Asszociációs szabályok

Nikházy László

Nagy adathalmazok kezelése

2010. március 10.

Mi az értelme?

- A $pelenka \rightarrow sör$ asszociációs szabály azt állítja, hogy azon vásárlói kosarak, amik tartalmaznak pelenkát, általában tartalmaznak sört is.
- Ha ez igaz, akkor a szupermarket extra profithoz juthat az alábbi módon: Óriási hírverés közepette csökkentsük a pelenka árát (mondjuk 15%-kal), miközben diszkréten megemeljük a sör árát (mondjuk 30%-kal), úgy hogy a pelenka árcsökkentéséből adódó profitcsökkenés kisebb legyen a sör áremeléséből adódó profitnövekedésnél.

Definíció

- Asszociációs szabály:

$$R\colon I_1 \overset{c,s}{\rightarrow} I_2 \text{ , ahol } I_1 \cap I_2 = \emptyset \text{ , és:}$$

- bizonyosság:

$$c = \frac{supp(I_1 \cup I_2)}{supp(I_1)} = conf(R)$$

- támogatottság:

$$s = supp(I_1 \cup I_2)$$

- Érvényes asszociációs szabály:

$$s \geq min\_supp\ , \qquad c \geq min\_conf$$

Érvényes asszociációs szabályok meghatározása

- Minden I gyakori termékhalmazt bontsunk fel két diszjunkt, nem üres részre: $I = I_1 \cup I_2$
- Ellenőrizzük, hogy teljesül-e:

$$\frac{supp(I)}{supp(I_1)} \geq min\_conf$$

- Ha igen, akkor $I_1 \rightarrow I_2$ érvényes asszociációs szabály
- Észrevétel:

  ha $I_1 \rightarrow I \backslash I_1$ nem érvényes, és $I_1' \subseteq I_1$-nek, akkor ${I_1}' \rightarrow I \backslash {I_1}'$ sem érvényes

Maximális következményű asszociációs szabály

- Ha $I_1 \rightarrow I_2$ érvényes asszociációs szabály, akkor
  - $I_1 \rightarrow I_2'$ is érvényes, minden $I_2' \subseteq I_2$-re
  - $I_1 \cup \{i\} \rightarrow I_2 \backslash \{i\}$ is érvényes minden $i \in I_2$-re
- Tehát minden asszociációs szabály „levezethető” a maximális következményrésszel rendelkező asszociációs szabályokból

Probléma: büfé

- Emberek 1/3-a vesz hamburgert, 1/3-a hot-dogot, 1/3-a mindkettőt.
- A kosarak 66%-a tartalmaz hot-dogot, ezek 50%-a majonézt is.
- Így a hot-dog → majonéz érvényes lehet, ezért a büfés csökkenti a hot-dog árát és emeli a majonézét.
- A várakozással ellentétben azonban a profit csökken, mert a hamburger fogyasztók is inkább hot-dogot vesznek.

A probléma forrása

- A bizonyosság a következményrész feltételes valószínűségét próbálja becsülni:

$$c \approx p(I_2|I_1) = \frac{p(I_1, I_2)}{p(I_1)}$$

- Ha $p(I_2|I_1) = p(I_2)$, vagyis ha I_1 és I_2 függetlenek, akkor a szabály nem hordoz hasznos információt (de ezt a bizonyosság és a támogatottság nem feltétlenül mutatja)
- Ötlet: vizsgáljuk a $\frac{p(I_2|I_1)}{p(I_2)}$ hányadost!

  Valószínűségek helyett persze relatív gyakoriságokkal → lift érték

Lift érték

- Definíció:

$$lift(I_1 \rightarrow I_2) = \frac{freq(I_1 \cup I_2)}{freq(I_1) \cdot freq(I_2)}$$

- Például, ha lift(sör→pelenka)=2, az azt jelenti, hogy a sört vásárlók körében dupla annyi a pelenkát vásárlók aránya, mint amúgy általában

Empirikus kovariancia és korreláció

- empirikus kovariancia:

$$cov(I_1 \to I_2) = freq(I_1 \cup I_2) - freq(I_1) \cdot freq(I_2)$$

emlékeztető: X, Y valószínűségi változók

$$cov(X, Y) = E[(X - EX)(Y - EY)] = E[X \cdot Y] - EX \cdot EY$$

- empirikus korreláció:

$$corr(I_1 \to I_2) = \frac{cov(I_1 \to I_2)}{\sigma_{I_1}\sigma_{I_2}} = \frac{freq(I_1 \cup I_2) - freq(I_1) \cdot freq(I_2)}{\sqrt{EI_1(1 - EI_1)} \cdot \sqrt{EI_2(1 - EI_2)}}$$

$$= \frac{freq(I_1 \cup I_2) - freq(I_1) \cdot freq(I_2)}{freq(I_1) \cdot freq(\overline{I_1}) \cdot freq(I_2) \cdot freq(\overline{I_2})}$$

(valószínűségi változókra: $corr(X, Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$)

Kontingenciatábla

	X	nem X	Σ
Y	$k_{1,1}$	$k_{1,2}$	$k_{1.}$
nem Y	$k_{2,1}$	$k_{2,2}$	$k_{2.}$
Σ	$k_{.1}$	$k_{.2}$	n

- rendelkezésre álló értékek

	I_1	nem I_1	Σ
I_2	$supp(I_1 \cup I_2)$		$supp(I_2)$
nem I_2			
Σ	$supp(I_1)$		n

A hiányzó értékek számíthatók.

A χ^2-statisztika

- A $T_n = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(k_{ij} - \frac{k_{i.}k_{.j}}{n}\right)^2}{\frac{k_{i.}k_{.j}}{n}}$ próbastatisztika eloszlása aszimptotikusan χ^2 eloszlású lesz, ha X és Y függetlenek.
- (Valszám: Legyen K_ε olyan, hogy $P(\chi^2 < K_\varepsilon) = 1 - \varepsilon$, ekkor ha $T_n < K_\varepsilon$, akkor $1 - \varepsilon$ szignifikanciával függetlenek.)
- Minél kisebb a próbastatisztika, annál inkább függetlenek az események.
- 2x2-es esetben $T_n = n \cdot corr^2$

Binomiális próba

- Tfh. I_1 és I_2 függetlenek, $P(I_1, I_2) = P(I_1) \cdot P(I_2)$. Legyen $Z_j = I_{1j} \cdot I_{2j}$
- $Z = \sum_{j=1}^{n} Z_j$ binomiális eloszlású val. vált. n és $P(I_1, I_2)$ paraméterekkel,

  $P(I_1, I_2) \approx freq(I_1) \cdot freq(I_2)$
- megfigyelések: $(z_1, \dots, z_n)$ ellentmondanak-e ennek?
- legyen olyan a próba, hogy ha valójában függetlenek, akkor ε

  valószínűséggel mondjuk azt, hogy nem függetlenek:

  $[l, u]$: legszűkebb intervallum, melyre $\sum_{k=l}^{u} P(Z = k) \leq 1 - \varepsilon$

  ha $\sum_{j=1}^{n} z_j = z \in [l, u]$, akkor a hipotézisünk az, hogy függetlenek,

  egyébként az összefüggőség mellett döntünk

Fisher-féle egzakt próba

- Adott $n, supp(I_1), supp(I_2)$. Ha egyenletes eloszlás szerint vannak szétszórva I_1 és I_2 termékek a kosarakban, akkor mennyi az esélye annak, hogy az I_1-et tartalmazó kosarakból X darabban lesz I_2?

$$P\big(X, n, supp(I_1), supp(I_2)\big) = \frac{\binom{supp(I_1)}{X}\binom{n - supp(I_1)}{supp(I_2) - X}}{\binom{n}{supp(I_2)}}$$

- p-érték: az adott esetnél extrémebb esetek valószínűségének összege

$$p_F(I_1 \to I_2) = \sum_{X' : P(X', n, s(I_1), s(I_2)) \leq P(s(I_1 \cup I_2), n, s(I_1), s(I_2))} P(X', n, s(I_1), s(I_2))$$

- minél kisebb a p-érték, annál kisebb valószínűséggel függetlenek

Asszociációs szabályok rangsora

- A gyakorlatban sok érvényes szabályt találunk -> rangsorolni kellene
- Három paraméter: támogatottság, bizonyosság, függetlenség
  - pl. súlyok rendelése a paraméterekhez

    mi szerint?
    - marketinges: támogatottság
    - statisztikus: függetlenség
  - függetlenségre sok paraméter – egymáshoz hogy viszonyulnak?

    empirikus korreláció, χ^2-statisztika, p-érték: ugyanaz a sorrend

    empirikus kovariancia, lift érték: más sorrendeket adhatnak

Általánosság, specialitás

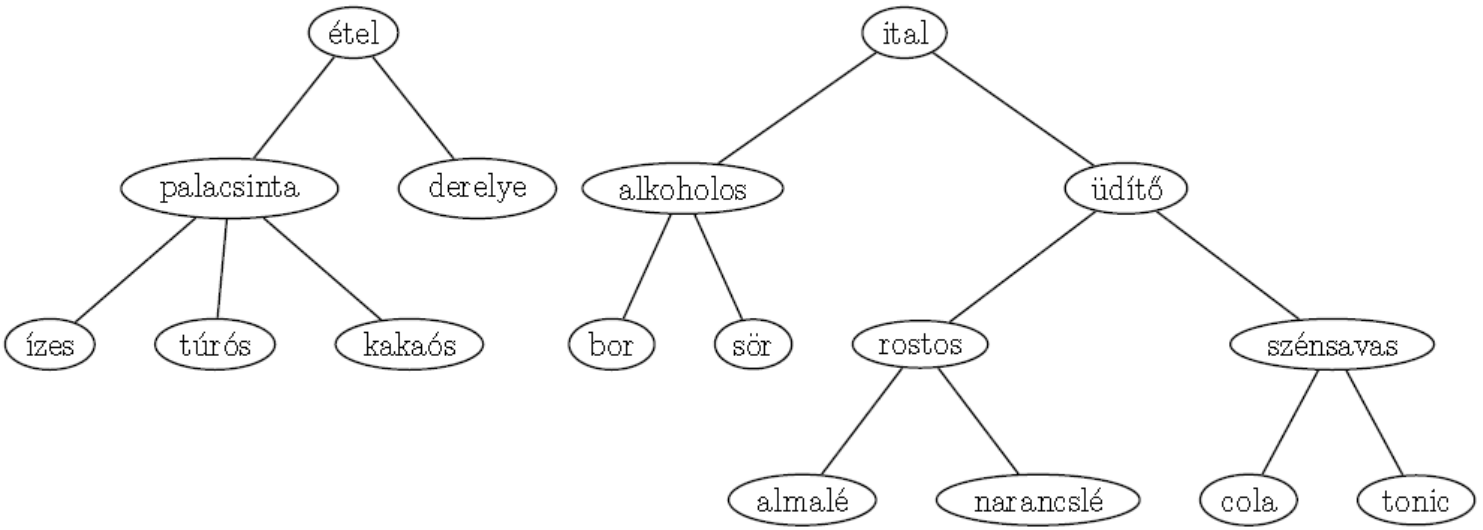
- Érdekes szabály mögé elbújva sok érdektelen szabály átmegy a szűrésen, és érdekesnek bizonyul.
- Legyen $I_1 \rightarrow I_2$ érvényes és érdekes, I_3egy olyan gyakori termékhalmaz, amely független I_1-től és I_2-től, és olyan nagy a támogatottsága, hogy $supp(I_1 \cup I_2 \cup I_3) \geq min\_supp$ is fennáll
- Ekkor könnyű belátni, hogy $I_1 \cup I_3 \rightarrow I_2$ is érvényes és érdekes asszociációs szabály lesz.
- A probléma kiküszöbölése: hagyjuk el a feltételrészből azt a részt, ami független a többi feltételtől és a következménytől is.

Hierarchikus asszociációs szabályok

- Előfordulhat, hogy termékkategóriák között vannak összefüggések

  Pl.: sört vásárlók 70%-a vesz valami chips-félét is

- Ismerni kell az elemek *taxonómiáját*
  - gyökeres, címkézett fa (fák)



Hierarchikus asszociációs szabályok

- Egy I kosár tartalmazza az I' elemhalmazt, ha

  $\forall i \in I' - re\ i \in I\ vagy\ \exists i' \in I, hogy\ i \in \H{o}s(i')$.

- Hierarchikus asszociációs szabály (def.):

  Legyen T a taxonómiában található termékek és kategóriák halmaza.

  $I_1 \overset{c,s}{\rightarrow} I_2$ hierarchikus asszociációs szabály, ha $I_1, I_2 \subseteq T,\ I_1 \cap I_2 = \emptyset$,

  továbbá egyetlen $i \in I_2$ sem őse egyetlen $i' \in I_1$-nek. c és s definíciója

  ugyanaz, mint a sima asszociációs szabálynál.

Hierarchikus asszoc. szabályok kinyerése

- Amikor a gyakori elemhalmazokat nyerjük ki (pl. apriori algoritmussal), akkor képzeletben töltsük fel a kosarakat az elemek ősével, amikor vizsgáljuk.
- Más megközelítés:

  kezdetben a gyökerekben található kategóriákkal határozzuk meg a gyakori elemhalmazokat, majd a következő lépésben vesszük a gyerekeiket stb.

Hierarchikus asszoc. szabályok érdekessége

- Lesznek semmitmondó szabályok:

  Pl.: élelmiszerbolt, háromféle (zsírszegény, félzsíros, normál) tej, az emberek egynegyede félzsíros tejet iszik, és:

$$tej \xrightarrow{80\%,4.8\%,2} zabpehely$$

$$zsírszegény\ tej \xrightarrow{80\%,1.2\%,2} zabpehely$$

- egy szabály nem érdekes, ha annak bizonyossága és támogatottsága nem tér el a nála általánosabb szabály paraméterei alapján becsült értékektől

Kategória asszociációs szabályok

- Ha az adatbázisban nem csak bináris attribútumok szerepelhetnek
- Minden olyan A attribútumot, amely k különböző értéket vehet fel ($k > 2$), helyettesítsünk k darab bináris attribútummal.
- Az így kapott bináris táblán már futtathatjuk a kedvenc asszociációs szabályokat kinyerő algoritmusunkat

A korreláció nem jelent implikációt

- Az asszociációs szabályok három paramétere közül egyik sem jelent okozatiságot
- Ha A és B között korreláció van, akkor lehet, hogy A okozza B-t, de lehet, hogy másféle kapcsolat áll fenn köztük. Az is lehet, hogy
  - B okozza A-t
  - egy harmadik C jelenség okozza A-t és B-t is

    pl.: „cipőben alvás fejfájást okoz”
  - A és B egymást is okozhatják kölcsönösen megerősítő módon
  - a korrelációt a véletlenek különös együttállása okozza (elsőfajú hiba)