

Adatbányászat labor

2011-2012 őszi félév

Tantárgyi követelmények

- 6 gyakorlati óra
 - minden óra végén e-mailben feladatok elküldése
 - 1 alkalommal lehet hiányozni
- 1 nagyházifeladat
 - Előzetes bemutató: ápr. 24. (5 perc)
 - Beadás: máj 8.
- kisZH (kb. 50 perc) – április 10/17.
- Jegy: $5 \cdot 10 + 20(\text{kisZH}) + 30(\text{nagyHF})$ pont
- Gyakorlati órák:
 - kéthetente: 4x45 perc
 - > szünet? kezdés?

Nap	Esemény
02/07	1. Labor (A + B)
02/14	2. Labor - A
02/21	2. Labor - B
02/28	3. Labor – A
03/06	3. Labor - B
03/13	4. Labor – A nagyHF kiadás
03/20	4. Labor – B nagyHF kiadás
03/27	5. Labor - A
04/03	5. Labor - B
04/10	KisZH (50 perc) 6. Labor - A
04/17	KisZH (50 perc) 6. Labor - B
04/24	nagyHF előzetes
05/01	Ünnep
05/08	NagyHF beadás KisZH pótlás

Elérhetőségek

Tantárgy honlapja:

http://dms.sztaki.hu/~daroczyb/adatb_bsc.php

A csoport:

Daróczy Bálint

Email: daroczyb@ilab.sztaki.hu

Személyesen: MTA SZTAKI Lágymányosi u. 11

B csoport:

Buza Krisztián

Email: buza@cs.bme.hu

Személyesen: BME I. épület, I.E. 217.3.

Honlap: <http://www.cs.bme.hu/~buza>

Röviden a tematikáról

- Klasszifikációs feladatok
- Klaszterezés
- Feature kiválasztás
- Kernel és egyéb trükkök

Mindez gyakorlati szempontból, alapvetően Weka és scriptek (és programozási nyelvek) segítségével.

Kicsit bővebben

Adatbányászat alapfeladatai (példák, megoldandó problémák stb.)

Adatreprezentációk (formátumok, ritka mátrix, pontosság)

Normalizáció

Klasszifikálási problémák: döntési fák, fa alapú algoritmusok

Klasszifikálás kiértékelése: prec, recall, f-m, roc, map

Klasszifikálási problémák: perceptron, logisztikus regresszió
(sztochasztikus gradiens módszer), dual és primal forma

Ensemble módszerek

Klaszterezés: K-means, DBSCAN, OPTICS, GMM, X-means

Bag-of-Words modellezés, Fisher információ, SuperVector

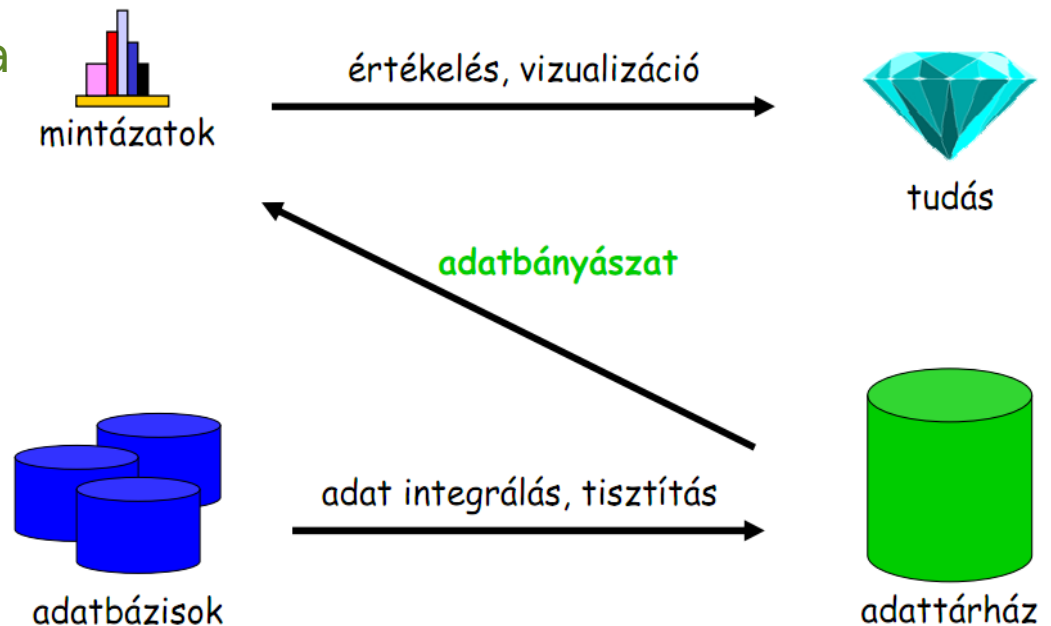
Attribútum és feature szelekció, PCA

Adatbányászat feladatai

Data Mining: érdekes avagy nem triviális, implicit, eddig ismeretlen és feltételezhetően hasznos információk vagy mintázatok kinyerése nagy adatbázisokból

Knowledge discovery in databases (KDD)

- ismeretszerzés adatbázisokból
- adat/mintázat vizsgálata
- összefüggések azonosítása
- stb.



adatbányászat: adatvezérelt mintázatkinyerés

Adatbányászat feladatai

Mivel az adatbányászat egy adott adattárház struktúráját feltételez (legyen az akár csak egy text fájl) s felhasznál statisztikai modelleket, az adatbányászati feladatok általában nem pusztán a már meglévő struktúrára épített elemek, hanem a teljes lánc felépítését is meghatározzák

pl. egy már adott TB+ méretű adatbázisban tárolt elemeken végzett hatékony tanítás szinte lehetetlen, ám bár elképzelhető, hogy más reprezentációban (pl. ritka mátrix vagy egy Bag-of-words modellben) már hatékonyan elvégezhető

Adatbányászat feladatai

Web és egyéb hálózat:

- keresés (szöveg,kép,videó,honlap,információ)
- ajánlás (reklám,cikk,kép...)
- log elemzés (telefonhálózat, ügyfélhálózat stb.)



Üzleti intelligencia:

- döntés segítő rendszerek (tőzsde,biztosítás)
- összefüggés keresés (hasonló ügyfél, hasonló megoldás)

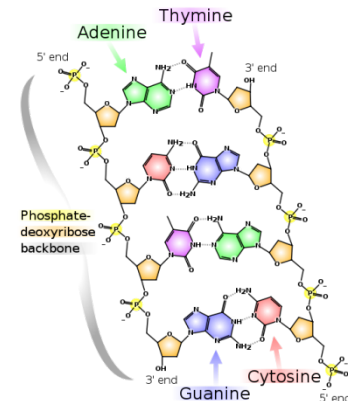


Mutlimédia:

- automatikus annotáció (kép,videó,zene stb.)
- hangfelismerés és beszédfelismerés

Orvosi kutatás és diagnosztika:

- szűrés (rák, TBC stb)
- élettani folyamatok elemzése
- génkutatás (génszekvenciák keresése, azonosítása, modellezése)
- gyógyszerkutatás



Adatbányászati szoftverek

Weka - open source java

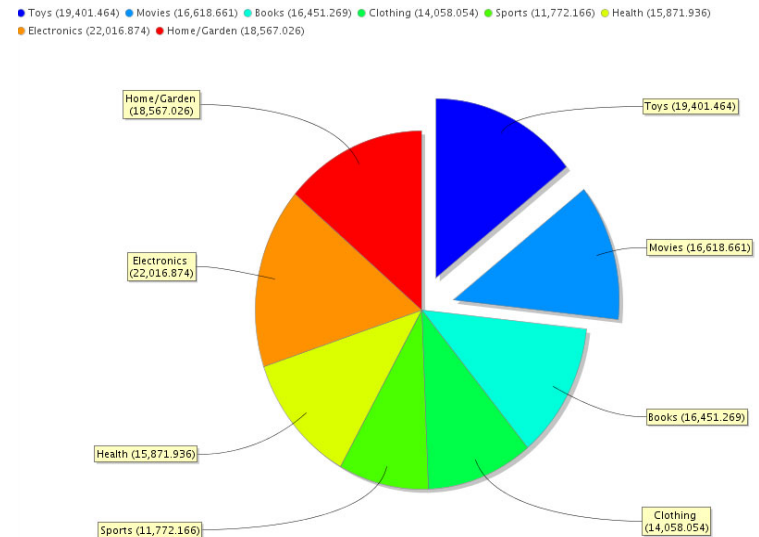
RapidMiner

Clementine (SPSS)

Darwin (Oracle)

SAS Enterprise Miner

IBM Intelligent Miner



Mielőtt ...

Adatok formázásához:

file másolása unix alatt (Windows alatt cygwin)

```
cp vmit valahova
```

file megnézése:

```
less valami
```

első oszlop kivágása:

```
cut -d' ' -f1,2 adatfajl
```

rendezés numerikusan:

```
cut -d' ' -f1 adatfajl | sort -n
```

azonos sorok eldobása:

```
cut -d' ' -f1 adatfajl | sort -n | uniq
```

Mielőtt ...

sorok megszámlálása:

```
cut -d' ' -f1 adatifajl | sort -n | uniq | wc -l
```

első oszlopbeli értékek megszámlálása:

```
cut -d' ' -f1 adatifajl | sort -n | uniq -c
```

ugyanaz awk-val:

```
awk '{v[$1]++}END{for(i in v) print i,v[i]}' adatifajl
```

első oszlop átlaga:

```
awk '{s+=$1}END{print s/NR;}' adatifajl
```

Mielőtt...

második oszlopbeli értékek átlaga első oszlop szerinti kulcs szerint

```
awk '{v[$1]+=$2;c[$1]++}END{for (i in v) print i,  
v[i]/c[i]}' adatfajl
```

file-ba írás echo-val

```
echo "AKARMI" > output
```

```
echo "MEG VALAMI" >> output
```

egy adott karakter kicserélése:

```
tr ',' '.' < input > output
```

két file konkatenálása:

```
cat file1 file2 > output
```

Egyéb hasznos eszközök:

du,df ,ls,paste,for,chmod,mc,sed,for,seq stb.

Arff fájl formátum

% comment

@RELATION <az adat azonosítója>

@ATTRIBUTE <feature neve> <típusa>

(NUMERIC / DATE /STRING vagy NOMINAL : pl. {-2,1,2,3})

@ATTRIBUTE class <típus>

(Csak nominális vagy numerikus lehet)

@DATA

<érték1>,<érték2> ...,<osztály>

Órai feladat 1:

http://www.ilab.sztaki.hu/~daroczyb/adatb_msc.php -> adatb1.zip

Formázzuk át arff-be a contact-lenses fájlt

WEKA

- „Weka 3: Data Mining Software in Java”
- Szabadon letölthető (GNU licenz):
<http://www.cs.waikato.ac.nz/ml/weka/>

Waikatoi egyetemen fejlesztik

- Classification
- Clustering
- Feature selection
- Association rules
- Some visualisation
- cvs és arff formátumok



Weka indítása

- `java -jar weka.jar`
- explorer felület
- Adat: próbáljuk meg betölteni az arff-be formázott fájlt!
- Classify -> mindenki válasszon egy típust

ZeroR

J48

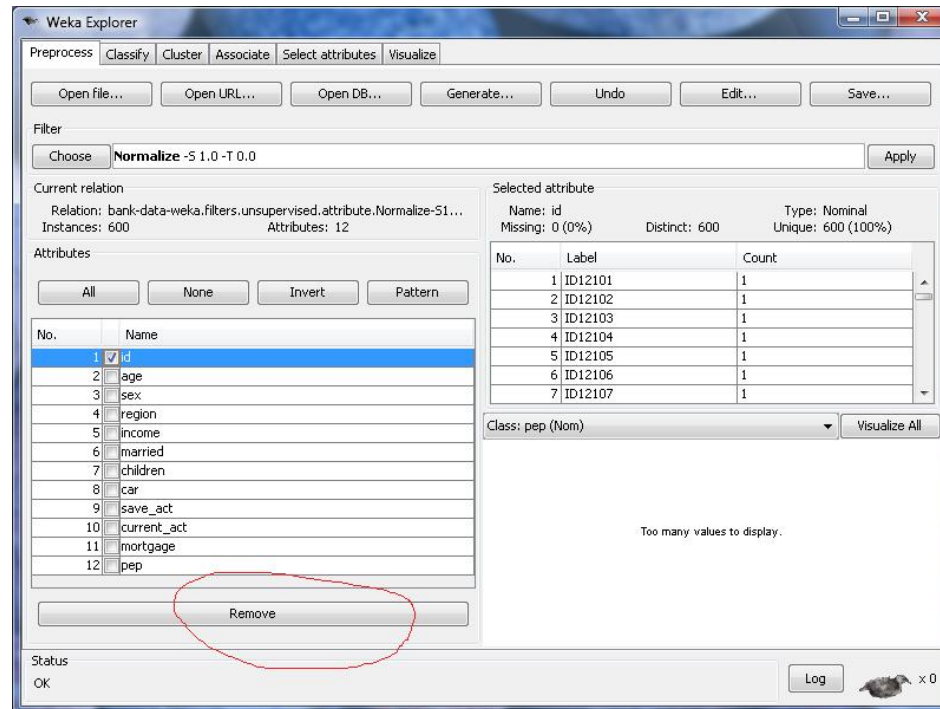
Bayes

stb...

Töltsük be a bank-data.arff-et!

Előfeldolgozás

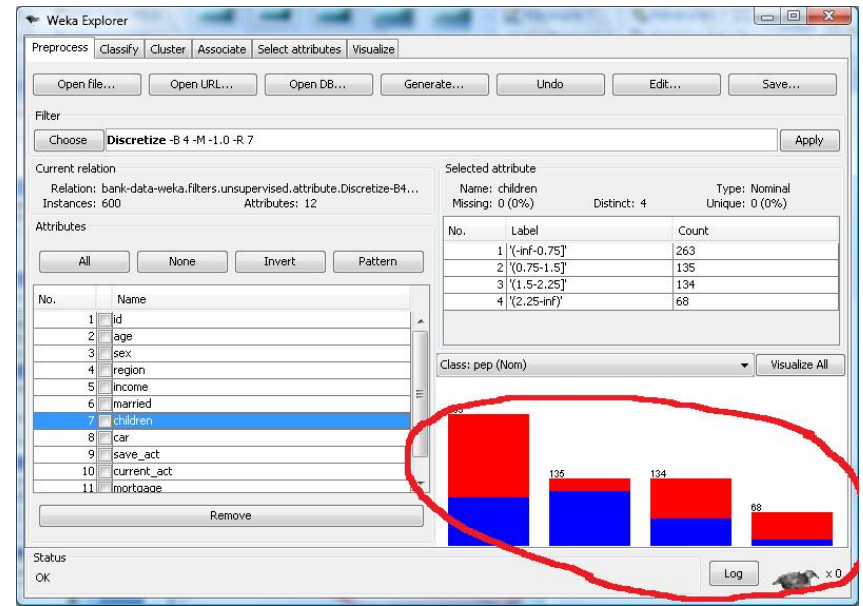
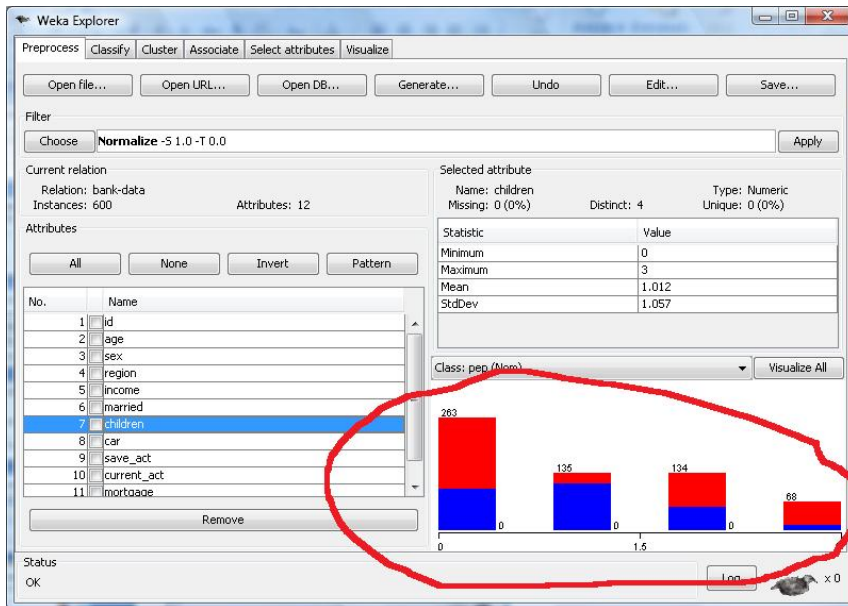
- Néha eldobunk pár változót
remove



Előfeldolgozás

- Egyes algoritmusok csak nominális változókon értelmezettek

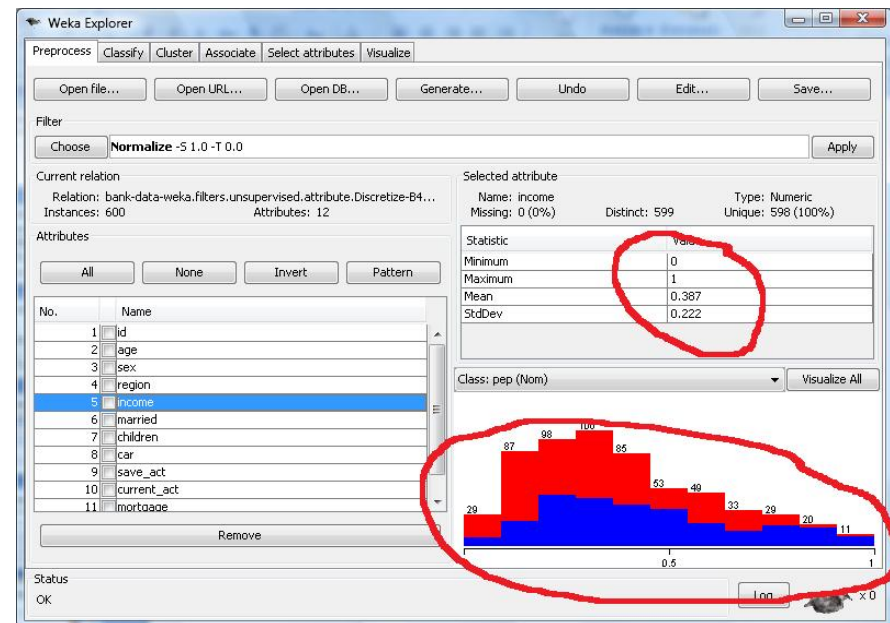
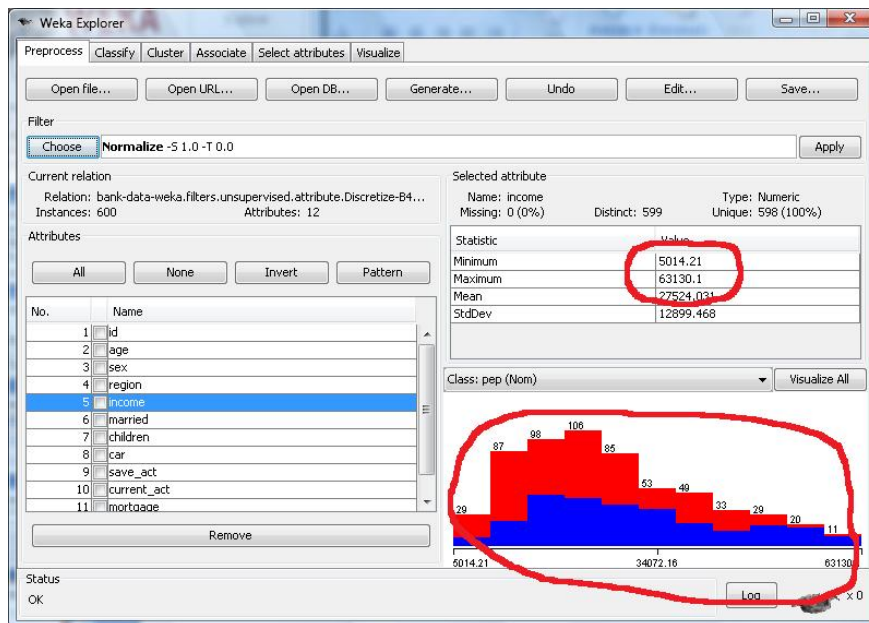
Preprocess -> filter -> discretize



Előfeldolgozás

- Normalizálás

preprocess -> filter -> normalize



Apriori kipróbálása

bank-data3.arff

- csak diszkrétizált változók
- összekapcsolási szabályok megtalálása
- PEP -> egy banki termék, melyről tudjuk az ügyfél megvette-e

Apriori

Órai feladat 2:

Bank-data:

- keressük az ügyfelek azon tulajdonságait, melyek gyakran egyszerre fordulnak elő
- ebben az esetben a tranzakciók az ügyfelek a keresett elemhalmazok pedig a feature-párok
- pl. a nők és középkorú ügyfelek gyakrabban rendelkeznek az adott termékkel (PEP) mint a vagy nem nők és középkorú vagy nők de nem középkorú ügyfelek

bank-data3.arff -> associate -> apriori

Normalizáció fontossága

Órai feladat 3:

hist_norm és minta_hist:

Az images könyvtár képein számolt 3×8 dimenziós RGB hisztogramjai

Rendezzük sorrendbe távolság szerint a minta_hist fájl-ban található kép hisztogramjától számított L2 távolság alapján!

Mi lehet a hiba? Mi történik ha L2 normalizáljuk a hisztogrammokat?

