MTA SZTAKI

Azonositási kódok statisztikai vizsgálata

Garádi János

A kórházakban ápolt személyek azonositására általában a születési év, hó, nap, a nem, valamint az anya nevének kezdőbetüje szolgál.

Ezek az adatok jól használhatók, mivel nem változnak meg az ember élete során. A vizsgálatok azonban azt mutatták, hogy ezek az adatok önmagukban a személyek csak mintegy 30 %-át azonositják egyértelmüen. Ezért a fenti azonositó kódhoz hozzávehetjük az állandó lakóhely megyéjét és a település jellegét (pl.: falu, város, illetve Budapesten a kerület). Ezek az adatok egy éven belül a lakosság mintegy 2 %-ánál változnak meg. Ezen felül az olyan kódokat is hozzávehetjük, amelyeknek a változása a lakosságnak legfeljebb 3 %-át érinti.

A statisztikai vizsgálathoz modellként az ugynevezett cellabetöltési problémát használjuk.

Ebben a dolgozatban egy cellabetöltési problémával foglalkozunk, amely az összetartozó rekordoknak (emberek) véletlenszerüen kisorsolt azonositó számok segitségével történő azonositása során merült fel.

A mi esetünk annyiban különbözik a klasszikus cellabetöltési problémától, hogy a különböző cellákba esés valószinüsége különböző.

A feladat leirása a következő:

Adott N számu ember, és n számu különböző azonositószám,

melyeket az N ember egymástól függetlenül $p_1, p_2 \ldots p_n$ valószínüséggel vesz fel (ahol n N).

Kérdés:

Mennyi a hibásan azonositott emberek várható száma? (Hibásan azonositottnak akkor tekintünk egy embert, ha található hozzá legalább még egy ember, amelynek ugyanaz az azonositója.)

Legyen az $\xi_i^{(m)}$ valószinüségi változó jelentése a következő :

$$\xi_i^{(m)} = \begin{cases} 1, & \text{ha az i-edik azonositó pontosan} \\ & m \text{ emberhez tartozik} \\ 0, & \text{különben} \end{cases} \qquad /1.1/$$

A $P(\xi_i^{(m)} = 1)$ valószinüség könnyen kiszámitható:

$$P(\xi_i^{(m)} = 1) = \binom{N}{m} p_i^m (1 - p_i)^{N-m} \qquad /1.2/$$

mert az egyes emberekhez egymástól függetlenül rendeltük az azonositókat.

Azoknak az azonositóknak az E_m várható száma, amelyhez pontosan m ember tartozik egyenlő:

$$E_m = \sum_{i=1}^{n} \binom{N}{m} p_i^m (1 - p_i)^{N-m} \qquad /1.3/$$

Konkrétan:

$$E_o = \sum_{i=1}^{n} (1 - p_i)^N$$

$$E_1 = N \cdot \sum_{i=1}^{n} p_i (1 - p_i)^{N-1} \qquad /1.4/$$

$$E_2 = \frac{N.(N-1)}{2!} \sum_{i=1}^{n} p_i^2 (1 - p_i)^{N-2}$$

$$E_3 = \frac{N.(N-1)(N-2)}{3!} \sum_{i=1}^{n} p_i^3 (1-p_i)^{N-3}$$

Vezessük be a következő jelölést: $p_i = \frac{\alpha_i}{N}$.

Fel kell tennünk, hogy $n \gg N$ - egyébként a rosszul azonosított emberek száma tulságosan nagy lenne (a konkrét feladatban $n \approx 150\ N$).

Emellett, a p_i valószinüségek, bár nem egyenlők, mégis egyenletesen kicsik, azaz ha $n \longrightarrow \infty$ és $n \cdot \max_i p_i$ egy független K korlát alatt marad. Ezért /1.4/ felhasználásával:

$$E_1 \approx \sum_{i=1}^{n} \alpha_i \cdot e^{-\alpha_i} \qquad /1.5/$$

Annak ellenére, hogy feltevésünk szerint:

$$\alpha_i < \frac{K}{n} \cdot N \approx 0 \,, \quad \text{az } e^{-\alpha_i}$$

nem helyettesithető egyszerüen 1-gyel, mert igy azt kapnánk, hogy a rekordok 100 %-a helyesen azonositható.

Az /1.4/ képletekből az előző feltevéseinket felhasználva, hogy a hibásan azonositott emberek H várható száma:

$$H = \sum_{m=2}^{\infty} m \cdot E_m$$

jól közelithető a $2 \cdot E_2$ -vel, mert a maradék:

$$\sum_{m=3}^{\infty} m \cdot E_m = \sigma\,(\alpha_i^2)$$

E_2 kiszámitásával már elhanyagolhatjuk az $e^{-\alpha_i}$ tényezőket, tehát:

$$H \approx 2 \cdot L_2 = 2 \cdot \frac{N.(N-1)}{2!} \sum_{i=1} \frac{\alpha_i^2}{N^2} \left(1 - \frac{\alpha_i}{N}\right)^{N-2} \approx$$

$$\approx \sum_{i=1}^{\infty} \alpha_i^2 e^{-\alpha_i} \approx \sum_{i=1}^{\infty} \alpha_i^2 + 0\left(\sum_{i=1} \alpha_i^3\right)$$

A $\sum_{i=1} p_i^2$ összeg a konkrét feladatnál könnyen kiszámitható, mert az azonositószámot 5 - a feladat természete miatt - egymástól független azonositó egyesitésével alakitottuk ki, ezért elegendő volt külön-külön megbecsülni ezen azonositok

$$p_{i_1}^{(1)}, p_{i_2}^{(2)}, \ldots, p_{i_5}^{(5)} \quad \text{valószinüségeit.}$$

Ekkor a függetlenség miatt:

$$\sum_{i=1}^{\infty} p_i^2 = \prod_{j=1}^{5} \sum_{i,j=1}^{\infty} \left(p_{ij}^{(j)}\right)^2$$

Térjünk át a $\xi^{(m)} = \sum_{i=1}^{n} \varepsilon_i^{(m)}$ valószinüségi változók szórás négyzetének kiszámitására.

$$D^2\left(\xi^{(m)}\right) = M\left(\xi^{(m)^2}\right) - M\left(\xi^{(m)}\right)^2 \qquad /1.6/$$

$$M\left(\xi^{(m)^2}\right) = M\left(\xi^{(m)}\right) + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} M\left(\varepsilon_i \varepsilon_j\right) =$$

$$= \sum_{i=1}^{n} \binom{N}{m} p_i^m (1-p_i)^{N-m} + \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \binom{N}{m}\binom{N-m}{m} \cdot p_i^m \cdot \qquad /1.7/$$

$$\cdot (1-p_i)^m \cdot p_j^m \cdot (1 - p_i - p_j)^{N-2m}$$

itt természetesen feltételeztük, hogy $N > 2.m$.

$$M\left(\xi^{(m)}\right)^2 = \sum_{i,j=1}^{n} \binom{N}{m}^2 p_i^m p_j^m (1-p_i)^{N-m} (1-p_j)^{N-m}$$

Könnyen belátható, a következő állitás:

Ha $m \geq 2$, és a $\max_i \alpha_i \longrightarrow 0$, hogy eközben

$$\frac{\max_i \alpha_i}{\min_i \alpha_i}$$

korlát alatt marad, akkor a

$$\frac{D^2\left(\xi^{(m)}\right)}{M\left(\xi^{(m)}\right)} \longrightarrow 1. \qquad /1.8/$$

Bizonyitás:

Azt kell belátnunk, hogy $\dfrac{D^2\left(\xi^{(m)}\right) - M\left(\xi^{(m)}\right)}{M\left(\xi^{(m)}\right)} \longrightarrow 0$

Az /1.6/ és az /1.7/ alapján:

$$D^2\left(\xi^{(m)}\right) - M\left(\xi^{(m)}\right) = \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \binom{N}{m}\binom{N-m}{m} p_i^m (1-p_i)^m \cdot$$

$$\cdot\ p_j^m (1-p_i-p_j)^{N-2m} - \sum_{i,j=1}^{n} \binom{N}{m}^2 p_i^m (1-p_i)^m \cdot \qquad /1.9/$$

$$\cdot\ p_j^m (1-p_i-p_j+p_i \cdot p_j)^{N-m}$$

Először vizsgáljuk az $i = j$ esetet: Ilyen tagból n darab van a kivonandóban. Könnyen belátható, hogy összegük rendje:

$$0\left(n \cdot \max_i \alpha_i^{2m}\right)$$

Ha az $i \neq j$, akkor az /1.9/-ben a kivonást tagonként elvégezve, és alkalmazva a binomiális tételt nyerjük, hogy egy tag rendje :

$$0\left(\frac{1}{n} \cdot \max_i \alpha_i^{2m-1}\right)$$

Tehát az összeg rendje:

$$0\left(n \cdot \max_i \alpha_i^{2m-1}\right)$$

Igy adódik, hogy az $\frac{1}{M\left(\xi^{(m)}\right)}$ rendje pedig : $0\left(\frac{1}{n \cdot \min_i \alpha_i^m}\right)$

Amiből következik állitásunk :

$$\frac{D^2\left(\xi^{(m)}\right) - M\left(\xi^{(m)}\right)}{M\left(\xi^{(m)}\right)} = 0\left(\max_i \alpha_i^{m-1}\right)$$

Q.E.D.

Számitásainkat két módon is alátámasztottuk számitogépes programok segitségével, reprezentativ mintavétel alapján.

Egyrészt közvetlen uton : sikerült találni a mintának egy olyan - viszonylag nagyméretü - részhalmazát, amelyben az összetartozó rekordok előfordulása logikailag lehetetlen volt. Ezen halmazon a logikailag megengedett azonositók száma 300-szorosa az azonositandó emberek számának. Ennek ellenére a rosszul azonositott emberek az összes emberek 5,5 %-át teszik ki. Ez 1100

olyan azonosítónak felel meg, amely pontosan 2 embert azonosít.

A szám tehát: $E\left(\xi^{(2)}\right) \approx 3t$.

A három vagy ennél több embert azonosító kódok száma már elenyészően kicsiny, azaz: 0.

A másik ellenőrzési mód közvetett: előző halmazunkhoz találtunk egy másik, vele kb. azonos méretű (N-elemű), tőle logikailag diszjunkt ember-halmazt, amelyen az azonosító kódok eloszlása (a p_i valószínüségek) - a probléma természetéből adódóan - ugyanaz.

Kérdés: mekkora a két halmaznak a hibás azonosításból adódó közös része.

A közös elemek várható száma könnyen meghatározható lenne, ha ismernénk az egyik halmaz által lefoglalt azonosítókhoz rendelt p_i valószinüségek összegét $P_{össz}$. A keresett várható érték ebben az esetben: $N \cdot P_{össz}$.

Ha csak az azonosítók eloszlását (a p_i valószinüségeket) ismerjük, de nem ismerjük, hogy az első halmaz konkrétan mely azonosítókat foglalja le, akkor a $P_{össz}$ várható értékét tudjuk kiszámitani.

$$P_{össz} = \sum_{i=1}^{n} p_i \, \varepsilon_i^{(1)} + \sum_{i=1}^{n} p_i \, \varepsilon_i^{(2)} + \dots$$

Az $\varepsilon_i^{(m)}$ definicióját lásd /1.1/-ben.

$$E\left(P_{össz}\right) = \binom{N}{1} \sum_{i=1}^{n} p_i^2 (1-p_i)^{N-1} + \binom{N}{2} \sum p_i^3 (1-p_i)^{N-2} + \dots$$

Tekintettel arra, hogy az $\alpha_i = p_i \cdot N$ számok nullához közeliek,

ezért elegendő csak az első tagot figyelembe venni, sőt az

$$(1 - p_i)^{i-1} \approx e^{-\alpha_i}$$

szorzót is elhanyagolhatjuk.

Így azt nyerjük, hogy:

$$L(P_{össz}) \approx \frac{2 \cdot E\left(\xi^{(2)}\right)}{N-1} ,$$

azaz a két halmaz hibás kódolásból adódó közös részének várható elemszáma kétszerese az egyik halmaz által kétszeresen lefoglalt kódok várható számának.

Az /1.8/ összefüggéshez hasonlóan belátható, hogy

$$D^2(P_{össz}) = N\left(\sum_{i=1}^{n} p_i^3 - \left(\sum_{i=1}^{n} p_i^2\right)^2\right) + O\left(\frac{N}{n^3}\right)$$

ha $n \longrightarrow \infty$ és $\max_i \alpha_i \longrightarrow 0$.

Ha a p_i-k egyenletességére nem kötünk ki semmilyen feltételt, akkor:

$$D^2(P_{össz}) = O\left(\frac{N}{n^2}\right) , \quad \text{ezért a}$$

$$D^2(N.P_{össz}) = O(n) .$$

Ugyanilyen nagyságrendű a közös rész elemszámának - a binomiális eloszlás alapján számított - szórásnégyzete, ha $P_{össz}$ ismert.

Fenti meggondolásaink érdekessége az, hogy a két halmazban közös azonositók elemeinek számára következtetni tudunk az egyik halmaz hibásan azonositott elemeinek számából - anélkül -, hogy ismernénk az azonositó kódok eloszlását.

A közös rész elemeinek tényleges (gépi uton nyert) száma : 1300.

A számitott értékeket ugy kaptuk, hogy különbözőnek tekintettük azokat a személyeket, akiknek állandó lakóhelye és foglalkozása különböző, majd megvizsgáltuk, hogy hány olyan 9 jegyü azonositó kód van, amely az ily módon különbözőnek tekintett személyek közül 1, 2, 3, stb. személyhez tartozik.

Megjegyezzük, hogy a belső vándorlás és az egy éven belüli foglalkozás változás - pl. nyugdijazás - viszonylag nagy száma miatt ez az eljárás nem teljesen korrekt, mégis hü képet nyujt az azonositó kódok statisztikai viselkedéséről.

A fentiek alapján a személyek azonositását egy 13 számjegyü azonositóval végeztük.