WEKA gyakorlat 1.

Bevezetés

Tartalomjegyzék

[1 WEKA alapfogalmak 3](#_Toc350375811)

[1.1 Weka adatábrázolás 3](#_Toc350375812)

[1.2 Weka műveletek 3](#_Toc350375813)

[1.3 Weka felületek 4](#_Toc350375814)

[1.4 Beállítások 4](#_Toc350375815)

[2 Adat beolvasása 5](#_Toc350375816)

[2.1 Osztálycímkétől független szűrők 5](#_Toc350375817)

[2.1.1 Adatok hozzáadása 5](#_Toc350375818)

[2.1.2 Hiányzó értékek helyettesítése (ReplaceMissingValues) 5](#_Toc350375819)

[2.1.3 Egyéb filterek 6](#_Toc350375820)

[2.2 Példa egy szűrő hatásának vizsgálatára 6](#_Toc350375821)

[2.3 Osztálycímkétől függő szűrők 9](#_Toc350375822)

# WEKA alapfogalmak

A [Weka](http://www.cs.waikato.ac.nz/~ml/weka/) egy gépi-tanuló algoritmusokat és a hozzájuk kapcsolódó funkciókat (adatkezelés, megjelenítés, stb.) megvalósító programcsomag, Java-ban íródott szabad szoftver.

## Weka adatábrázolás

A Weka minden adathalmazt mátrixként ábrázol, mely az objektumaink azonosítója és jellemzői mellett tartalmazhatja az osztálycímké(ke)t is.

* **dataset** = mátrix
  + lehet teljesen (sűrű) vagy részlegesen kitöltött (ritka)
  + ARFF fájl (sűrű tárolás)
    - fejléc (header) és adat (data)
  + sparse ARFF fájl (ritka tárolás)
    - 0 értékek nincsenek rögzítve
* **attribute** = tulajdonság (dimenzió/jellemző/"oszlopvektor")
  + név is adható neki, de pozíciójával hivatkozzuk (first=1,2,...,n=last)
  + értékkészlete kötött, pl.:
    - **numeric**: szám (egész és lebegőpontos is)
    - **nominal**: halmaz elemei, pl. {yes, no}
    - **string**: tetszőleges szöveg
    - **date**: dátum
  + ha egy funkció igényel osztálycímkét (**class attribute**), akkor az alapértelmezésben az utolsó (last) tulajdonság
* **instance** = vektor ("sorvektor")
  + egy objektum tulajdonságait írja le
  + az ARFF fájl egy sora

## Weka műveletek

A Weka segítségével módosíthatjuk az adathalmazunkat; jellemzőkiválasztási, csoportosítási, vagy osztályozási módszereket alkalmazhatunk.

* **filter** = módosító művelet
  + sorokat vagy oszlopokat módosít (instance vs attribute)
  + osztálycímkét igényel vagy sem (supervised vs unsupervised)
* **classifier** = osztályozó
  + tanításhoz osztálycímkét igényel, az összes többi jellemzőt felhasználja
  + a tanítás eredménye modellfájlba menthető későbbi alkalmazásra

## Weka felületek

A Weka programozói és több felhasználói felülettel is használható.

* GUI
  + Experimenter (dialógus alapú)
  + KnowledgeFlow (drag-n-drop)
* CLI (konzol)
* Java

A kisebb feladatokhoz a szemléletesebb, grafikus módszert érdemes választani. A nagyobb adatállományokat feldolgozó és nagyobb számítási igényű feladatokhoz a Java programba ágyazott módszert célszerű használni.

## Beállítások

Tipikus probléma a Java-s programok esetén az induláskor rögzített maximális rendelkezésre álló memóriaméret (heap size), mely a munkavégzés közepén Java OutOfMemoryException-nel kísért összeomlásban végződik. A Weka sem kivétel, így célszerű a rendelkezésre álló memória méretét megnövelni, pl.:

* Windows: a C:\Progra~1\Weka-3-6\RunWeka.ini fájl 'maxheap' paraméterét megnövelni "1400m"-re.
  + CLASSPATH beállítása: add $PATH$\weka.jar
* Linux: Weka indítása: weka -m 1400m

# Adat beolvasása

Hajtsa végre a következő lépéseket:

* Explorer indítása
* Preprocess fülön Openfile-al fájl megnyitása
  + Keressük meg az adatfájlok helyét, a Weka-3-6\data könyvtárt.
* Nyissuk pl. meg a soybean adathalmazt

Előfeldolgozásban két féle szűrő

* unsupervised – nem használ fel osztályadatokat
* supervised – felhasznál osztályadatokat

## Osztálycímkétől független szűrők

### Adatok hozzáadása

Új (üres) attribútum hozzáadása

* weka.filters.unsupervised.attribute.Add

Új attribútumérték hozzáadása egy attribútumhoz

* weka.filters.unsupervised.attribute.AddValues

Új attribútum hozzáadása többinek kombinációjaként

* weka.filters.unsupervised.attribute.AddExperssion

Zaj hozzáadása

* weka.filters.unsupervised.attribute.AddNoise

Azonosító hozzáadása

* weka.filters.unsupervised.attribute.AddID

### Hiányzó értékek helyettesítése (ReplaceMissingValues)

Keressük meg a következő szűrőfüggvényt

weka.filters.unsupervised.attribute.ReplaceMissingValues

**Leírás**

*Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.*

**Opciók**

* ignoreClass -- The class index will be unset temporarily before the filter is applied.

**Használat**

* Szerkesztés: Choose utáni mezőre kattintás
* Nézzük meg a More és a Capabilities gombokra mit jelenít meg a Weka
* osztálycímkét vegyük ki a futtatandók közül
* Futtassuk a szűrőt (Apply gomb)

**Feladatok**:

* Keressünk olyan adatsort, ahol hiányzó numerikus értékek vannak (labor)!
* Nézzük meg, hogy hogyan változik az attribútum átlaga és a szórása?
* Hogyan tárolja a hiányzó értékeket Weka az arff fájlban?

### Egyéb filterek

Dátum transzformálása más formátumúra

* weka.filters.unsupervised.attribute.ChangeDateFormat

Eltolás – középérték nullára állítása

* weka.filters.unsupervised.attribute.Center

.

## Példa egy szűrő hatásának vizsgálatára

Két érték összevonása:

weka.filters.unsupervised.attribute.MergeTwoValues

**Leírás**

Merges two values of a nominal attribute into one value.

**Opciók**

* attributeIndex -- Sets which attribute to process. This attribute must be nominal ("first" and "last" are valid values)
* firstValueIndex -- Sets the first value to be merged. ("first" and "last" are valid values)
* secondValueIndex -- Sets the second value to be merged. ("first" and "last" are valid values)

**Feladat**

* Töltsük be a credit-g adatsort.
* Futtassunk rá egy döntési fa osztályozó algoritmust (J48)
* Nézzük meg az eredményeket és értékeljük ki.

Correctly Classified Instances 733 73.3 %

Incorrectly Classified Instances 267 26.7 %

Kappa statistic 0.3264

Mean absolute error 0.3293

Root mean squared error 0.4579

Relative absolute error 78.3705 %

Root relative squared error 99.914 %

Total Number of Instances 1000

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.851 0.543 0.785 0.851 0.817 0.685 good

0.457 0.149 0.568 0.457 0.506 0.685 bad

Weighted Avg. 0.733 0.425 0.72 0.733 0.724 0.685

=== Confusion Matrix ===

a b <-- classified as

596 104 | a = good

163 137 | b = bad

2. Feladat

* new\_car / used\_car összevonása

Correctly Classified Instances 734 73.4 %

Incorrectly Classified Instances 266 26.6 %

=== Confusion Matrix ===

a b <-- classified as

595 105 | a = good

161 139 | b = bad

3. Feladat

* new\_car / used\_car összevonása
* furniture/equipment / domestic appliance összevonása

Correctly Classified Instances 732 73.2 %

Incorrectly Classified Instances 268 26.8 %

=== Confusion Matrix ===

a b <-- classified as

598 102 | a = good

166 134 | b = bad

4. Feladat

* new\_car / used\_car összevonása
* furniture/equipment / domestic appliance / radio\_tv összevonása

Correctly Classified Instances 738 73.8 %

Incorrectly Classified Instances 262 26.2 %

=== Confusion Matrix ===

a b <-- classified as

597 103 | a = good

159 141 | b = bad

Sikerül-e valamilyen összevonással további javulást elérni?

(749 elérhető)

## Osztálycímkétől függő szűrők

Diszkretizálás – az intervallum típusú attribútumok felosztása az osztálycímkétől függően

* weka.filters.supervised.attribute.Discretize

Binárissá konvertálás – minden kategorikus értékhez egy-egy új attribútumot definiál.

* weka.filters.supervised.attribute.NominalToBinary

Mintavételezés – példányokon való szűrés, úgy, hogy megmaradjon az eredeti eloszlása az osztálycímkéknek (stratified), bias a normális eloszlás felé megadható.

* weka.filters.supervised.instance.Resample