Adatbányászati technikák
2. kisHF

(kiadás: 2012. április 5.)

Adatbányászati technikák
2. kisHF

- Töltse le a wine.arff fájlt innen:
  http://repository.seasr.org/Datasets/UCI/arff/wine.arff
- Ezen a fájlok borokra vonatkozó adatokat tartalmaz.
- Jelenítse meg a fájl tartalmát pl. Windows Commander-ben F3-mal vagy Parancssorban a
  type wine.arff | more
  paraccsal és tekintse meg a fájl elején komment sorokban megadott információkat arról, hogy milyen adatokat tartalmaz a fájl.

Adatbányászati technikák
2. kisHF

- **1. részfeladat (6 pont)**
  - Nyissa meg WEKA-ban a fájl-t! Az Explorer felület Preprocess fülén állítsa be, hogy az első attribútum az osztályattribútum (a képernyő bal oldalán, kb. középen).
  - A Cluster fülön klaszterezze az adatbázist:
    - Állítsa a cluster mode-t „classes to cluster evaluation”-re, és válassza ki az osztályattribútumot (class)
    - Klaszterezze az adatbázist k-Means (SimpleKMeans) és Hierarchical Clusterer eljárásokkal

Adatbányászati technikák
2. kisHF

- **1. részfeladat (6 pont, folytatás)**
  - A hierarchikus klaszterező mind három tanult eljárását próbálja ki (linkType-t állítsa SINGLE-re, COMPLETE-ra, ill. AVERAGE-ra)
  - A klaszterek számát állítsa 3-ra illetve 5-re.
  - Adja meg hogy fentiek szerinti összesen 8 kipróbált eljárás (k-Means és hierarchikus mindhárom változata, 3 illetve 5 klaszter keresése mellett) közül melyik eredményezi a legjobb klaszterezést (incorrectly clustered instances alapján).

Adatbányászati technikák
2. kisHF

- **2. Részfeladat (4 pont) - programozás**
  - Készítsen egy JAVA programot, amely a WEKA API-n keresztül (a WEKA-t függvénykönyvtárként használva) megoldja az 1. részfeladatot, majd kiírja a standard outputra (konzol-ra), hogy melyik klaszterező algoritmus klaszterezi legjobban az adatbázist
- **Bónuszpontért (+1 pont):**
  - Ebben a feladatban a klaszterezés „jóságát” az alapján mértük, hogy a klaszterező által talált csoportok mennyire hasonlítanak az adatbázisbeli, emberi szakértő által megadott csoportokhoz (osztályokhoz). Lehet-e más módon mérni a klaszterezés jóságát?

Adatbányászati technikák - 2. kisHF

- 1. részfeladat megoldásaként beadandó:
  - A 8 darab kipróbált klaszterező eljárás kimenete (Clusterer output mezőben megejelenő szöveg),
  - a válasz arra a kérdésre, hogy melyik klaszterező a legjobb
- 2. részfeladat megoldásaként beadandó:
  - JAVA forráskód
- **Dolgozzon igényesen**, különben: pontlevonás!