Asociacijos taisyklės

Asociacija ir klasifikacija

Asociacijos taisyklės yra panašios struktūros kaip ir klasifikavimo taisyklės. Tik šiuo atveju nėra akivaizdaus klasės kintamojo. Todėl tiek asociacijos taisyklės prielaida tiek jos išvada gali būti sudaryta iš daugelio atributų.

Asociacijos taisyklių konstravimas - nekontroliuojamo mokymo uždavinys.

Charakteringas asociacijų paieškos pavyzdys yra vadinamasis pirkėjo krepšelio uždavinys.

Pirkėjo krepšelis

Analizuojant prekybos centro pardavimų duomenis, stengiamasi nustatyti kokios prekės dažniausiai perkamos kartu.

Pavyzdys. Duomenys apie 5 pirkėjus

Nr.	duona	pienas	degtukai	alus	sūris	sultys
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Galima pastebėti, kad pirkėjai, perkantys degtukus, dažniausiai perka ir alų. Tokį pastebėjimą atspindinti asociacijos taisyklė yra

$$\{\text{degtukai}\} \longrightarrow \{\text{alus}\}\,.$$

Asociacijos taisyklė

Aibę $I = \{i_1, i_2, \ldots, i_n\}$ sudaro visi galimi nagrinėjamų duomenų elementai (prekių rūšys). Visų I poaibių aibę žymėsime $\mathcal{P}(I)$.
N įrašų imtis

$$T = \{t_1, t_2, \ldots, t_N\}, \quad t_i \in \mathcal{P}(I).$$

Įrašų, į kuriuos įeina elementų rinkinys X, skaičius $\sigma(X)$ vadinamas rinkinio X *dažniu* :

$$\sigma(X) = |\{t_i \mid X \subset t_i,\ t_i \in T\}| \ .$$

Apibrėžimas. *Asociacijos taisykle* $X \longrightarrow Y$ *vadinsime implikaciją*

$$X \subset t \Longrightarrow Y \subset t \, .$$

Čia $X, Y \in \mathcal{P}(I),\ X \cap Y = \emptyset;\ t$ *- bet kuris duomenų aibės įrašas.*

Asociacijos taisyklės apimtis ir tikslumas

Apibrėžimas. *Pagal imties T duomenis sukonstruotos asociacijos taisyklės*

$$r: \quad X \longrightarrow Y$$

apimtis $a(r)$ ir tikslumas $\theta(r)$ yra lygūs

$$a(r) = \frac{\sigma(X \cup Y)}{|T|},$$

$$\theta(r) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

Čia $|T| = N$ - įrašų skaičius imtyje T.

Asociacijos taisyklių konstravimo uždavinys

Pagal turimą imtį reikia rasti asociacijos taisykles r, kurių apimtis $a(r) \geq a_{\min}$ ir tikslumas $\theta(r) \geq \theta_{\min}$, čia $a_{\min}$ ir $\theta_{\min}$ yra iš anksto pasirinktieji apimties ir tikslumo rėžiai.

Jei imtyje sutinkamų elementų yra $|I| = n$, tai iš viso galima sukonstruoti

$$R(n) = 3^n - 2^{n+1} + 1$$

asociacijos taisyklių $X \longrightarrow Y$ su netuščiais elementų rinkiniais X ir Y. Pavyzdžiui, $R(20) = 3484687250$.

Dauguma algoritmų asociacijos taisykles generuoja dviem etapais:

Asociacijos taisyklių konstravimo uždavinys

Pagal turimą imtį reikia rasti asociacijos taisykles r, kurių apimtis $a(r) \geq a_{\min}$ ir tikslumas $\theta(r) \geq \theta_{\min}$, čia $a_{\min}$ ir $\theta_{\min}$ yra iš anksto pasirinktieji apimties ir tikslumo rėžiai.

Jei imtyje sutinkamų elementų yra $|I| = n$, tai iš viso galima sukonstruoti

$$R(n) = 3^n - 2^{n+1} + 1$$

asociacijos taisyklių $X \longrightarrow Y$ su netuščiais elementų rinkiniais X ir Y. Pavyzdžiui, $R(20) = 3484687250$.

Dauguma algoritmų asociacijos taisykles generuoja dviem etapais:

1. **Dažnų rinkinių radimas.** Randami visi elementų rinkiniai X, kurių santykinis dažnis imtyje T yra ne mažesnis už pasirinktą taisyklės apimties rėžį $a_{\min}$, t.y. $\sigma(X) \geq N \cdot a_{\min}$.

Asociacijos taisyklių konstravimo uždavinys

Pagal turimą imtį reikia rasti asociacijos taisykles r, kurių apimtis $a(r) \geq a_{\min}$ ir tikslumas $\theta(r) \geq \theta_{\min}$, čia $a_{\min}$ ir $\theta_{\min}$ yra iš anksto pasirinktieji apimties ir tikslumo rėžiai.

Jei imtyje sutinkamų elementų yra $|I| = n$, tai iš viso galima sukonstruoti

$$R(n) = 3^n - 2^{n+1} + 1$$

asociacijos taisyklių $X \longrightarrow Y$ su netuščiais elementų rinkiniais X ir Y. Pavyzdžiui, $R(20) = 3484687250$.

Dauguma algoritmų asociacijos taisykles generuoja dviem etapais:

1. **Dažnų rinkinių radimas**. Randami visi elementų rinkiniai X, kurių santykinis dažnis imtyje T yra ne mažesnis už pasirinktą taisyklės apimties rėžį $a_{\min}$, t.y. $\sigma(X) \geq N \cdot a_{\min}$.
2. **Taisyklių generavimas**. Iš rastųjų dažnų rinkinių konstruojamos asociacijos taisyklės, kurių tikslumas ne mažesnis už $\theta_{\min}$. Tokias taisykles vadinsime *tvirtomis*.

Apriori principas

Jei duomenų elementų aibės I dydis yra n, tai bet kuris iš $2^n - 1$ netuščių jos poaibių yra potencialus dažnas rinkinys. Visų jų dažnių radimui imtyje T reikėtų $O\left(|T|2^n \max_i |t_i|\right)$ palyginimo operacijų. Potencialių dažnų rinkinių skaičių galima sumažinti remiantis vadinamuoju *Apriori* principu. Jis labai paprastas.

***Apriori* principas**. *Bet kuris dažno elementų rinkinio poaibis taip pat yra dažnas.*

Jis išplaukia iš akivaizdžios dažnio *antimonotoniškumo* savybės.

Teorema. *Jei* $X \subset Y \subset I$, *tai* $\sigma(X) \geq \sigma(Y)$.

Tiesioginė *Apriori* principo išvada:

jei elementų rinkinys X *nėra dažnas, tai ir bet kuris jį apimantis rinkinys taip pat nėra dažnas.*

Apriori algoritmas dažnų rinkinių radimui

1. $k := 1$
2. $F_k := \{i \mid i \in I,\ \sigma(\{i\}) \geq N \cdot a_{\min}\}$ - randami dažni 1 elemento rinkiniai
3. **repeat**
4. $k = k + 1$
5. $C_k = \texttt{genApriori}(F_{k-1})$ - potencialūs dažni rinkiniai iš k elementų
6. **for** $\forall\, t \in T$ **do**
7. $C_k(t) = \{c \mid c \in C_k,\ c \subset t\}$ - įrašui t priklausantys C_k rinkiniai
8. **for** $\forall\, c \in C_k(t)$ **do**
9. $\sigma(c) = \sigma(c) + 1$ - dažnio prieaugis
10. **end for**
11. **end for**
12. $F_k = \{c \mid c \in C_k,\ \sigma(c) \geq N \cdot a_{\min}\}$ - dažni rinkiniai iš k elementų
13. **until** $F_k = \emptyset$
14. Rezultatas = $\cup F_k$

Funkcija $\texttt{genApriori}(F_{k-1})$ pagal turimą $k-1$ elemento dažnų rinkinių aibę F_{k-1}, remiantis *Apriori* principu, generuoja visus galimai dažnus k elementų rinkinius. Toliau (6 - 11 eilutės) skaičiuojami nustatytų kandidatų į dažnus rinkinius dažniai.

Apriori algoritmas dažnų rinkinių radimui

Pagrindiniai algoritmo efektyvumą įtakojantys faktoriai

1. **Apimties rėžis**. Didinant apimties rėžį, mažėja kandidatų į dažnus rinkinius ir dažnų rinkinių skaičius. Tuo pačiu greitėja ir pats algoritmas.
2. **Elementų skaičius (dimensija)**. Didesniam elementų skaičiui reikia daugiau sąnaudų jų dažnių skaičiavimui. Be to, didesnės dimensijos imtyse ir dažnų rinkinių gali būti daugiau.
3. **Imties dydis**. Skaičiuojant dažnius, yra tikrinami visi imties įrašai. Todėl akivaizdu, kad didinant įrašų skaičių, laiko sąnaudos didėja.
4. **Vidutinis įrašo dydis**. Įrašams ilgėjant, algoritmas lėtėja. Tai pasireiškia dvejopai. Visų pirma - ilgesni įrašai paprastai turi daugiau dažnų rinkinių. Antra - rinkinių paieška ilgesniuose įrašuose užima daugiau laiko.

Asociacijos taisyklių generavimas

Generuojamos asociacijos taisyklės r, sudarytos iš dažno k elementų rinkinio Y poaibių ir tenkinančios tikslumo sąlygą $\theta(r) \geq \theta_{\min}$

$$r: \quad X \longrightarrow Y \setminus X, \quad X \subset Y.$$

Jos apimtis ir tikslumas yra

$$a(r) = \frac{\sigma(Y)}{N}, \quad \theta(r) = \frac{\sigma(Y)}{\sigma(X)}.$$

Visi dažno rinkinio Y poaibiai taip pat yra dažni ir jų dažniai jau buvo apskaičiuoti pirmajame algoritmo etape, be to, $a(r) \geq a_{\min}$.
Jei $X \neq \emptyset$ ir $X \neq Y$, tokių taisyklių galima sudaryti $2^k - 2$. Šį kandidatų į tikslias taisykles skaičių galima sumažinti.

Asociacijos taisyklių generavimas

Teorema. *Jei* $A \subset B \subset C \subset I$ *, tai asociacijos taisyklė*

$$r_1 : \quad A \longrightarrow C \setminus A$$

yra ne tikslesnė už taisyklę

$$r_2 : \quad B \longrightarrow C \setminus B\,,$$

t.y., teisinga nelygybė $\theta(r_1) \leq \theta(r_2)$ *.*

Įrodymas. Asociacijos taisyklių r_1 ir r_2 tikslumai yra lygūs

$$\theta(r_1) = \frac{\sigma(C)}{\sigma(A)}\,, \quad \theta(r_2) = \frac{\sigma(C)}{\sigma(B)}\,.$$

Dėl dažnio antimonotoniškumo $\sigma(A) \geq \sigma(B)$. Todėl $\theta(r_1) \leq \theta(r_2)$.

□

Asociacijos taisyklių generavimo pavyzdys

Tarkime $Y = \{abcd\}$ yra dažnas rinkinys ir

$$\theta(\{acd\} \longrightarrow \{b\}) \geq \theta_{\min}\,, \quad \theta(\{abd\} \longrightarrow \{c\}) \geq \theta_{\min}\,,$$
$$\theta(\{bcd\} \longrightarrow \{a\}) < \theta_{\min}\,, \quad \theta(\{abc\} \longrightarrow \{d\}) < \theta_{\min}\,.$$

Galima sudaryti $\binom{4}{2} = 6$ taisykles, kurių išvadoje bus 2 elementų rinkinys. Pagal teoremą penkios iš šių taisyklių

$$\{cd\} \longrightarrow \{ab\}\,, \quad \{bd\} \longrightarrow \{ac\}\,, \quad \{bc\} \longrightarrow \{ad\}\,,$$
$$\{ac\} \longrightarrow \{bd\}\,, \quad \{ab\} \longrightarrow \{cd\}$$

nėra tikslios - jų tikslumas $\leq \theta_{\min}$. Tad lieka apskaičiuoti tik taisyklės

$$\{ad\} \longrightarrow \{bc\}$$

tikslumą.

"Blogos" asociacijos taisyklės pavyzdys

Didelėje elektronikos parduotuvėje iš 10000 pirkėjų 6000 pirko kompiuterinius žaidimus, 7500 pirko DVD filmus, o 4000 pirko ir žaidimus ir filmus. Pasirinkus apimties ir tikslumo rėžius $a_{\min} = 0,3$ ir $\theta_{\min} = 0,6$, buvo suformuluota asociacijos taisyklė:

$$r_1: \quad X = \{\text{kompiuteriniai žaidimai}\} \longrightarrow Y = \{\text{DVD filmai}\}$$

Ši taisyklė yra tvirta, nes

$$a(r_1) = \frac{4000}{10000} = 0,4 > 0,3 \quad \text{ir} \quad \theta(r_1) = \frac{4000}{6000} \approx 0,67 > 0,6\,.$$

Tačiau taisyklė r_1 nėra logiška. Iš tikrųjų, 75% pirkėjų perka DVD filmus. Bet tik 67% kompiuterinių žaidimų pirkėjų perka ir DVD filmus. Vadinasi prekės X buvimas krepšelyje net sumažina Y tikimybę !

Asociacijos taisyklės kokybės matai

Tegul asociacijos taisyklė $r: \quad X \longrightarrow Y$ gauta pagal imties $T = \{t_1, t_2, \ldots, t_N\}$ duomenis. Jos *įtakos faktorius* lygus

$$ITF(r) = ITF(X, Y) = \frac{N \cdot \theta(r)}{\sigma(Y)}\,.$$

Nesunku pastebėti, kad įtakos faktorius $ITF(X, Y)$ iš tikrųjų yra tikimybių santykio

$$\frac{P(Y \mid X)}{P(Y)}$$

įvertis. Todėl jį galima interpretuoti taip:

$$ITF(X, Y) \begin{cases} = 1, \text{jei } X \text{ ir } Y \text{ nepriklausomi;} \\ > 1, \text{jei tarp } X \text{ ir } Y \text{ yra tiesioginė koreliacija;} \\ < 1, \text{jei tarp } X \text{ ir } Y \text{ yra atvirkštinė koreliacija.} \end{cases}$$

Elektronikos parduotuvės pavyzdyje $ITF(r_1) \approx 0,89 < 1\,.$

Asociacijos taisyklės kokybės matai

Dažnai asociacijos taisyklės $r: \quad X \longrightarrow Y$ vertinimui naudojami binarinių vektorių **x** $= (x_1, x_2, \ldots, x_N)$ ir **y** $= (y_1, y_2, \ldots, y_N)$ panašumo koeficientai. Čia i - tosios koordinatės yra rinkinių X ir Y patekimo į įrašą $t_i \in T$ indikatoriai.

Gerą taisyklę atitinkantys vektoriai **x** ir **y** yra "panašūs", t.y. jų panašumo koeficientas turėtų būti kuo didesnis.

Sąryšį tarp **x** ir **y** nusako dažnių lentelė

$x_i \backslash y_i$	0	1	
0	k_{00}	k_{01}	k_{0+}
1	k_{10}	k_{11}	k_{1+}
	k_{+0}	k_{+1}	N

Pastebėsime, kad

$$k_{1+} = \sigma(X)\,, \quad k_{+1} = \sigma(Y)\,, \quad k_{11} = \sigma(X \cup Y)\,.$$

Asociacijos taisyklės kokybės matai

Pavadinimas	Apibrėžimas
Įtakos faktorius	$\frac{N \cdot k_{11}}{k_{1+}k_{+1}}$
Žakardo	$\frac{k_{11}}{k_{1+} + k_{+1} - k_{11}}$
Kosinusas	$\frac{k_{11}}{\sqrt{k_{1+}k_{+1}}}$
Koreliacijos	$\frac{k_{11}k_{00} - k_{10}k_{01}}{\sqrt{k_{1+}k_{+1}k_{0+}k_{+0}}}$