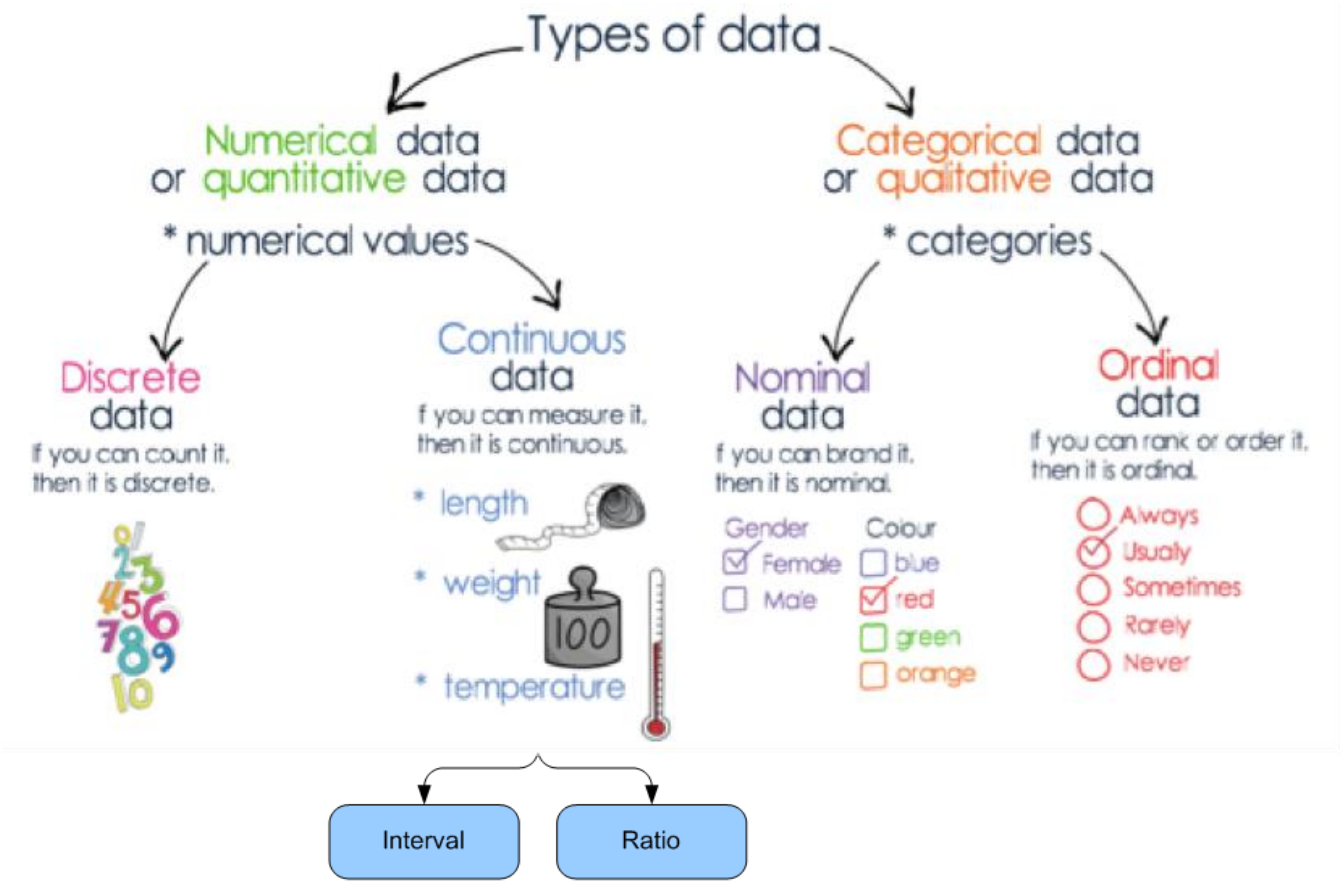
Bevezető alapismeretek

Fogarassyné Vathy Ágnes *vathy.agnes@mik.uni-pannon.hu*

Adattípusok



Adatok típusai (1)

Az adatok milyen értékeket vehetnek fel?

- Folytonos: A folytonos változók egy adott skálán tetszőleges értékeket vehetnek fel. A skála bármely két értéke között végtelen sok újabb érték helyezkedik el, s a folytonos típusú változók ezen értékek bármelyikét felvehetik.
  - Pl: hőmérséklet, magasság, …
  - Általában lebegőpontos változókkal reprezentáljuk.

- Kategorikus (diszkrét): az adatok a mérési skálán nem vehetnek fel tetszőleges értékeket, a felvehető értékek jól elkülönülnek egymástól, köztük „rés” van. Véges vagy megszámlálható végtelen sok értéke lehet.
  - Pl: nem, beosztás, irányítószám

Adatok típusai (skálatípusok)

Az értékek hogyan viszonyulnak egymáshoz?

- Felsorolás típusú (nominális): csak az vizsgálható, hogy 2 érték azonos-e vagy különbözik
  - pl.: ID, szemszín, irányítószám
- Rendezett típusú (ordinális, sorrendi): az adatok nagyság szerint sorba rendezhetőek
  - pl.: rangsorolás, fokozat, magasság, mint {magas, átlagos, alacsony}.
- Intervallumskálázott: a sorrendiség mellet a különbség is meghatározható, de az arány nem – a 0 pont megválasztása tetszőleges
  - pl.: dátum, hőmérséklet Celsiusban vagy Fahrenheitben
- Arányskálázott: a felvett értékek aránya is jelentést hordoz
  - pl.: abszolút hőmérséklet (Kelvin), hosszúság, idő

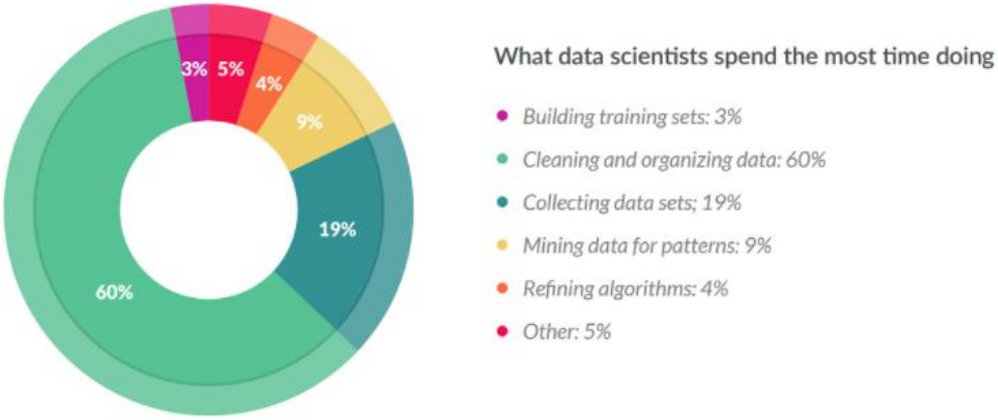
Adatok típusai

Attributum típusa	Leírás	Példák	Műveletek
Névleges (nominális)	Egy névleges attributum értékei csak különböző nevek, azaz csak ahhoz nyújt elegendő információt, hogy egy objektumot megkülönböztessünk egy másiktól. (=, ≠)	irányítószám, dolgozó azonosító, szemszín, nem: {*férfi, nő*}	módusz, entropia, kontingencia korreláció, χ^2 érték
Sorrendi (ordinális)	Egy rendezett attributum értékei ahhoz nyújtanak elegendő információt, hogy rendezzük az objektumokat. (<, >)	ásványok keménysége {*jó, jobb, legjobb*}, fokozat, házszám	medián, percentilis, rang korreláció, széria próba, előjel ill. előjeles rangösszeg próba
Intervallum	Egy intervallum attributumnál az értékek közötti különbségek is jelentéssel bírnak. (+, -)	naptári dátumok, hőmérséklet Celsiusban ill. Fahrenheitben	átlag, szórás, Pearson féle korreláció, *t* és *F* próba
Hányados	Hányados változónál a különbségnek és a hányadosnak egyaránt van értelme. (*, /)	abszolút hőmérséklet, pénzügyi mennyiség, kor, tömeg, hossz, elektromos áram	mértani és harmónikus közép, százalék variáció

Fontosabb adatelőkészítési technikák

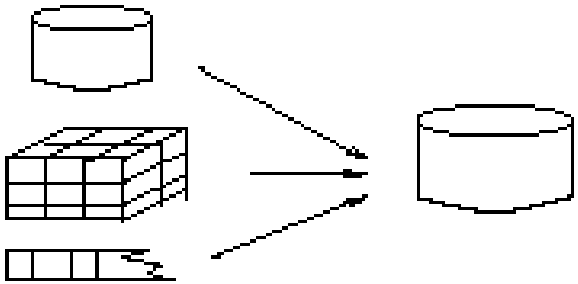
Adatok előkészítése

- Adatelőkészítés:
  - Adatintegráció
  - Adattisztítás
  - Adattranszformáció
  - Adatredukció
  - Adatdiszkretizáció

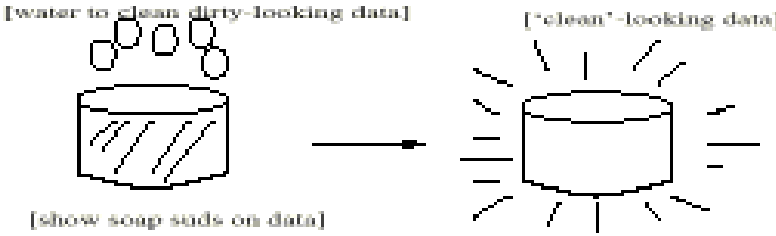


Az adatelőkészítés fő feladatai 1.

- Adatintegráció
  - különféle adatbázisok, adatkockák, vagy fájlok egyesítése



- Adattisztítás
  - hiányzó értékek kitöltése, zajos adatok simítása, outlier-ek azonosítása vagy törlése, inkonzisztenciák feloldása, adathibák karanténba helyezése

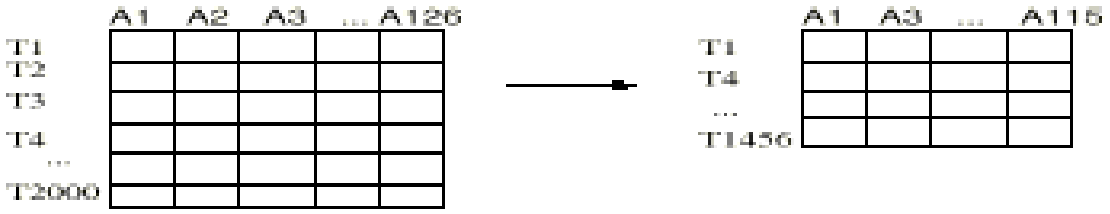


Az adatelőkészítés fő feladatai 2.

- Adattranszformáció
  - adatok konvertálása új alakra
    - pl: normalizáció és aggregáció

$$-2, 32, 100, 59, 48 \rightarrow -0.02, 0.32, 1.00, 0.59, 0.48$$

- Adatredukció
  - Az adatok kisebb mértékű reprezentációja, de azonos, vagy hasonló elemzési eredmény.



- Adatdiszkretizáció
  - Az adatok egy részének különleges jelentőségű redukciója + transzformációja, főként numerikus adatokon.

Adatok előkészítése

- Adatok átkódolása
- Adatok diszkretizálása
- Adatok vödrözése
- Attribútumértékek skálázása

Dummy változó – one-hot-encoding

- dummy változó: csak két lehetséges értéke van
  - kvalitatív (minőségi) változó
    - pl. igen/nem; tag/nem tag; stb.
  - kódolása kvantitatív módon történik: 0 és 1 formában
    - pl. igen=1 nem=0

- one-hot encoding: kódolási technika, amely tetszőleges kategorikus változót dummy változókká alakít át
  - Miért van erre szükség? – például regressziós módszerek alkalmazása esetén.

ID	Név	Szak
1	Kiss Aranka	gazdaságinformatikus
2	Tóth János	programtervező informatikus
3	Nagy Péter	mérnökinformatikus
4	Nagy Ábel	programtervező informatikus

ID	Név	Szak_GI	Szak_PI	Szak_MI
1	Kiss Aranka	1	0	0
2	Tóth János	0	1	0
3	Nagy Péter	0	0	1
4	Nagy Ábel	0	1	0

Dummy változók száma

Hány dummy változóra van szükség?

Triviális kódolás

- Minden lehetséges értéknek egy dummy változó

ID	Név	Szak_GI	Szak_PI	Szak_MI
1	Kiss Aranka	1	0	0
2	Tóth János	0	1	0
3	Nagy Péter	0	0	1
4	Nagy Ábel	0	1	0

Referencia kódolás

- Referencia csoport: minden dummy változó 0 értéket vesz fel

ID	Név	Szak_GI	Szak_PI
1	Kiss Aranka	1	0
2	Tóth János	0	1
3	Nagy Péter	0	0
4	Nagy Ábel	0	1

- dummy változó csapda: függetlennek tekintett változók között multikollinearitás[1] áll fenn
  - triviális kódolás esetén az egy változó értéke meghatározható a többi változó értékéből
- *Olyan modellekben (pl. regresszió), amelyek függetlenséget feltételeznek a prediktív változók között, k értékű kategorikus változót $k-1$ dummy változóval szabad csak kódolni!!!*

1: multikollinearitás: egyik változó értéke megjósolható a többi változó lineáris kombinációjaként

Label encoding

- Label encoding: a kategorikus változókat (abc sorrend alapján) egyedi egész értékekkel helyettesíti
  - a megfeleltetés egy-az-egyhez típusú és illesztéssel (*fit, fit_transform*) jön létre
  - mivel egy-az-egyhez megfeleltetés jön létre, ezért létezik hozzá dekóder is, ami az ellentétel konverziót valósítja meg

ID	Név	Szak
1	Kiss Aranka	gazdaságinformatikus
2	Tóth János	programtervező informatikus
3	Nagy Péter	mérnökinformatikus
4	Nagy Ábel	programtervező informatikus

ID	Név	Szak
1	Kiss Aranka	0
2	Tóth János	2
3	Nagy Péter	1
4	Nagy Ábel	2

Kihívások:

- a kategorikus értékeket gyakorlatilag ordinális értékekre cseréltük
- fenn áll a veszélye, hogy a kialakítandó modellek beépítik a sorrendi kapcsolatokat
  - India < Japán ???

One-hot encoding vs. Label encoding

Mikor melyiket érdemes használni?

- One-hot encoding:
  - a kategorikus értékek nem ordinálisak
  - a kategorikus értékek száma viszonylag alacsony
- Label encoding:
  - a kategorikus értékek abc sorrend szerint ordinálisak (pl. általános iskola, gimnázium, Bsc, …)
  - a kategorikus értékek száma meglehetősen magas, ezért a one-hot encoding nagy memóriaigényű lenne
- Vegyük figyelembe az alkalmazandó modell sajátosságait!

Ordinal encoding

- Ordinal encoding: az ordinális sztring értékű változókat megadott sorrend alapján egyedi egész értékekkel helyettesíti
  - a megfeleltetés egy-az-egyhez típusú és illesztéssel (*fit, fit_transform*) jön létre
  - mivel egy-az-egyhez megfeleltetés jön létre, ezért létezik hozzá dekóder is, ami az ellentétel konverziót valósítja meg

ID	Név	Iskolai végzettség
1	Bármi Áron	Egyetem BSc
2	Pál Kata	Egyetem MSc
3	Tóth Aranka	Középiskola
4	Kis Péter	Egyetem MSc
5	Kakukk Tomi	Általános iskola

ID	Név	Szak
1	Bármi Áron	3
2	Pál Kata	4
3	Tóth Aranka	2
4	Kis Péter	4
5	Kakukk Tomi	1

Adatok előkészítése

- Adatok átkódolása
- Adatok diszkretizálása
- Adatok vödrözése
- Attribútumértékek skálázása

Diszkretizáció

Diszkretizáció: folytonos értékek kategorikus intervallumokba osztása

- = adatredukció + transzformáció
- *pl: életkor -> korcsoport kategóriák*
- gyakran szükséges, mert
  - számos algoritmus csak kategorikus adatokkal dolgozik
  - a tényleges érték nem fontos csak a „nagyságrendje”

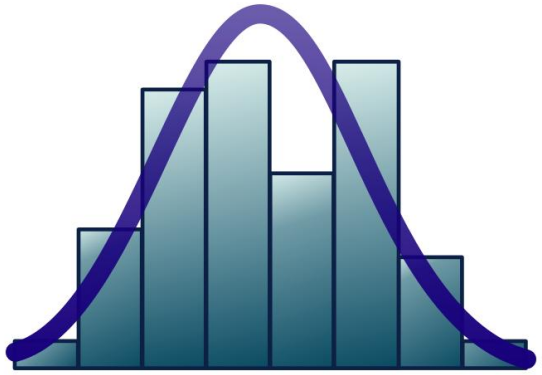
- diszkretizációs módszerek pl.:
  - vödrözés alkalmazása
  - klaszterezés
  - entrópia alapú diszkretizálás
  - szakterületi szabályok alapján
  - intuitív feldarabolás (kívánatos lehet a természetesebb határok érdekében)

Adatok előkészítése

- Adatok átkódolása
- Adatok diszkretizálása
- Adatok vödrözése
- Attribútumértékek skálázása

Binning – vödrözés

- Egyenlő szélességű partícionálás:
  - Távolság alapú vödrözés
  - A tartományt N egyenlő intervallumra osztja: egyforma rács.
  - Ha *A* és *B* az attribútum legnagyobb és legkisebb értékei, akkor az intervallum: $W = (B - A)/N$
  - De az outlier-ek meghatározóak!

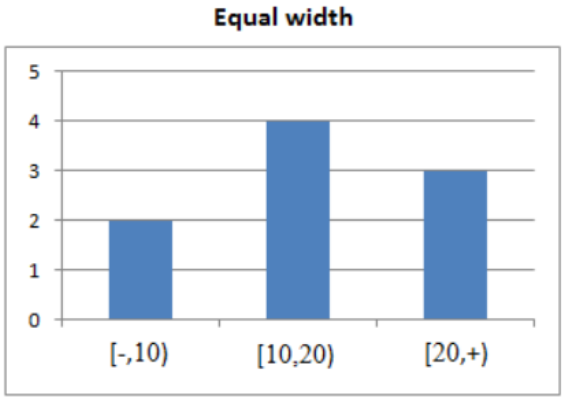


- Egyenlő mélységű partícionálás:
  - Gyakoriság alapú vödrözés
  - A tartományt N intervallumra osztja, mindegyik megközelítőleg ugyanannyi elemet tartalmaz

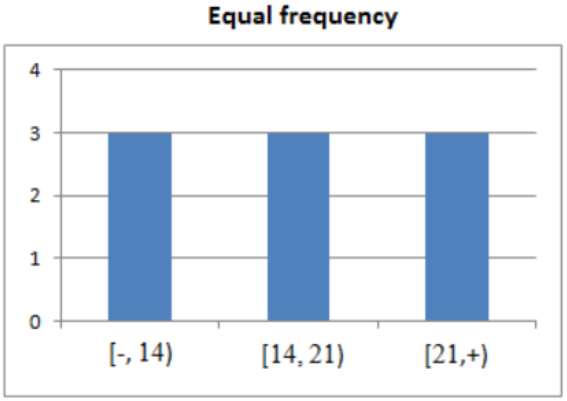
„Vödrözési” módszer az adatok simítására

Rendezett adatok: 0, 7, 12, 16, 16, 18, 24, 26, 28

- Partícionálás egyenlő szélességű vödrökbe:
  - Bin 1 [0-10): 0, 7
  - Bin 2 [10-20): 12, 16, 16, 18
  - Bin 3 [20, +): 24, 26, 28



- Partícionálás egyenlő mélységű vödrökbe:
  - Bin 1: 0, 7, 12
  - Bin 2: 16, 16, 18
  - Bin 3: 24, 26, 28



Adatok előkészítése

- Adatok átkódolása
- Adatok diszkretizálása
- Adatok vödrözése
- Attribútumértékek skálázása

Attribútum értékének skálázása

Tulajdonság skálázása (feature scaling) vs. normalizálás (normalization):

- Normalizálás: az adatok eloszlása megváltozik
  - hasznos ha tudjuk, hogy az adatok nem normál eloszlásúak
- Skálázás: az adatok tartománya változik

Módszerek:

- min-max skálázás
- z-score standardizálás
- L-norma skálázás
- normalizáció decimális skálázással
- Log transzformáció

További lehetőségek:
https://pub.towardsai.net/feature-transformation-and-scaling-techniques-f9645cb538e

Módszerek

- min-max skálázás: *A* attribútum v értékét transzformálja lineáris módon a $[new\_min_A, new\_max_A]$ tartományra:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- L-norma átalakítás:
  - L1-normalizáció: (Least Absolute Deviation – L1 norma alapján)
    - a normált adatok abszolútértékének összege 1
    - kevésbé érzékeny az outlierekre
  - L2-normalizáció: (Least Squares – L2 norma alapján)
    - a normált adatok négyzeteinek összege 1

Módszerek

- z-score standardizálás:

$$v' = \frac{v - avg_A}{s_A}$$

ahol avg_A az A attribútum átlagát és s_A az A attribútum szórását jelenti.

  - megmutatja milyen távol vannak az értékek az átlagtól

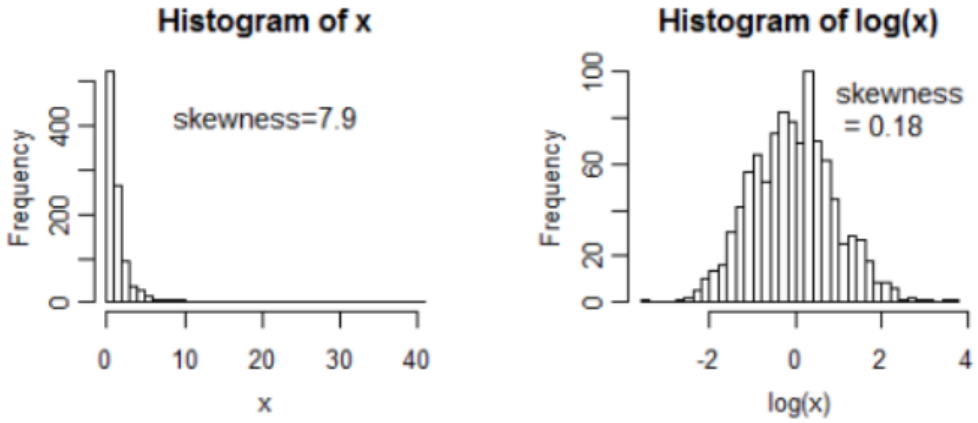
- Decimális skálázás:

$$v' = \frac{v}{10^j}$$

ahol j az a legkisebb egész szám, amire $\max(|v'|) < 1$

Módszerek

- Log transzformáció: $y = log_b(x)$



Example distribution before (left) and after (right) log transformation

- a ferde adatok kezelésére, azért, hogy csökkentsük a változatosságot és kevésbé ferdévé tegyük az adatokat
- segít megfelelni a normalitás követelményének azon algoritmusokban, ahol ez elvárás

Min-max skálázás – példa

Telefonálási adatok:

ÜgyfélID	Beszéd (mp)	SMS (db)	Adatforgalom (MB)
1	15000	20	800
2	5600	1	4500
3	4200	33	1500
4	18000	12	1200
5	8500	7	600

Kiszámítható 2 ügyfél távolsága?

- pl.: ügyfél1 és ügyfél2 távolsága:

$$d_{(1,2)} = \sqrt{(15000-5600)^2+(20-1)^2+(800-4500)^2} = 10102$$

Jó megoldás?

Min-max skálázás – példa

Távolságmátrix min-max skálázás nélkül:

	1	2	3	4	5
1	0	10102,00	10822,67	3026,56	6503,09
2	10102,00	0	3310,74	12831,61	4860,04
3	10822,67	3310,74	0	13803,28	4393,25
4	3026,56	12831,61	13803,28	0	9518,93
5	6503,09	4860,04	4393,25	9518,93	0

Min-max skálázás – példa

Min-max skálázás:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

ÜgyfélID	Beszéd (mp)	SMS (db)	Adatforgalom (MB)
1	15000	20	800
2	5600	1	4500
3	4200	33	1500
4	18000	12	1200
5	8500	7	600

- pl. a *Beszéd* attribútumot [0,1]-re skálázva:
  - 2-es ügyfél esetében:

$$= \frac{5600 - 4200}{18000 - 4200} * (1 - 0) + 0 = 0{,}1014$$

Min-max skálázás – példa

Eredeti adatok:

ÜgyfélID	Beszéd (mp)	SMS (db)	Adatforgalom (MB)
1	15000	20	800
2	5600	1	4500
3	4200	33	1500
4	18000	12	1200
5	8500	7	600

Skálázott adatok a [0,1] tartományra:

ÜgyfélID	Beszéd (másodperc)	SMS (db)	Adatforgalom (MB)
1	0,7826	0,5938	0,0513
2	0,1014	0,0000	1,0000
3	0,0000	1,0000	0,2308
4	1,0000	0,3438	0,1538
5	0,3116	0,1875	0,0000

Beszédes? – Mit olvashatunk le???

Min-max skálázás – példa

Távolságmátrix skálázás nélkül:

	1	2	3	4	5
1	0	10102,00	10822,67	3026,56	6503,09
2	10102,00	0	3310,74	12831,61	4860,04
3	10822,67	3310,74	0	13803,28	4393,25
4	3026,56	12831,61	13803,28	0	9518,93
5	6503,09	4860,04	4393,25	9518,93	0

Távolságmátrix min-max skálázással:

	1	2	3	4	5
1	0	1,31	0,90	0,35	0,62
2	1,31	0	1,27	1,28	1,04
3	0,90	1,27	0	1,20	0,90
4	0,35	1,28	1,20	0	0,72
5	0,62	1,04	0,90	0,72	0,0