesebben: utolsó fólia).
- A housing_labeled.arff esetében ismert az ingatlanok árának mediánja az adott városrészben, a housing_unlabeled.arff esetében nem.
- Az árak 10 diszkrét kategória valamelyikébe tartoznak.

Adatbányászati technikák
I. kisHF

- **1. Részfeladat (6 pont)**
  - Nyissa meg WEKA-ban a housing_labeled.arff fájlt!
  - Készítsen J48 algoritmussal döntési fát, amely képes megbecsülni a megadott adatok alapján, az árat!
  - A döntési fa készítése előtt, a döntési fa paramétereinek beállításánál, kapcsolja be a ReducedErrorPruning opciót, és az egy levélre eső példányok minimális számát (minNumObj) állítsa 5-re!
  - Rajzolja ki Weka-val az elkészített döntési fát és adja meg a döntési fa confusion mátrix-át! A test options mezőben melyiket lehetőséget kell kiválasztani, hogy a confusion mátrix-beli értékek becslése megbízható (fair) legyen?

Adatbányászati technikák
I. kisHF

- **2. Részfeladat (4 pont) - programozás**
  - Készítsen egy JAVA programot, amely a WEKA API-n keresztül (a WEKA-t függvénykönyvtárként használva) osztályozza a housing_unlabeled.arff fájlban lévő példányokat!
  - Osztályozó algoritmusként használja ismét a J48-t az előbbi paraméterekkel!
  - Írassa ki a standard outputra, hogy melyik példányt hogyan osztályozta!
    Legyen azonosítható, hogy melyik osztálycímke melyik példányhoz tartozik, azaz ne csak az osztályozó algoritmus által adott osztálycímkét írassa ki, hanem a példányt is, vagy legalább annak sorszámát!

Adatbányászati technikák - I. kisHF

- 1. részfeladat megoldásaként beadandó:
  - az elkészült döntési fa ábrája (screen shot),
  - a confusion mátrix,
  - a válasz arra a kérdésre, hogy a test options közül mely lehetőséget választotta
- 2. részfeladat megoldásaként beadandó:
  - JAVA forráskód
  - Az, hogy a döntési fa mely osztályba sorolja be az ismeretlen címkéjű példányokat
- **Dolgozzon igényesen**, különben: pontlevonás! (Pl. döntési fa ábráját megfelelően nagyítsa ki, hogy látszódjon, mi van az egyes csomópontokban, stb.)

Változók jelentése

- 1. CRIM per capita crime rate by town
- 2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- 3. INDUS proportion of non-retail business acres per town
- 4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- 5. NOX nitric oxides concentration (parts per 10 million)
- 6. RM average number of rooms per dwelling
- 7. AGE proportion of owner-occupied units built prior to 1940
- 8. DIS weighted distances to five Boston employment centres
- 9. RAD index of accessibility to radial highways
- 10. TAX full-value property-tax rate per $10,000
- 11. PTRATIO pupil-teacher ratio by town
- 12. B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- 13. LSTAT % lower status of the population
- 14. MEDV Median value of owner-occupied homes in $1000's