Gyakorló feladatok adatbányászati technikák tantárgyhoz

Buza Krisztián
Számítástudományi és Információelméleti Tanszék
Budapesti Műszaki és Gazdaságtudományi Egyetem

Klaszterezés kiértékelése

Feladat: Egy klaszterező algoritmus által adott csoportosítást referencia-csoportokhoz (osztályokhoz) viszonyítunk. Az alábbi táblázat mutatja az egyes klaszterekbe tartozó példányok számát osztályonként. Számítsa ki a klaszterező algoritmus eredményének aggregált F-measure-jét! Mikor teljesít az algoritmus jól: ha a kiszámított F-measure nagy vagy ha kicsi?

	1. klaszter	2. klaszter	3. klaszter	4. klaszter
1. osztály	40	20	15	15
2. osztály	5	20	5	10
3. osztály	5	10	5	25

Klaszterezés kiértékelése

Megoldás: 1. lépés: Úgy tekintjük, hogy a *j*-dik klaszter válasz egy olyan lekérdezésre, amelyre a helyes válasz az *i*-dik osztálybeli példányok halmaza. Kiszámoljuk az összes osztály-klaszter párra a precision-t és recall-t.

Jelölés: *precision(i,j)* ill. *recall(i,j*) = j-dik klaszter i-dik osztályra vonatkozó precision-je ill. recall-ja

	1. klaszter	2. klaszter	3. klaszter	4. klaszter
1. osztály	40	20	15	15
2. osztály	5	20	5	10
3. osztály	5	10	5	25

Klaszterezés kiértékelése

Megoldás (folytatás):

precision(1,1) = 40 / **50**
recall(1,1) = 40 / 90
precision(1,2) = 20 / **50**
recall(1,2) = 20 / 90

precision(1,3) = 15 / **25**
recall(1,3) = 15 / 90
precision(1,4) = 15 / **50**
recall(1,4) = 15 / 90

	1. k.	2. k.	3. k.	4. k.	Összesen
1. oszt.	40	20	15	15	90
2. oszt.	5	20	5	10	40
3. oszt.	5	10	5	25	45
Összesen	**50**	**50**	**25**	**50**	

Klaszterezés kiértékelése

Megoldás (folytatás):

precision(2,1) = 5 / **50**
recall(2,1) = 5 / 40
precision(2,2) = 20 / **50**
recall(2,2) = 20 / 40

precision(2,3) = 5 / **25**
recall(2,3) = 5 / 40
precision(2,4) = 10 / **50**
recall(2,4) = 10 / 40

	1. k.	2. k.	3. k.	4. k.	Összesen
1. oszt.	40	20	15	15	90
2. oszt.	5	20	5	10	40
3. oszt.	5	10	5	25	45
Összesen	**50**	**50**	**25**	**50**	

Klaszterezés kiértékelése

Megoldás (folytatás):

precision(3,1) = 5 / **50**
recall(3,1) = 5 / 45
precision(3,2) = 10 / **50**
recall(3,2) = 10 / 45

precision(3,3) = 5 / **25**
recall(3,3) = 5 / 45
precision(3,4) = 25 / **50**
recall(3,4) = 25 / 45

	1. k.	2. k.	3. k.	4. k.	Összesen
1. oszt.	40	20	15	15	90
2. oszt.	5	20	5	10	40
3. oszt.	5	10	5	25	45
Összesen	**50**	**50**	**25**	**50**	

Klaszterezés kiértékelése

Megoldás (folytatás):

F-measure-t számolunk az osztály-klaszter-párokra

F = 2*precision*recall / (precision + recall)

precision(1,1) = 40 / **50**

recall(1,1) = 40 / 90

f(1,1) = 2*0.8*0.44 / (0.8+0.44) = 0.57

precision(1,2) = 20 / **50**

recall(1,2) = 20 / 90

f(1,2) = 2*0.4*0.22 / (0.4+0.22) = 0.28 ...

Klaszterezés kiértékelése

Megoldás (folytatás):

f(1,1) = **0.57**	f(2,1) = 0.11	f(3,1) = 0.11
f(1,2) = 0.28	f(2,2) = **0.44**	f(3,2) = 0.21
f(1,3) = 0.26	f(2,3) = 0.15	f(3,3) = 0.14
f(1,4) = 0.21	f(2,4) = 0.22	f(3,4) = **0.53**

Aggregált F-measure:

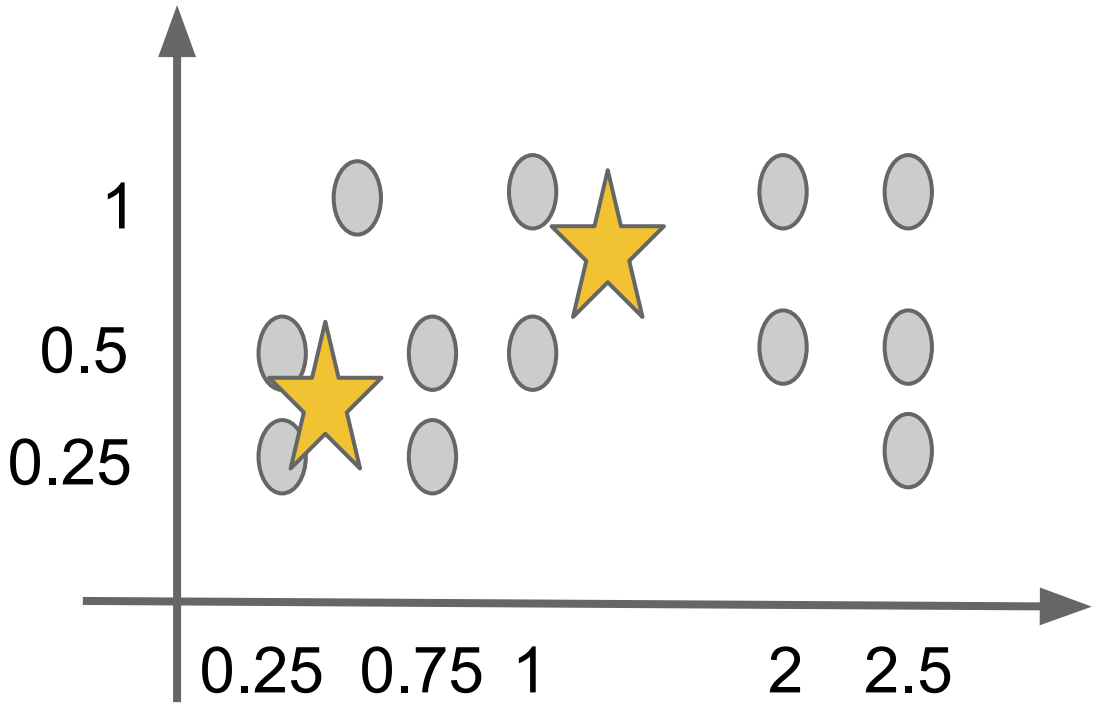
F = (90/175)***0.57** + (40/175)***0.44** + (45/175)***0.53** = **0.53**

- minden osztályra megnézzük, hogy melyik klaszter adja a maximális F-measure-t
- súlyozottan átlagolunk az osztályok mérete alapján

(lásd még: Milos Radovanovic: Representations and Metrics in High-Dimensional Data Mining, Novi Sad, 2011)

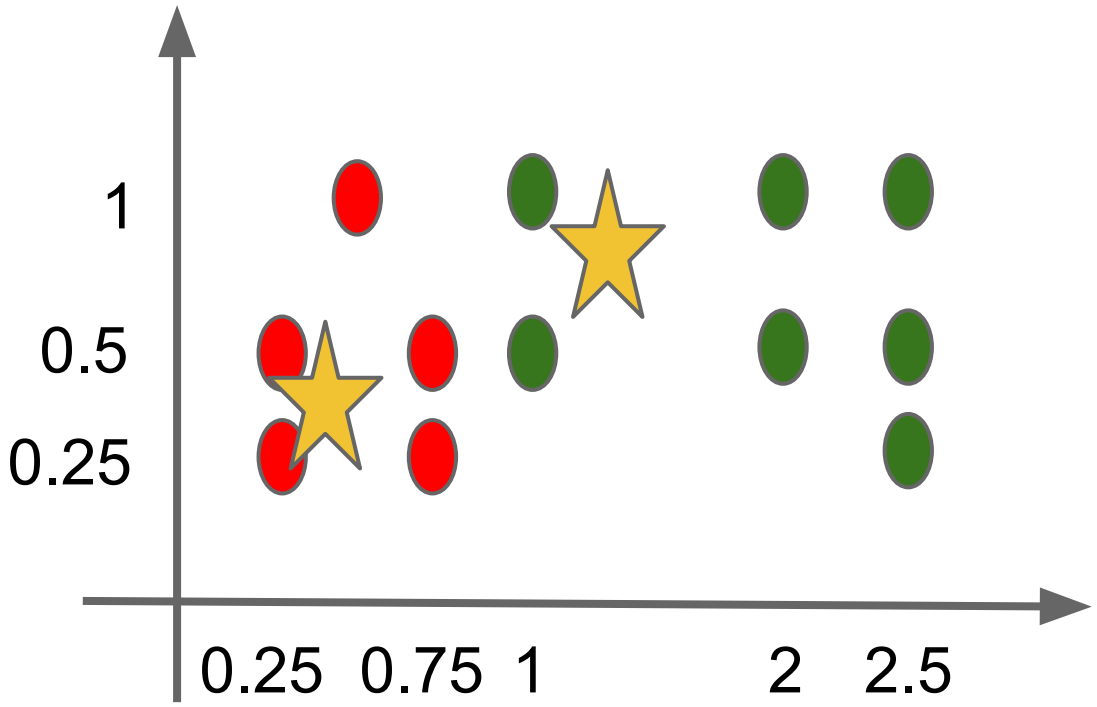
k-Means

Feladat: A k-Means i-dik iterációja után az ábrán jelölt klaszterközéppontok adódtak. Hajtsa végre az i+1-dik iterációt, és adja meg az i+1-dik iteráció végén az új klaszterközéppontokat!



k-Means

Megoldás: 1. lépés: minden pontot (példányt, rekordot...) a hozzá legközelebbi klaszterközépponthoz rendelünk hozzá.



k-Means

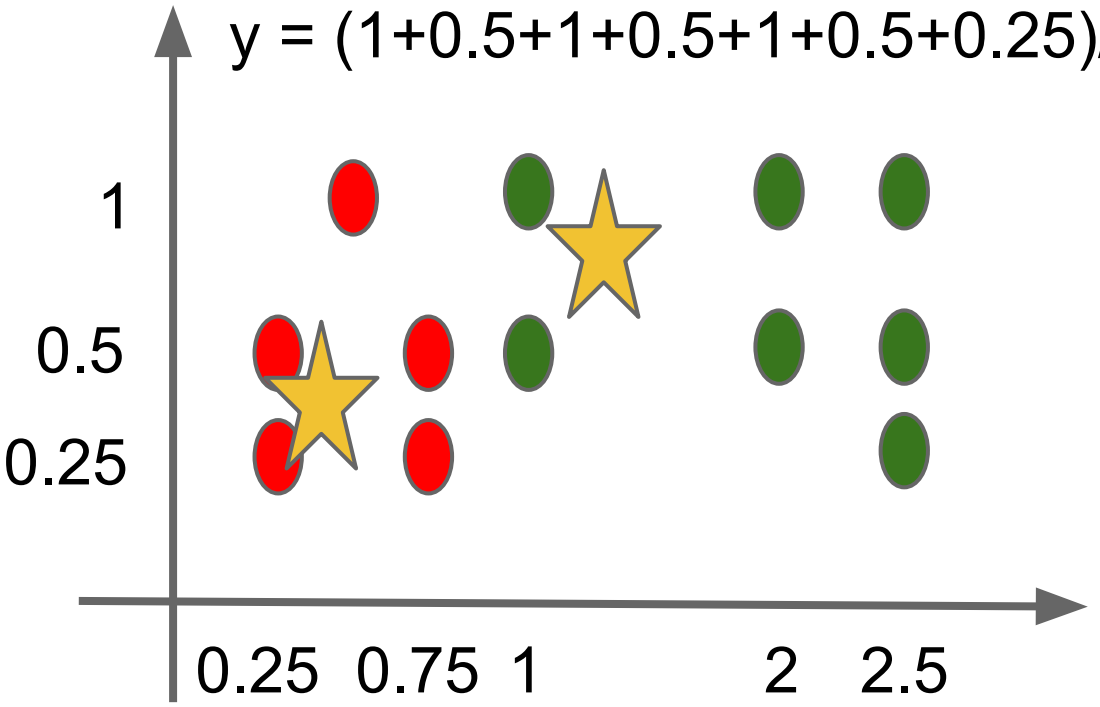
Megoldás: 2. lépés: kiszámoljuk az új középpontokat

Piros klaszer: x = (0.25+0.25+0.5+0.75+0.75)/5 = 0.5

y = (0.25+0.5+1+0.25+0.5)/5=0.5

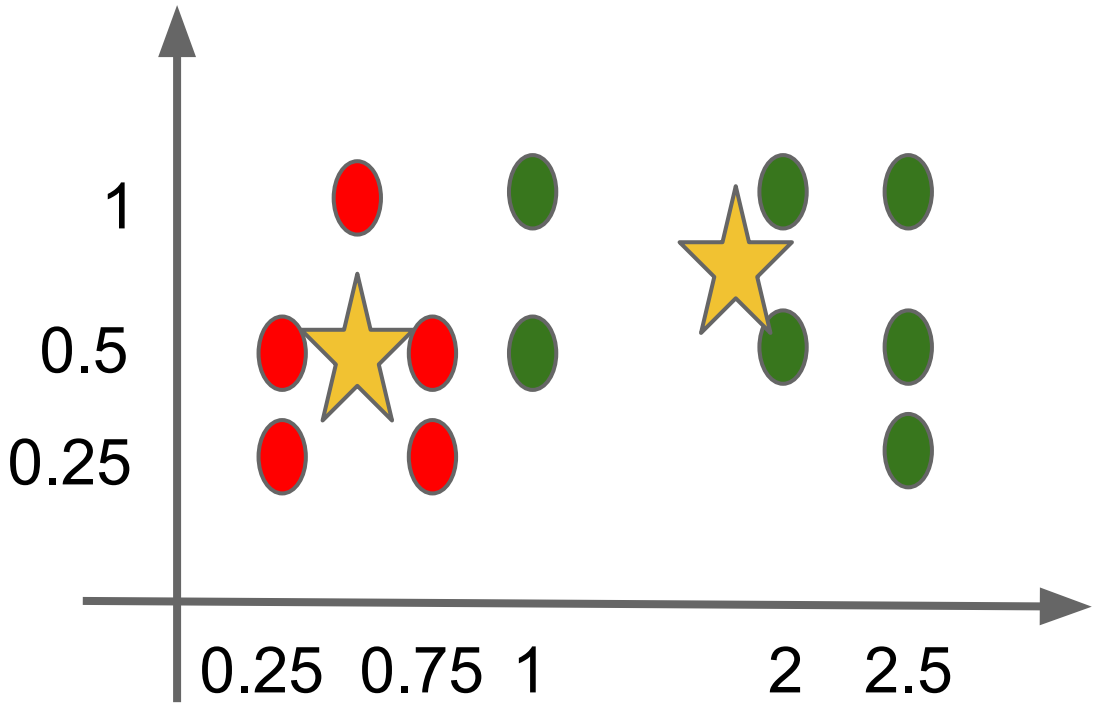
Zöld klaszter: x = (1+1+2+2+2.5+2.5+2.5)/7=1.92

y = (1+0.5+1+0.5+1+0.5+0.25)/7=0.68



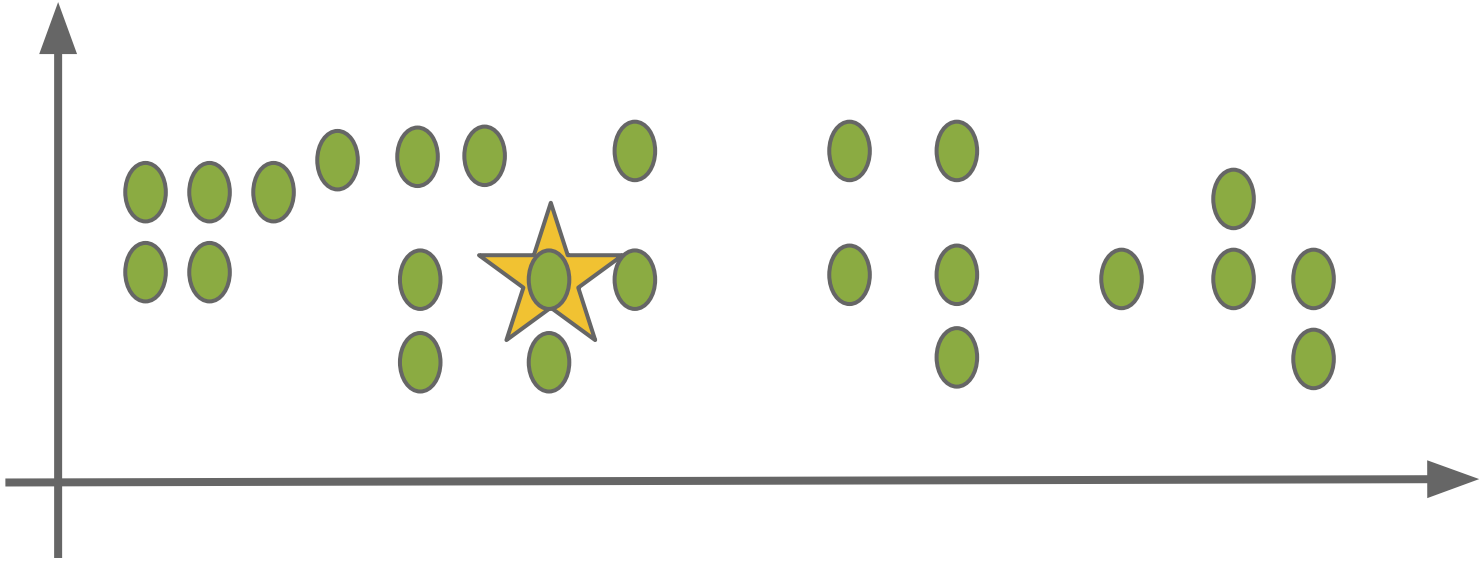
k-Means

Új klaszterközéppontok:



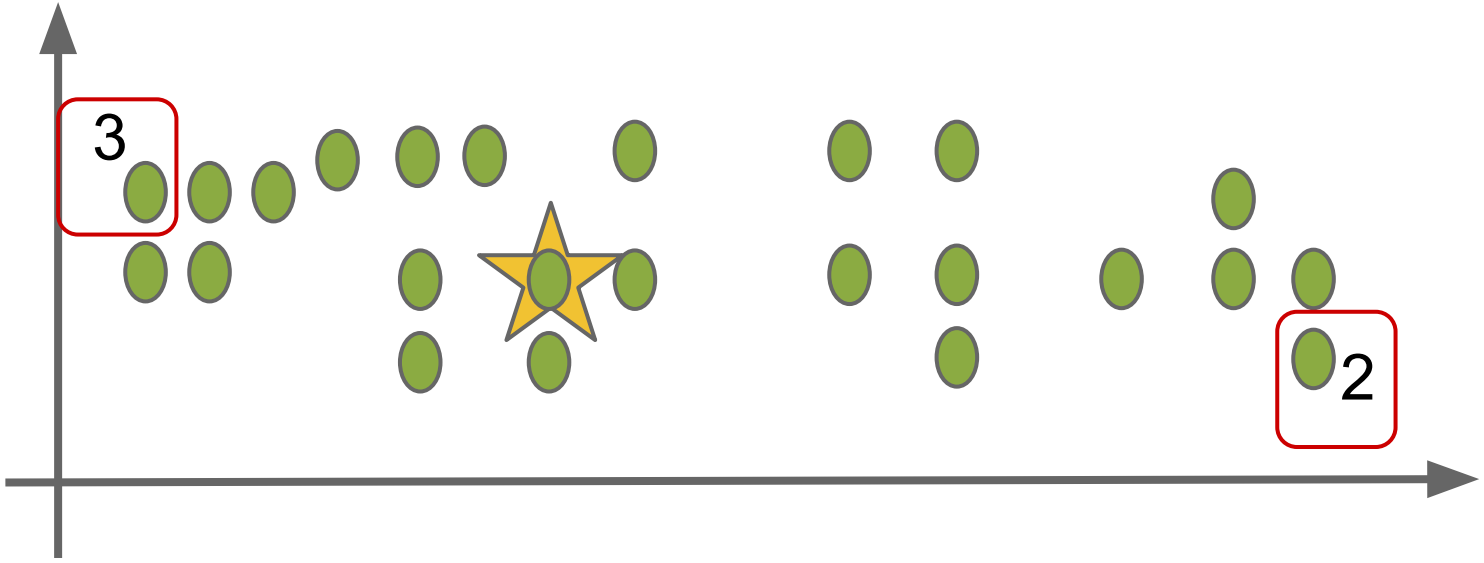
Furthest First

Feladat: A Furthest First algoritmussal 3 klasztert keresünk. Az adatpéldányok egy 2-dimenziós tér pontjainak feleltethetők meg (2 numerikus attribútummal rendelkeznek). Az első klaszterközéppontnak az ábrán jelölt pontot választottuk. Adja meg a FurthestFirst által választott további klaszterközéppontokat!



Furthest First

Feladat: A Furthest First algoritmussal 3 klasztert keresünk. Az adatpéldányok egy 2-dimenziós tér pontjainak feleltethetők meg (2 numerikus attribútummal rendelkeznek). Az első klaszterközéppontnak az ábrán jelölt pontot választottuk. Adja meg a FurthestFirst által választott további klaszterközéppontokat!



Hierarchikus klaszterező

Feladat: Adott az alábbi távolság mátrix. Mutassa meg egy dendogram segítségével, hogyan klaszterezi a példányokat a Single Linkage, Complete Linkage és Average Linkage!

	1	2	3	4	5
1	0	0.5	1.5	2	6
2	0.5	0	1	4	7
3	1.5	1	0	5.5	6
4	2	4	5.5	0	2
5	6	7	6	2	0

Hierarchikus klaszterező

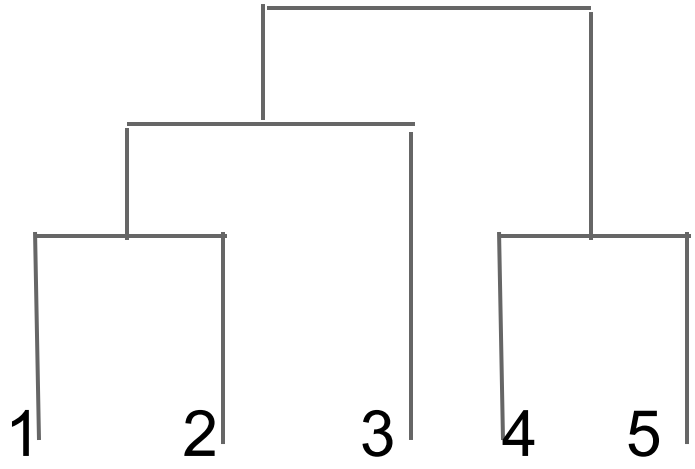
- Kezdetben minden példány külön klaszter
- Iteratíve mindig a két leghasonlóbb klasztert egyesíti

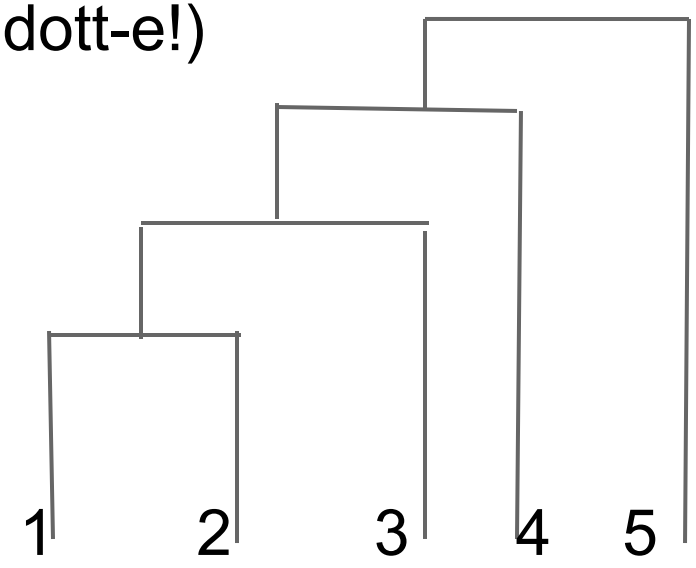
  Leállási feltétel:

  hasonlósági küszöbszám vagy klaszterek száma
- Két klaszter hasonlósága:
  - Single Link: leghasonlóbb elempár alapján
  - Complete Link: legkevésbé hasonló elempár alapján
  - Average Link: páronkénti hasonlóságok/távolságok átlaga

Hierarchikus klaszterező

A feladatban távolság mátrix adott: hasonló elempárok távolsága kicsi, különbözőké nagy. (Figyeljünk, hogy hasonlósági vagy távolság mátrix adott-e!)





Single link: jobboldai vagy baloldali
Average link: baloldali
Complete link: baloldali

Asszociációs szabályok

Feladat: adott az alábbi tranzakciós adatbázis. Számolja ki a {csavarhúzó, kalapács} elemhalmaz támogatottságát (support-ját)!

TID	Elemek
1	csavarhúzó, kalapács, vezeték
2	csavarhúzó, tégla, csempe
3	csavarhúzó, szög, kalapács
4	csavarhúzó, alátét, csavar
5	csavarhúzó, szög
6	szög, kalapács, vezeték

Asszociációs szabályok

Megoldás: {csavarhúzó, kalapács} elemhalmaz

abszolút támogatottsága: 2

relatív támogatottsága: 2/6 = 0.33

TID	Elemek
1	csavarhúzó, kalapács, vezeték
2	csavarhúzó, tégla, csempe
3	csavarhúzó, szög, kalapács
4	csavarhúzó, alátét, csavar
5	csavarhúzó, szög
6	szög, kalapács, vezeték

Asszociációs szabályok

Feladat: adott az alábbi tranzakciós adatbázis. Számolja ki a {csavarhúzó, kalapács} $\rightarrow$ {szög} asszociációs szabály támogatottságát, konfidenciáját és lift-mutatóját!

TID	Elemek
1	csavarhúzó, kalapács, vezeték
2	csavarhúzó, tégla, csempe
3	csavarhúzó, szög, kalapács
4	csavarhúzó, alátét, csavar
5	csavarhúzó, szög
6	szög, kalapács, vezeték

Asszociációs szabályok

Megoldás:

{csavarhúzó, kalapács} → {szög}

támogatottsága:
{csavarhúzó, kalapács, szög} támogatottsága, azaz 1 ill. 1/6 (abszolút ill. relatív támogatottság)

konfidencia: 1 / 2 = 0.5

(2-szer fordul elő {csavarhúzó, kalapács}, de ezek közül csak egy tranzakcióban szerepel szög)

lift-mutató:
(1/6) / ((2/6)*(3/6)) = 1

TID	Elemek
1	csavarhúzó, kalapács, vezeték
2	csavarhúzó, tégla, csempe
3	csavarhúzó, szög, kalapács
4	csavarhúzó, alátét, csavar
5	csavarhúzó, szög
6	szög, kalapács, vezeték

Asszociációs szabályok

Ábrázolja prefx-fában az elöbbi tranzakciós adatbázist. Az elemek sorrendezésnek válassza

a) az ABC-szerinti sorrend fordítottját

b) az alábbit: csavarhúzó, tégla, vezeték, kalapács, csempe, szög, alátét