Adatok előfeldolgozása

Kulcsár Géza

Budapest, 2012. február 23.

Miért van szükség előfeldolgozásra?

Az adatbányászat a nagy mennyiségű adatokban rejlő információk feltárása.

Milyen formában áll az adat rendelkezésre?

Általában:

Miért van szükség előfeldolgozásra?

Az adatbányászat a nagy mennyiségű adatokban rejlő információk feltárása.

Milyen formában áll az adat rendelkezésre?

Általában:

- Hiányos

Miért van szükség előfeldolgozásra?

Az adatbányászat a nagy mennyiségű adatokban rejlő információk feltárása.

Milyen formában áll az adat rendelkezésre?

Általában:

- Hiányos
- Zajos

Miért van szükség előfeldolgozásra?

Az adatbányászat a nagy mennyiségű adatokban rejlő információk feltárása.
Milyen formában áll az adat rendelkezésre?
Általában:

- Hiányos
- Zajos
- Inkonzisztens

Hogyan előfeldolgozzunk?

- Milyen az adat? $\rightarrow$ leírás, statisztika
- Hiányos, zajos, inkonzisztens $\rightarrow$ adattisztítás
- Több adatforrás $\rightarrow$ integráció
- Nagy adatmennyiség $\rightarrow$ redukció

Az adat leírása, jellemzői

Attribútum típusok

- I. Kategória típusú attribútumok: csak az egyenlőség vizsgálható
- II. Sorrend típusú attribútumok: $>, <, =$ eldönthető, azaz van teljes rendezés
- III. Intervallum típusú attribútumok: az elemek csoportot alkotnak
- IV. Arány skálájú attribútum: van zérus elem is

Az adat leírása, jellemzői

Attribútum típusok

- I. Kategória típusú attribútumok: csak az egyenlőség vizsgálható
- II. Sorrend típusú attribútumok: $>, <, =$ eldönthető, azaz van teljes rendezés
- III. Intervallum típusú attribútumok: az elemek csoportot alkotnak
- IV. Arány skálájú attribútum: van zérus elem is

Az adat leírása, jellemzői

Középértékek

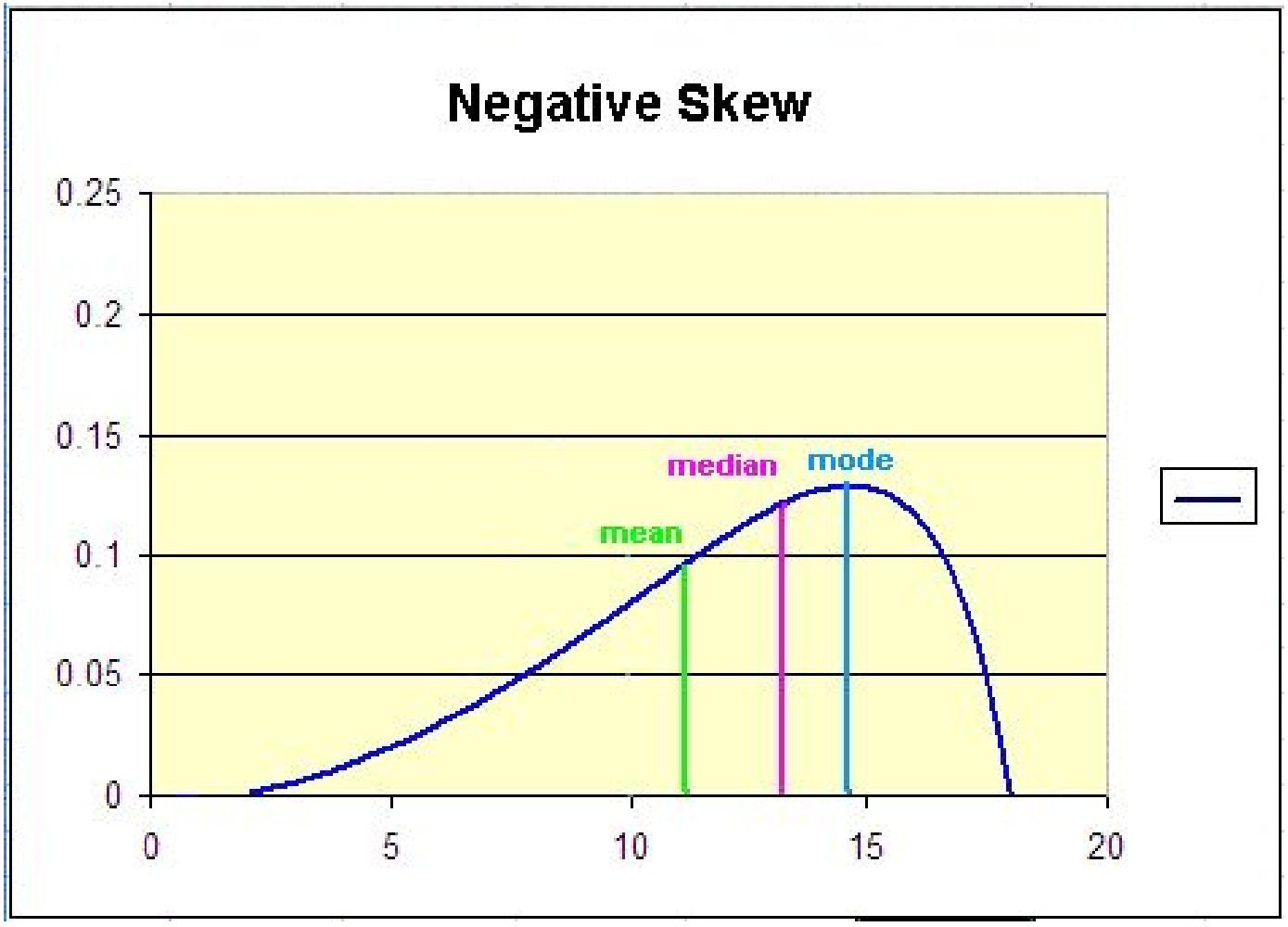
- Átlag (súlyozott átlag)

- Medián
  Nehezen számolható (nem lehet darabolni a feladatot, *holisztikus* mérték.)
  De ismert intervallumszámosságoknál jól becsülhető! Hogyan?

- Módusz
  Lehet több is

- Ferdeség (skewness): $\gamma_1 = \frac{E[(X-m)^3]}{(E[(X-m)^2])^{\frac{3}{2}}}$

Példa negatív ferdeségű adatra



Az adat leírása, jellemzői

Az értékek eloszlása

Többet tudhatunk meg az adathalmazról, ha nem csak a centralitásáról tájékozódunk.

- Negyedelőpontok (számosság szerint):
  Q_1: 25%-hoz, Q_3: 75%-hoz legközelebb eső adatpont a rendezett sorban

- $IQR = Q_3 - Q_1$
  Mit nem tudunk még?

Az adat leírása, jellemzői

Az értékek eloszlása

Többet tudhatunk meg az adathalmazról, ha nem csak a centralitásáról tájékozódunk.

- Negyedelőpontok (számosság szerint):
  Q_1: 25%-hoz, Q_3: 75%-hoz legközelebb eső adatpont a rendezett sorban

- $IQR = Q_3 - Q_1$
  Mit nem tudunk még?

- *Ötszámos jellemzés:* Minimum, Q_1, Medián, Q_3, Maximum

- Egyszerű lehetőség outlierek (különcök) azonosítására:
  $d(x, Q) > 1,5IQR$

Az adat leírása, jellemzői

Szórás

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

- σ^2: szórásnégyzet, variancia (tkp. második centrális momentum)
- σ: szórás
- Könnyen, elosztottan számítható, új adat érkezésekor felhasználhatjuk az eddigi értéket

Az adat leírása, jellemzői

Hasonlósági mértékek

Mennyire hasonlít egymásra két elem? Mekkora a *távolságuk?*

- Legegyszerűbben (ha azonos típusúak az attribútumaink): nemegyezések aránya az összes attribútum számához
- Intervallum típusú attribútumok esetén sokkal jobb:
  Minkowski-norma: $L_p(\overline{z}) = (|z_1|^p + ... + |z_m|^p)^{\frac{1}{p}}$

Az adat leírása, jellemzői

Hasonlósági mértékek

Mennyire hasonlít egymásra két elem? Mekkora a *távolságuk?*

- Legegyszerűbben (ha azonos típusúak az attribútumaink): nemegyezések aránya az összes attribútum számához
- Intervallum típusú attribútumok esetén sokkal jobb:
  Minkowski-norma: $L_p(\overline{z}) = (|z_1|^p + ... + |z_m|^p)^{\frac{1}{p}}$
- $p = 2$? $p = 1$?

Az adat leírása, jellemzői

Hasonlósági mértékek

Mennyire hasonlít egymásra két elem? Mekkora a *távolságuk?*

- Legegyszerűbben (ha azonos típusúak az attribútumaink): nemegyezések aránya az összes attribútum számához
- Intervallum típusú attribútumok esetén sokkal jobb:
  Minkowski-norma: $L_p(\overline{z}) = (|z_1|^p + ... + |z_m|^p)^{\frac{1}{p}}$
- $p = 2$? $p = 1$?
- $p = 2$: euklideszi norma, $p = 1$: Manhattan-norma

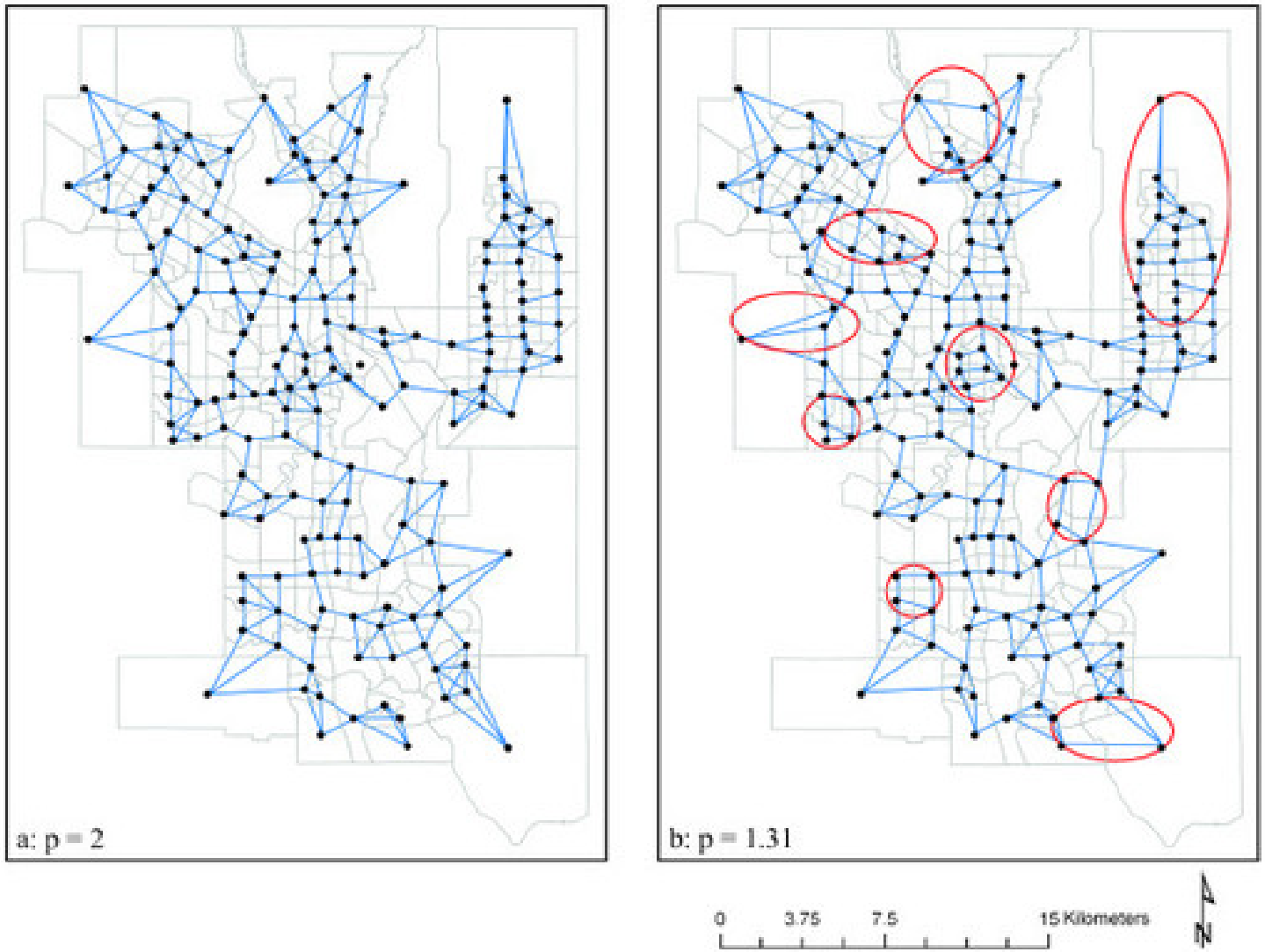
Az adat leírása, jellemzői

Hasonlósági mértékek

Mennyire hasonlít egymásra két elem? Mekkora a *távolságuk?*

- Legegyszerűbben (ha azonos típusúak az attribútumaink): nemegyezések aránya az összes attribútum számához
- Intervallum típusú attribútumok esetén sokkal jobb:
  Minkowski-norma: $L_p(\overline{z}) = (|z_1|^p + ... + |z_m|^p)^{\frac{1}{p}}$
- $p = 2$? $p = 1$?
- $p = 2$: euklideszi norma, $p = 1$: Manhattan-norma
- Vegyes attribútumok: csoportosítás, majd súlyozott távolságátlag
- Speciális esetek pl.: szerkesztési távolság, bezárt szög alapú hasonlóság

Különböző p értékek hatása a szomszédságokra



Adattisztítás

Mit kezdjünk a hiányzó adattal?

- Elhagyjuk a rekordot, vagy kézzel tömjük be a lyukakat: nem túl hatékony
- Dedikált konstans a hiány jelzésére: félrevezetheti az adatbányász alkalmazást
- A teljes attribútum, vagy az adott osztály átlagának, móduszának behelyettesítése
- Következtetés a hiányzó értékre (regresszió, döntési fa...)

Adattisztítás

Mit kezdjünk a zajos adattal?
Egy értékhalmaz esetén zajon véletlenszerű hibából fakadó hibás értékeket értünk.

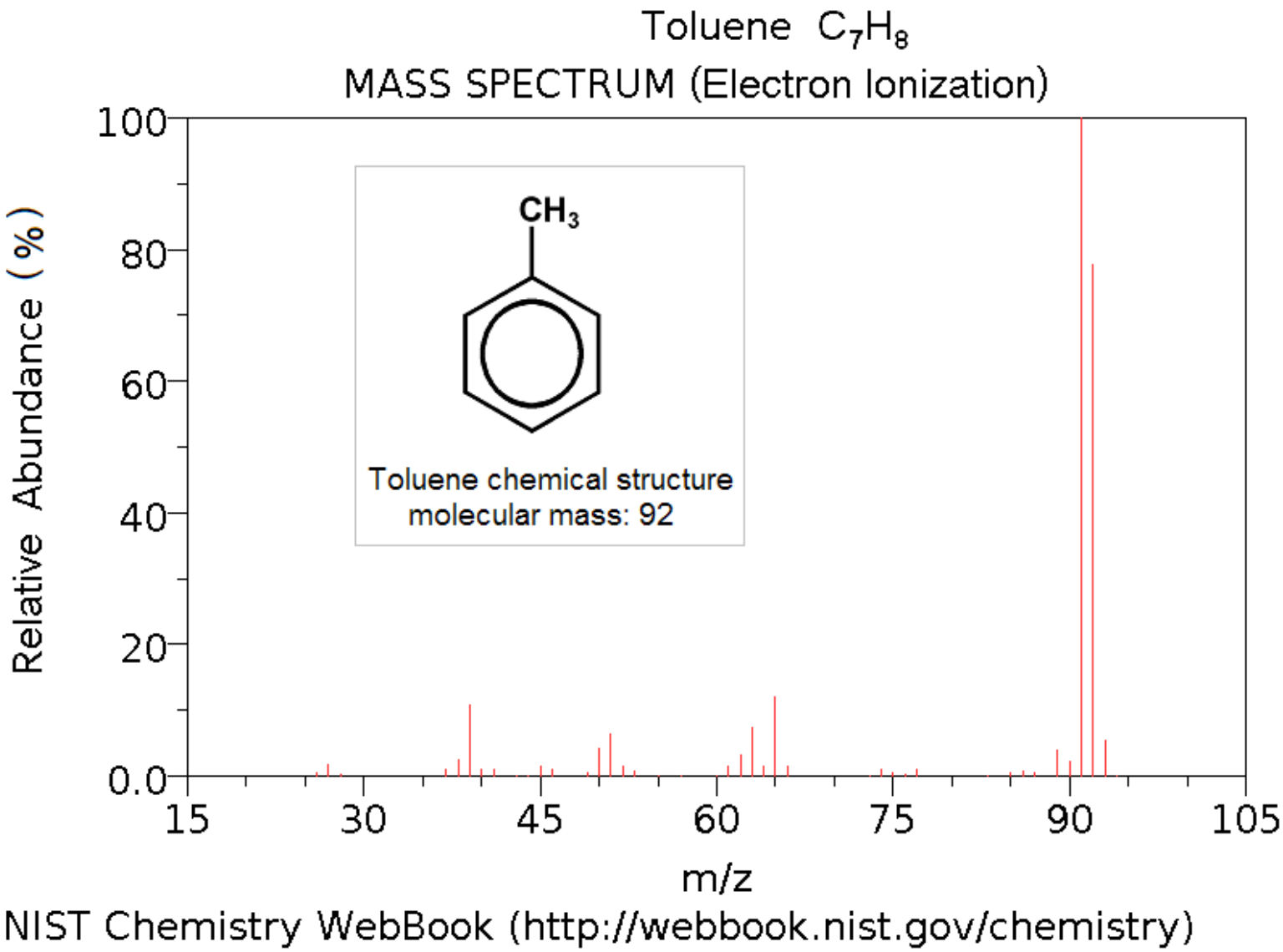
- Binning: rögzített számosságú vagy szélességű "vödrök" értékeit átlagukkal, mediánjukkal vagy a legközelebbi szélsőértékkel helyettesítjük
  Túl egyszerűnek látszik, de gyakorlatban is használható, pl. MS, NMR kísérletek

- Regresszió: az adat illesztése egy függvényhez

- Klaszterezés: hasonlóság alapú csoportosítás

- Ezek a technikák már több célra használhatók, egyben redukciós és diszkretizációs eljárások is

Toluol ionizációs tömegspektruma



Adatintegráció

Különböző források egyesítése

- Azonosítási probléma, redundancia $\rightarrow$ korrelációanalízis

- Pearson-féle korrelációs együttható: $r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B}$
  Nem jelent implikációt
  Milyen értékeket vehet fel?

Adatintegráció

Különböző források egyesítése

- Azonosítási probléma, redundancia $\rightarrow$ korrelációanalízis
- Pearson-féle korrelációs együttható: $r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B}$
  Nem jelent implikációt
  Milyen értékeket vehet fel?
- Kategória típusú attribútumokra: χ^2 statisztika (Pearson)
  A c-féle, B r-féle értéket vehet fel. Egy c x r-es táblázatot töltünk ki az eseménypárok együttes előfordulásával.
  $\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$, ahol o_{ij} a megfigyelt, e_{ij} pedig a várt együttes előfordulás
  $e_ij = \frac{\#(A=a_i)\#(B=b_i)}{N}$

Adattranszformáció

Rengeteg diverz technikával készíthetjük elő az adatokat az adatbányászathoz.

- Zajszűrés (már láttuk)
- Aggregáció (pl. havi adatokból éves kimutatás)
- Általánosítás (fogalmi, numerikus hierarchiák)
- Normalizálás
- Új attribútum létrehozása

Adattranszformáció

Normalizálás

- Min-max: $v_n = \frac{v - min}{max - min}(max_n - min_n) + min_n$
- z-score: $v_n = \frac{v - \overline{A}}{\sigma_A}$
- Decimális skálázás a céltartományhoz igazítva
- A min-max normalizálás és a decimális skálázás nem robusztus az érkező új értékekre

Adatredukció

Ez tulajdonképpen a feldolgozandó adat méretének csökkentése érdekében végzett transzformáció.

- Aggregáció
- Attribútum-részhalmaz választása(inkrementálisan, dekrementálisan, döntési fákkal)
- Méretcsökkentés (alternatív ábrázolás, modellezés)
- Dimenziócsökkentés (ez is új reprezentáció, de valamilyen kódolás, leképezés, "tömörítés")
  Hasznos, ha nagyon nagy az attribútumhalmaz számossága (pl. szópárok gyakorisága az interneten: $n \approx 10^9$)

Adatredukció

Dimenziócsökkentés: DWT (Discrete Wavelet Transform)

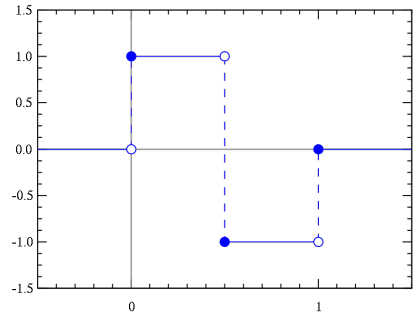
- Az adatrekordra vektorként tekintünk, a DWT ezt egy azonos hosszúságú vektorrá transzformálja. De akkor mire jó?

Adatredukció

Dimenziócsökkentés: DWT (Discrete Wavelet Transform)

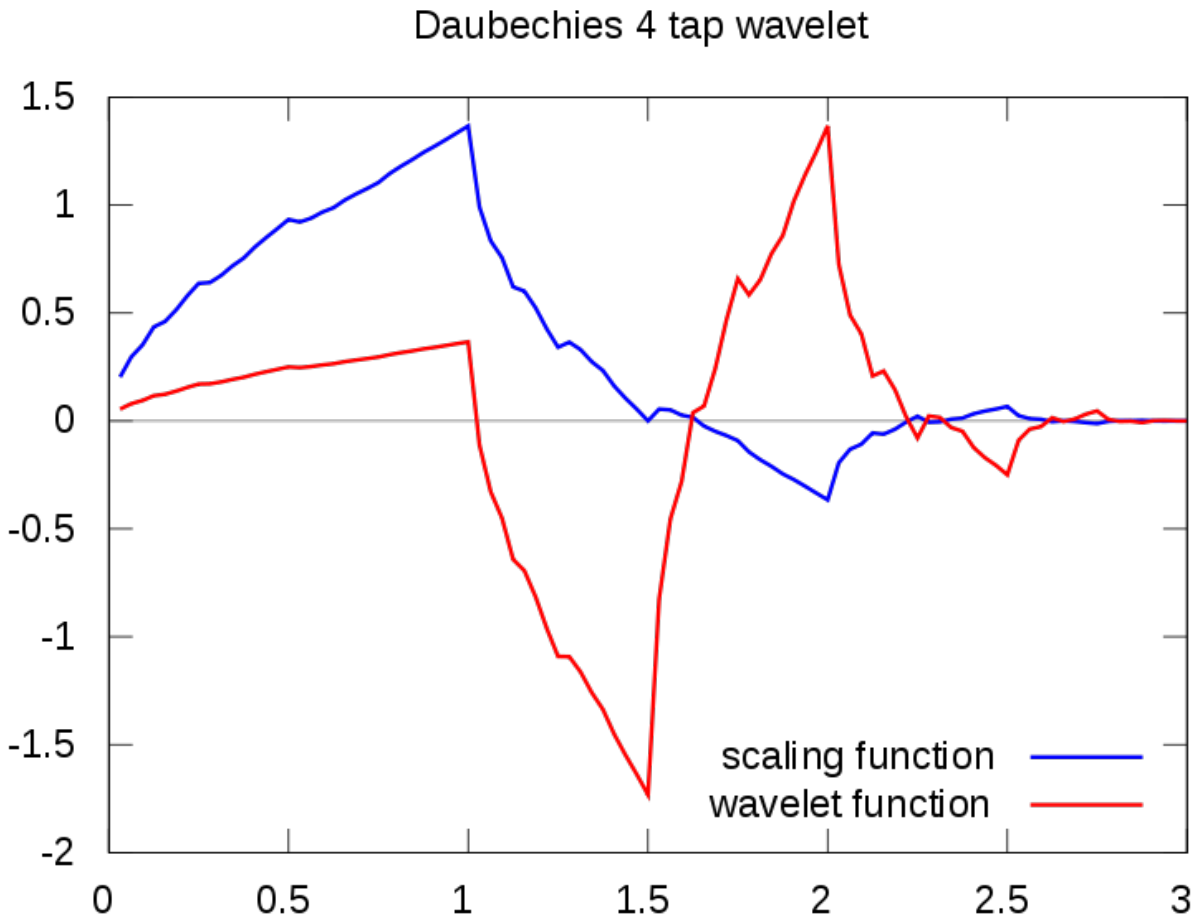
- Az adatrekordra vektorként tekintünk, a DWT ezt egy azonos hosszúságú vektorrá transzformálja. De akkor mire jó?
- **Az új vektornak néhány együtthatójából is jól közelíthető az eredeti adat.**
- A kiinduló vektor mérete 2 hatványa legyen (padding szükséges lehet). Egy transzformáció során két függvényt alkalmazunk szomszédos adatpárokra, iteratívan, minden ciklusban felezve ezzel az adathalmaz méretét. Az egyik függvény jellemzően simítja az adatot, a másik pedig a különbségeket erősíti.
- A futás során előálló, kijelölt értékek lesznek a transzformált vektor együtthatói. A transzformáció inverzét végrehajtva az adat visszaállítható.

De mi az a wavelet (”hullámocska”)?

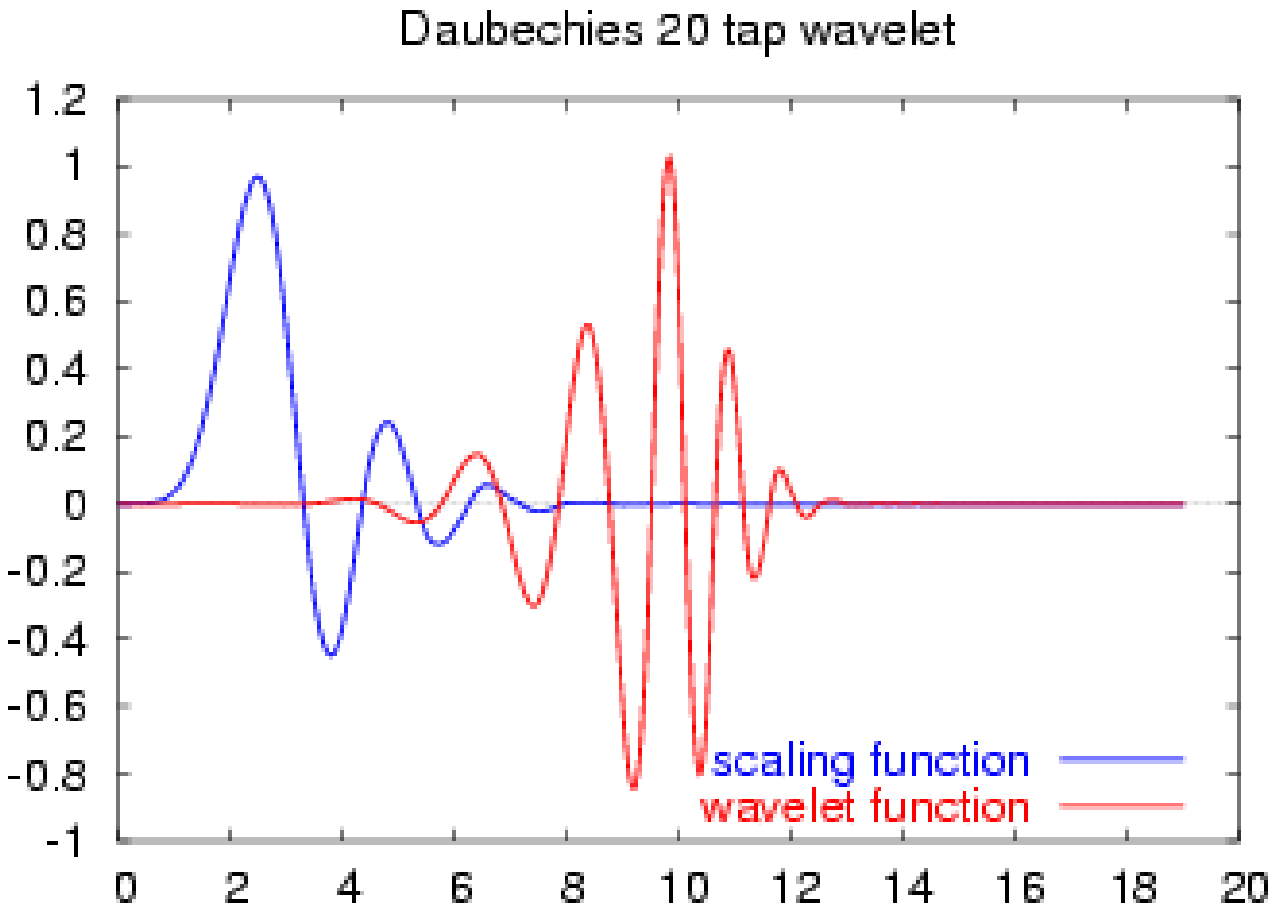


- A képen a legegyszerűbb diszkrét wavelet, a Haar-wavelet látható. (Haar Alfréd, 1909)
- A waveletek egy ortonormált bázisban történő leírást tesznek lehetové. (Két függvény: wavelet függvény az ortogonalitáshoz, és egy skálázó függvény az ortonormalitáshoz.
- A Haar-wavelet rekurzívan páronkénti különbségeket ad meg, illetve a teljes adatsor összegét.
- A DWT elonye a DFT-vel szemben, hogy nem csak a frekvenciát ábrázolja, hanem a lokalitást is.

Vannak összetettebb waveletek is! (Ingrid Daubechies, 1988)



Vannak összetettebb waveletek is! (Ingrid Daubechies, 1988)



DWT vs. DFT az egységimpulzus példáján

$$(1,0,0,0) = \frac{1}{4}(1,1,1,1) + \frac{1}{4}(1,1,-1,-1) + \frac{1}{2}(1,-1,0,0) \qquad \text{Haar DWT}$$

$$(1,0,0,0) = \frac{1}{4}(1,1,1,1) + \frac{1}{2}(1,0,-1,0) + \frac{1}{4}(1,-1,1,-1) \qquad \text{DFT}$$

$$\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$$

$$\left(\frac{1}{2},\frac{1}{2},0,0\right) \qquad \text{2-term truncation}$$

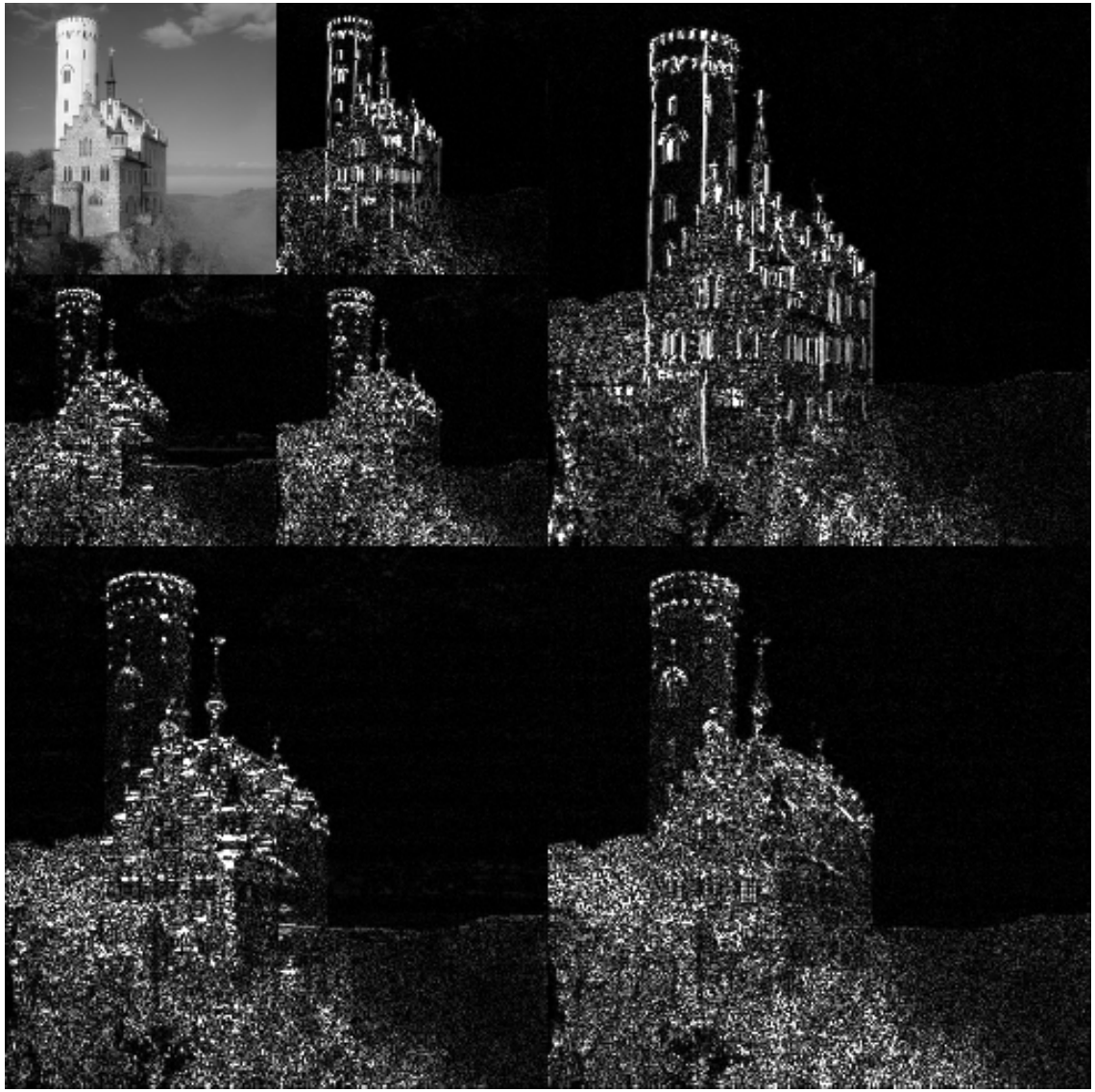
$$(1,0,0,0)$$

$$\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right)$$

$$\left(\frac{3}{4},\frac{1}{4},-\frac{1}{4},\frac{1}{4}\right) \qquad \text{2-term truncation}$$

$$(1,0,0,0)$$

DWT a gyakorlatban - JPEG2000



JPEG JFIF vs. JPEG2000

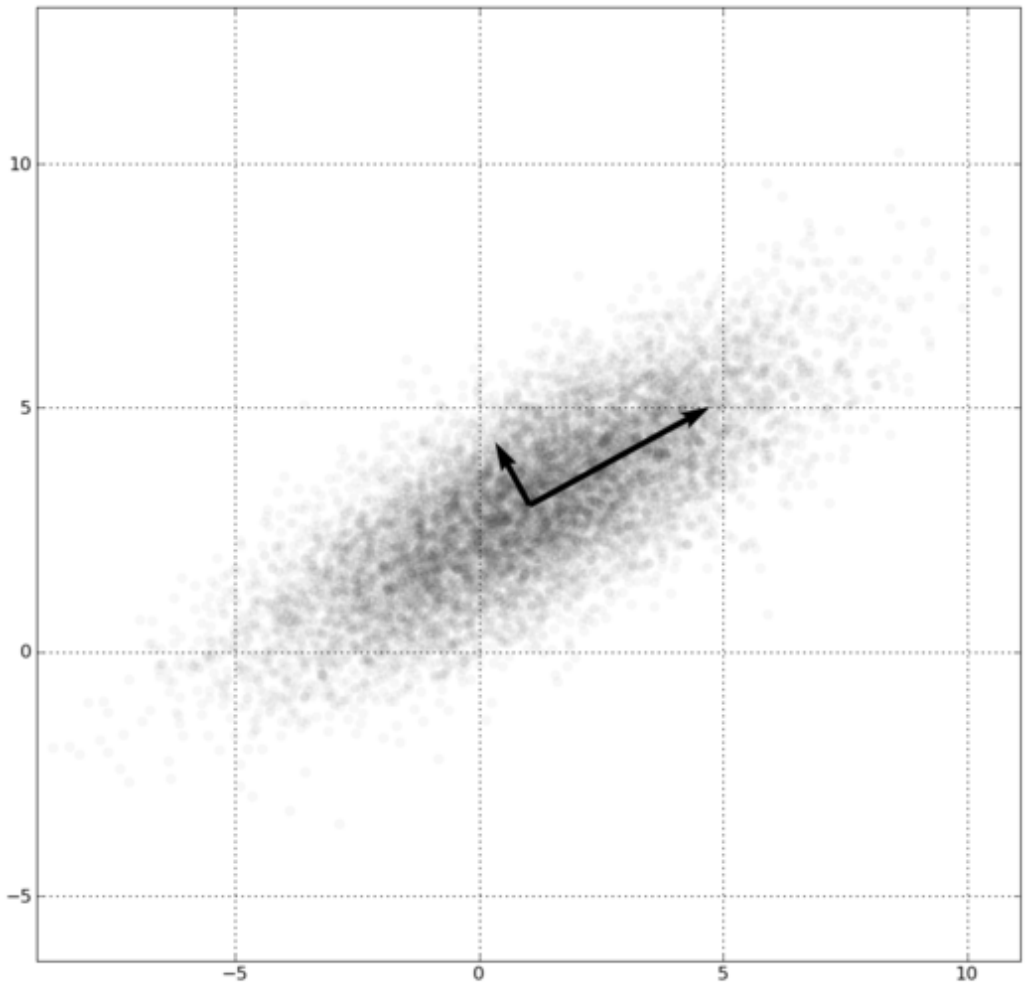
Adatredukció

Dimenziócsökkentés: PCA (Principal Component Analysis - főkomponens-analízis, szinguláris felbontás, Karhunen-Loeve módszer)

Először: intuitíven

- A PCA-val egy új ortogonális bázist találhatunk az eredeti adathalmazhoz.
- Ezeket a vektorokat, amiket főkomponenseknek nevezünk, szignifikancia szerint sorba rendezhetjük, kezdve a legerősebb szórás irányába mutatóval.
- A leggyengébb komponensek elhagyásával is az eredeti adat jó közelítését kapjuk.
- (Bizonyos értelemben a leheto legjobb közelítést, mint azt látni fogjuk.)

PCA Gauss-eloszlásra



Adatredukció

Dimenziócsökkentés: PCA (Principal Component Analysis - főkomponens-analízis, szinguláris felbontás, Karhunen-Loeve módszer)
Másodszor: alapos(abb)an - PCA SVD használatával

- **Definíció**
  Ortogonális mátrix: Az U négyzetes mátrix ortogonális, ha létezik U^{-1} inverze és $U^{-1} = U^T$

Adatredukció

Dimenziócsökkentés: PCA (Principal Component Analysis - főkomponens-analízis, szinguláris felbontás, Karhunen-Loeve módszer)
Másodszor: alapos(abb)an - PCA SVD használatával

- **Definíció**
  Ortogonális mátrix: Az U négyzetes mátrix ortogonális, ha létezik U^{-1} inverze és $U^{-1} = U^T$
- Ortogonális mátrix által reprezentált lineáris transzformáció nem változtatja a vektorok hosszát.
- **Tétel**
  Minden $M \in \mathbb{R}^{m \times n}$ mátrixnak létezik szinguláris érték felbontása (SVD): $M = U\Sigma V^T$, ahol $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{m \times n}$, és Σ bal felső részmátrixa egy $r \times r$ diagonális mátrix, a főátlójában csökkenő sorrendben pozitív elemekkel, mindenhol máshol pedig 0 elemekkel.

Adatredukció

Dimenziócsökkentés: PCA (Principal Component Analysis - főkomponens-analízis, szinguláris felbontás, Karhunen-Loeve módszer)
Miért jó ez nekünk a dimenziócsökkentésnél?

- $M = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

Adatredukció

Dimenziócsökkentés: PCA (Principal Component Analysis - főkomponens-analízis, szinguláris felbontás, Karhunen-Loeve módszer)
Miért jó ez nekünk a dimenziócsökkentésnél?

- $M = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$
- Vegyük csak a k legnagyobb súlyú diadikus szorzatot! $M_k = U_k \Sigma_k V_k^T$, M_k rangja k
- **Tétel**
  $||M - M_k||_F = min||M - N||_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2}$, ahol a minimumot a k rangú mátrixok N halmazán vesszük.
- Az M mátrix $U_k\Sigma_k$-val közelítheto, ahol V_k^T sorai alkotják a bázist.

Adatredukció

Méretcsökkentés

- Parametrikus módszerek: nem kell eltárolni az adatot, csak néhány paramétert
  Példa: regresszió

- Nemparametrikus módszerek: klaszterezés, mintavételezés
  - SRSWOR, SRSWR (Simple Random Sample WithOut/With Replacement): s rekordot "húzunk", az utóbbinál ugyanaz többször is húzható
  - Ha van klaszter- vagy osztályinformáció, azt felhasználhatjuk

Adatredukció

Mintavételezés hibája a méret függvényében

- Első közelítés: $hiba(m) = \mathbb{P}(|\frac{Y_x}{m} - p_x| \geq \epsilon)$, ahol m a minta mérete, Y_x az x elem előfordulásainak száma a mintában, p_x pedig x előfordulásának valószínűsége
- **Csernov-korlátból:** $m \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\sigma}$, ahol σ a kívánt hibakorlát
- Pl. ha 0,01 vagy nagyobb eltérés esélyét 0,01 alá akarjuk csökkenteni, 27000 mintát kell vennünk
- Ez a megközelítés nem szerencsés, mert kisebb p valószínuség mellett kisebb hibát ad
- Jobb: $hiba(m) = \mathbb{P}(\frac{Y_x}{mp} \geq 1 + \epsilon) + \mathbb{P}(\frac{Y_x}{mp} \leq \frac{1}{1+\epsilon})$

Diszkretizálás és hierarchiák

Csökkenti, egyszerűsíti, ugyanakkor megfoghatóbbá teszi az adatot.

- Lehet *top-down*, illetve *bottom-up* megközelítésű (splitting ill. merging), aszerint, hogy az osztópontok kezdeti halmazához továbbiakat veszünk, vagy kezdetben minden pont osztópont, és ezt csökkentjük
- Lehet *felügyelt*, ha a diszkretizáláshoz használunk osztályinformációt
- Létrehozhatunk több szintű hierarchiát is rekurzív diszkretizálással

Diszkretizálás és hierarchiák

Módszerek numerikus adatokon

- Binning (már láttuk)
- Az 1R algoritmus diszkretizáló eljárása
- Entrópia alapú diszkretizálás
- ChiMerge
- Klaszterezés
- Intuitív feldarabolás (kívánatos lehet a természetesebb határok érdekében)

Diszkretizálás és hierarchiák

Entrópia alapú diszkretizálás

- Felügyelt, top-down
- Entrópia: $H(D) = -\sum_{C_i \in C} p_i \log_2 p_i$, ahol p_i a C_i osztályú elemek relatív gyakorisága D-ben
- Úgy választunk vágópontot, hogy $\frac{|D_1|}{|D|} H(D_1) + \frac{|D_2|}{|D|} H(D_2)$ minimális legyen
- Ez annak az információnak a mértéke, amennyi hozzáadásával tökéletessé tehető lenne a vágás osztályozási képessége
- Rekurzívan, amíg el nem érünk valamilyen kívánt határt (hiányzó információban, intervallumok számosságában)

Diszkretizálás és hierarchiák

ChiMerge

- Felügyelt, bottom-up diszkretizálás a már megismert χ^2 teszt használatával
- A táblázat oszlopai azok a szomszédos intervallumok, amiknek kíváncsi vagyunk az összevonhatóságára, sorai pedig az osztályok
- Alacsony χ^2 értékek esetén az intervallumok osztályfüggetlenek és összevonhatók
- Megállási kritériumok: χ^2, intervallumok száma

Diszkretizálás és hierarchiák

Ökölszabály az intuitív diszkretizáláshoz

- **3-4-5 szabály**

- A decimális értékek MSD-je alapján határozzuk meg a következő szintű intervallumok számát
  3, 6, 7 és 9 különbözo érték esetén: 3 részre
  2, 4 és 8 különbözo érték esetén: 4 részre
  1, 5 és 10 különözo érték esetén: 5 részre
  osztunk

- Előtte megfontolható az extrémumok kihagyása (pl. 5%-95%), de utána szükség lehet új intervallumokat létrehozni nekik

Köszönöm a figyelmet!