2021. szeptember 14.

"Jó klaszterező algoritmus nem létezik"

<u>Jelölés</u>

Legyen $\mathcal{C}$ egy klaszterezése X-nek és legyenek $u, v \in X$ különbözők

$u \sim_{\mathcal{C}} v$ ha u és v ugyanabban a $\mathcal{C}$-beli klaszterben vannak

$u \not\sim_{\mathcal{C}} v$ ha különbözőkben

Az X halmazon a d függvényt távolságnak nevezzük, ha

$$d(x,y) \geq 0$$

$$d(x,y) = 0 \iff x = y$$

$$d(x,y) = d(y,x)$$

alapból nem követeljük meg a háromszög-egyenlőtlenséget

$$[\ d(x,y) + d(y,z) \geq d(x,z)\ ]$$

Klaszterezés függvény

(szándékosan nem algoritmus)

Ez egy olyan függvény, amely tetszőleges (X, d) párhoz az X egy $\mathcal{C}$ klaszterezését rendeli

Néhány viszonylag természetes elvárás egy épkézláb klaszterezési függvénytől:

Jelölje F azt a bizonyos épkézláb klasz-terezési függvényt

① skála-invariancia

tetszőleges (X, d) párra és $\lambda > 0$ valós számra

$$F(X, d) = F(X, \lambda d)$$

② konzisztencia

$F(X, d) = F(X, d')$

akkor az esetben ha $\mathcal{C} = F(X, d)$

és $d(u, v) \geq d'(u, v)$ $u \sim_{\mathcal{C}} v$ esetén

$d(u, v) \leq d'(u, v)$ $u \not\sim_{\mathcal{C}} v$ esetén

mindig

(3) teljesség

Tetszőleg $\mathcal{C}$ klaszterezéshez van olyan

d távolságfüggvény, hogy $F(X, d) = \mathcal{C}$

<u>Kleinberg lehetetlenségi tétel</u>

Nem létezik olyan klaszterezés függvény, amely skála-invariáns, konzisztens és teljes

(igazából olyan se ... ld. biz.)

Bizonyítás

Kicsit többet bizonyítunk:
Ha F skála-invariáns és konzisztens, akkor az értékkészlete nem tartalmazhat két olyan klaszterezést, amelyek közül az egyik valódi finomítása másiknak.
(egy klaszterezés akkor finomítása egy másiknak, ha ez utóbbi klaszterei előállnak mind

az előbbi kiosztásainak uniójaként)

Indirekt módon finomítunk

Tfh létezik olyan F skála-invariáns és konzisztens klaszterezési fv., amelynek értékkészlete tartalmaz két olyan klaszterezést, hogy az egyik valódi finomítása a másiknak.

Precízebben az utóbbi:

Valamely X legalább kételemű halmazra, és X-en értelmezett d_1 és d_2 távolság-függvényekre mondjuk $\mathcal{C}_1 = F(X, d_1)$ valódi finomítása $\mathcal{C}_2 = F(X, d_2)$-nek.

Bevezetünk néhány mennyiséget:

$$m_1 = \min \{ d_1(x_i, x_j) \mid x_i \sim_{\varepsilon_1} x_j \}$$

$$M_1 = \max \{ d_1(x_i, x_j) \mid x_i \nsim_{\varepsilon_1} x_j \}$$

$$m_2 = \min \{ d_2(x_i, x_j) \mid x_i \sim_{\varepsilon_2} x_j \}$$

$$M_2 = \max \{ d_2(x_i, x_j \mid x_i \nsim_{\varepsilon_2} x_j \}$$

Legyenek még $a_1 < b_1$ és $a_2 < b_2$ olyanok:

$$a_1 \leq m_1, \quad b_1 \geq M_1, \quad a_2 \leq m_2, \quad b_2 \geq M_2$$

Bevezetünk egy új d távolságfüggvényt:

$$d(x_i, x_j) = \begin{cases} \frac{a_1 a_2}{b_1} \left(\leq a_2\right) & x_i \sim_{e_1} x_j \\ a_2 & x_i \sim_{e_2} x_j \quad x_i \not\sim_{e_1} x_j \\ b_2 & x_i \not\sim_{e_2} x_j \end{cases}$$

F konzisztens $\Rightarrow$ $F(X, d) = F(X, d_2) = e_2$

Még egy d' távolságfüggvény jön:

$$d' = \frac{b_1}{a_2} d$$

F skálainvariáns $\Rightarrow$ $\boxed{F(X, d') = F(X, d) = \mathcal{C}_2}$

Másrészt

ha $x_i \sim_{\mathcal{C}_1} x_j$, akkor

$$d'(x_i, x_j) = \frac{b_1}{a_2} d(x_i, x_j) = \frac{\cancel{b_1}}{\cancel{a_2}} \frac{a_1 \cancel{a_2}}{\cancel{b_1}} = a_1$$

ha pedig $x_i \nsim_{\mathcal{C}_1} x_j$, akkor

$$d'(x_i, x_j) = \frac{b_1}{a_2} d(x_i, x_j) \geq \frac{b_1}{a_2} a_2 = b_1$$

F konzisztens $\Rightarrow$ $\boxed{F(X, d') = F(X, d_1) = \mathcal{C}_1}$

A két beszerezett összefüggés egyszerre nem állhat fenn csak ha $\mathcal{C}_1 = \mathcal{C}_2$ már pedig ez nem igaz hiszen $\mathcal{C}_1$ valódi finomítása $\mathcal{C}_2$-nek

Megjegyzés

A három feltételből bármely kettő kielégíthető:

Algoritmus (három is)

Definiáljunk egy súlyozott élű teljes gráfot az objektumok halmazán, ahol az élsúlyok az objektumok távolságai, és

futtassuk ezen a gráfon a Kruskal algo-
ritmust de ne végig, hanem

(A) csak addig, amíg a komponensek
száma el nem ér egy előre adott
k értéket → konzisztens, skálainvari-
áns

(B) csak addig, amíg a két legközelebbi
komponens között a min távolság

nem nagyobb, mint egy adott r pozitív szám $\rightarrow$ teljes, konzisztens

Ⓒ csak addig, amíg a két legközelebbi komponens között a min. távolság nem nagyobb, mint αM ahol M az X átmérője (legnagyobb távolság) és $\alpha < 1$ rögzített poz. szám

$\rightarrow$ teljes, skálainvariáns