

BEVEZETÉS

Mi az Apriori algoritmus?

- Egy adatbányászati algoritmus asszociációs szabályok keresésére.
- Azonosítja a gyakran együtt előforduló elemeket egy nagy adathalmazban.

Hol használják?

- Leggyakrabban: bevásárlókosár-elemzés (szupermarketek, e-kereskedelem).
- Csalásfelismerés (banki tranzakciók).
- Orvosi diagnosztika (tünetek és betegségek kapcsolatai).
- Weboldalak elemzése (milyen tartalmak érdeklik a felhasználókat).

ELŐNYÖK ÉS HÁTRÁNYOK

//

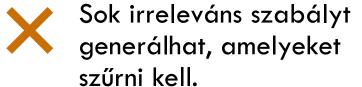
Egyszerű és könnyen érthető algoritmus.



Kiválóan alkalmas üzleti elemzésre (pl. ajánlórendszerek).



Nagy adathalmazoknál lassú, mert sok kombinációt kell kiszámítani.



ALAPFOGALMAK

Gyakori elemhalmazok: Azok az elemek, amelyek gyakran szerepelnek együtt a tranzakciókban.

Support: Egy elemhalmaz előfordulásának aránya az összes tranzakcióhoz képest.

$$support(A) = \frac{A - t \ tartalmaz\'o \ tranzakci\'ok \ sz\'ama}{\ddot{o}sszes \ tranzakci\'o \ sz\'ama}$$

Confidence: Ha A megtörténik, milyen gyakran követi B?

$$conf(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)}$$

Lift: A és B mennyire kapcsolódnak egymáshoz?

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{support(B)}$$

ALGORITMUS LÉPÉSEI



PÉLDA

Tranzakció sorszáma	Vásárolt termékek
T1	A, B, C
T2	A, B
Т3	A, B, C
T4	С
T5	A, C

Adottak a fenti tranzakciók. Keressük ki a ritka elemhalmazokat. Hozzuk létre az asszociációs szabályokat. Értékeljük a szabályokat confidence és lift alapján.

PARAMÉTEREK MEGVÁLASZTÁSA

Tranzakció sorszáma	Vásárolt termékek
T1	A, B, C
T2	A, B
Т3	A, B, C
T4	С
T5	A, C

Választott paraméterek:

- Minimum support érték: 50%, azaz a terméknek a tranzakciók legalább 3/5-ben szerepelnie kell.
- Minimum confidence érték: 70%.

GYAKORI ELEMHALMAZOK ELŐÁLLÍTÁSA

Tranzakció sorszáma	Vásárolt termékek
T1	A, B, C
T2	A, B
Т3	A, B, C
T4	С
T5	A, C

Számoljunk support értéket. Ami 0.5 alatt lenne, azt eldobnánk.

Termék	Support (%)
A	4/5 = 0.8 (80%)
В	3/5 = 0.6 (60%)
C	4/5 = 0.8 (80%)

GYAKORI ELEMHALMAZOK ELŐÁLLÍTÁSA

Tranzakció sorszáma	Vásárolt termékek
T1	A, B, C
T2	A, B
Т3	A, B, C
T4	C
T5	A, C

Számoljunk support értéket 2-es párokra. {B, C} nem üti meg a min. supportot, eldobjuk.

Termék	Support (%)
A, B	3/5 = 0.6 (60%)
A, C	3/5 = 0.6 (60%)
В, С	2/5 = 0.4 (40%)



GYAKORI ELEMHALMAZOK ELŐÁLLÍTÁSA

Tranzakció sorszáma	Vásárolt termékek
T1	A, B, C
T2	A, B
Т3	A, B, C
T4	С
T5	A, C

Számoljunk support értéket 3-as párokra. Mivel az egyetlen generálható érték az {A, B, C}, azonban {B, C} már nem gyakori, így garantált, hogy ez sem lesz az. Csak olyanokat generálunk, amelyek részhalmazai gyakoriak. Ha ilyet nem tudunk, megállunk.

Termék	Support (%)
A, B, C	2/5 = 0.4 (40%)

ASSZOCIÁCIÓS SZABÁLYOK GENERÁLÁSA

Mivel csak a 2-es csoportokban volt a min. *support*ot meghaladó csoport, így azokat vizsgáljuk.

1. A \rightarrow B:

$$conf(A \Rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{0.6}{0.8} = 0.75$$

2. A \rightarrow C:

$$conf(A \Rightarrow C) = \frac{support(A \cup C)}{support(A)} = \frac{0.6}{0.8} = 0.75$$

Meg lehet vizsgálni a fordított eseteket is $(B \rightarrow A)$, ebben az esetben a confidence 100% lesz. $C \rightarrow A$ esetén ugyanúgy 0.75 lesz.

Mindegyik meghaladta a minimális 70%-os confidence értéket.

VÉLETLEN EGYBEESÉS?

Hogy megbizonyosodjunk, valóban nem véletlenül alakultak így a vásárlások, tudunk *lift* értéket számolni. Azaz vizsgáljuk meg mennyire erős a kapcsolat. Ha a:

lift>1 \rightarrow akkor a két termék közötti kapcsolat **erősebb**, mint amit véletlenszerűen várnánk.

 $lift \le 1 \rightarrow$ nincs erős kapcsolat a két termék között: **függetlenek** egymástól.

Például:

$$lift(A\Rightarrow B)=rac{conf(A\Rightarrow B)}{support(B)}=rac{0.75}{0.6}=1.25>1$$
, tehát erős kapcsolat van közöttük.

$$lift(A\Rightarrow C)=rac{conf(A\Rightarrow C)}{support(C)}=rac{0.75}{0.8}=0.9<1$$
, tehát egymástól függetlenek, véletlenszerűek.