Asociačné pravidlá

Asociačné pravidlá sú jednou z najpopulárnejších metód hĺbkovej analýzy (data miningu). Úspešne sa aplikujú v analýze nákupného košíka, finančných dát, cenzusu a pod. Asociačné pravidlá nám umožňujú získať obraz o vzťahoch medzi prvkami v danej množine dát, pričom výsledky sa veľmi ľahko interpretujú. Pravidlá totiž predstavujú konštrukciu IF THEN, ktorú nájdeme vo všetkých programovacích jazykoch a dajú sa vyjadriť v prirodzenom jazyku. Asociačné pravidlá spopularizoval začiatkom 90. rokov Agraval v súvislosti s analýzou nákupného košíka [4]. Asociačné pravidlá sa radia medzi symbolické metódy strojového učenia a rovnako ako stromy sú šité na mieru kvalitatívnym/kategoriálnym dátam. Kvantitatívne/numerické dáta vyžadujú diskretizáciu - nahradenie numerických hodnôt intervalmi hodnôt, s ktorými sa potom narába ako s kategorickou premennou.

Nasledujúca časť ponúka teoretický rozbor asociačných pravidiel z matematického hľadiska.

Nech $D = \{T_1, T_2, \ldots, T_n\}$ je množina n transakcií a nech I je množina položiek, $I = \{i_1, i_2, \ldots, i_m\}$. Každá transakcia je množinou položiek, t. j. $T_i \subseteq I$. Asociačné pravidlo je implikácia v tvare $X \Rightarrow Y$, kde $X, Y \subset I$, a $X \cap Y = \varnothing$; X sa nazýva predpoklad/podmienka (antecedent/body) a Y záver/následok (consequent/head) pravidla. Vo všeobecnosti, množina položiek, ako napríklad X alebo Y, sa nazýva položková množina (itemset).
Príkladom T_i môže byť jeden nákup, pričom D predstavuje celú databázu, t. j. všetky nákupy za určité časové obdobie. Výsledkom analýzy sú pravidlá tvaru AK podmienka POTOM následok (IF body THEN head). Pravidlá sú určované na základe početností, s akými sa podmienka a následok vyskytujú v dátach.
Príklad asociačného pravidla z analýzy nákupného košíka:
$\{chlieb, syr\} \Rightarrow \{maslo\}$, *confidence* = 57%, *support* = 21%
Interpretácia pravidla je nasledovná: Zákazníci, ktorí si na jeden nákup kúpia chlieb a syr si s 57% pravdepodobnosťou kúpia aj maslo, pričom 21% nákupov obsahovalo chlieb, syr aj maslo.
Nech $P(X)$ je pravdepodobnosť výskytu položkovej množiny X v D a nech $P(Y|X)$ je podmienená pravdepodobnosť výskytu položkovej množiny Y, za podmienky výskytu položkovej množiny X.
Pre položkovú množinu $X \subseteq I$, *support*(X) je definovaná ako časť transakcií $T_i \in D$ takých, že $X \subseteq T_i$, t. j. $P(X)$ = *support*(X). Podpora (*support*) pravidla $X \Rightarrow Y$ je definovaná ako *support*($X \Rightarrow Y$) = $P(X \cup Y)$. Inak povedané pravidlo $X \Rightarrow Y$ má podporu s, ak s% transakcií v databáze obsahuje $X \cup Y$. V podstate podpora reprezentuje frekvenciu výskytu danej množiny položiek v databáze.
Ďalšou dôležitou charakteristikou asociačného pravidla $X \Rightarrow Y$ je miera reliability nazývaná spoľahlivosť (*confidence*), *confidence*($X \Rightarrow Y$) je definovaná ako $P(Y|X) = P(X \cup Y)/P(X)$ = *support*($X \cup Y$)/*support*(X). Inak povedané pravidlo $X \Rightarrow Y$ má spoľahlivosť c, ak c% transakcií v databáze, ktoré obsahujú položku X, taktiež obsahujú položku Y. Spoľahlivosť pravidla je pravdepodobnosť výskytu pravej strany pravidla za podmienky výskytu ľavej strany, t. j. je to percentuálny podiel pravidiel, ktorých ľavá strana je X a pravá Y zo všetkých, ktorých ľavá strana je X.
Ďalšou dôležitou charakteristikou asociačného pravidla $X \Rightarrow Y$ je miera zaujímavosti nazývaná zdvih (*lift*), *lift*($X \Rightarrow Y$) je definovaný ako *lift*($X \Rightarrow Y$) = $P(Y|X)/P(Y)$ = *confidence*($X \Rightarrow Y$)/*support*(Y). Táto miera určuje koľko krát častejšie sa X a Y vyskytujú spolu, než by to bolo, keby boli štatisticky nezávislé. Ak je miera *lift*($X \Rightarrow Y$) > 1 indikuje to, že sa X a Y vyskytujú častejšie spolu ako zvlášť.

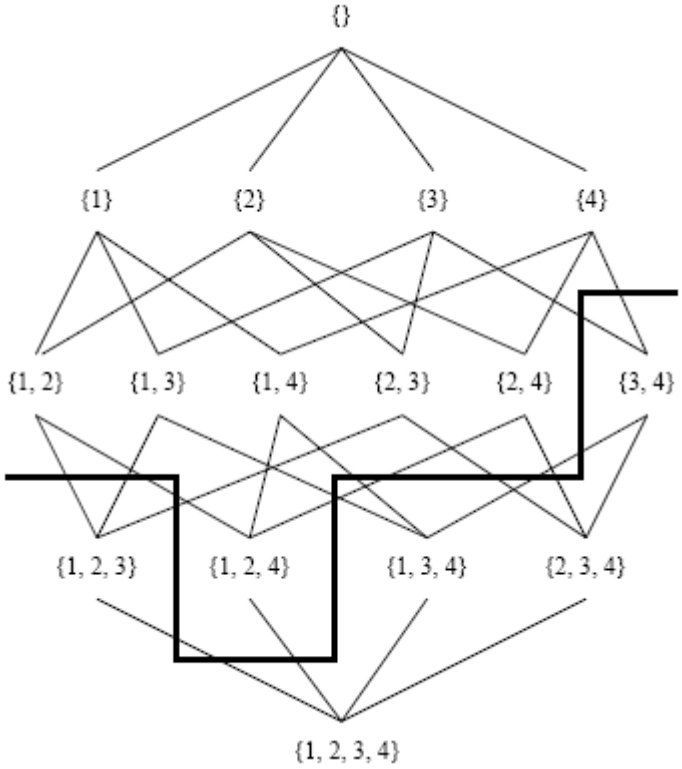
Cieľom asociačnej analýzy je nájsť všetky pravidlá, ktorých miery sú väčšie alebo rovné ako špecifikovaná podpora (*minimum support*) a spoľahlivosť (*minimum confidence*) [4], [5]. Tieto dve miery vypovedajú o početnosti výskytu pravidla v databáze (*support*) a o sile pravidla (*confidence*). K- položková množina s podporou väčšou ako určené minimum sa nazýva frekventovaná/veľká/často opakujúca sa množina položiek. Celý proces získavania pravidiel, definovaný nižšie, pozostáva z dvoch krokov, kde sa najskôr získajú všetky frekventované množiny položiek, z ktorých sa potom vygenerujú samotné pravidlá.

Definícia 1
1. Daná je množina položiek I, vstup pozostáva z množiny transakcií D, kde každá transakcia T je neprázdna podmnožina položiek množiny I, $T \subseteq I$.
2. Daná je množina položiek $T \subseteq I$ a množina transakcií D, definujeme podporu T ako $support(T) = P(T)$.
3. Nastavením minimálnej podpory $min\ s$, kde $0 \leq min\ s \leq 1$, definujeme frekventované množiny položiek ako množiny položiek kde $support(T) \geq min\ s$.

Definícia 2
1. Asociačné pravidlo je pár disjunktných množín položiek, predpokladu $X \subseteq I$ a záveru $Y \subseteq I$, kde $X \Rightarrow Y$ a $X \cap Y = \varnothing$.
2. Definujeme podporu asociačného pravidla $X \Rightarrow Y$ ako $support(X \Rightarrow Y) = P(X \cup Y)$.
3. Definujeme spoľahlivosť asociačného pravidla $X \Rightarrow Y$ ako $confidenceD(X \Rightarrow Y) = P(Y|X)$.
4. Nastavením minimálnej spoľahlivosti $min\ c$, kde $0 \leq min\ c \leq 1$, definujeme silné pravidlá ako pravidlá kde $confidence(X \Rightarrow Y) \geq min\ c$.

Najznámejší algoritmus Apriori navrhol Agrawal v súvislosti s analýzou nákupného košíka [6]. Algoritmus je založený na hľadaní už spomínaných frekventovaných položkových množín, ktoré predstavujú kombinácie/konjunkcie kategórií atribútov splňujúce podmienku minimálnej podpory.
Apriori algoritmus využíva generovanie kombinácii do šírky, t. j. najskôr sa vygenerujú kombinácie dĺžky jedna, potom všetky kombinácie dĺžky dva atd., pričom sa nesmú v kombinácii opakovať atribúty a položky sú usporiadané lexikograficky. Pri generovaní vlastne prehľadávame priestor všetkých prístupných kombinácií. Pri generovaní sa využíva vlastnosť frekventovaných položkových množín – každá podmnožina frekventovanej množiny položiek musí byť tiež frekventovaná, t. j. vo všeobecnosti na hľadanie kombinácii dĺžky k, sa využíva toho, že už poznáme $(k - 1)$ - prvkové frekventované množiny. Algoritmus nevyužívajúci túto vlastnosť frekventovaných množín by musel preskúmať všetkých 2^m možných podmnožín množiny položiek I, $I = \{i_1, i_2, \ldots, i_m\}$.
Z príkladu na obrázku (obr. 1) vidíme, že z kandidátov frekventovaných jednoprvkových množín $C1$ stanovenú minimálnu podporu spĺňajú všetky jednoprvkové množiny, t. j. $L1 = \{\{1\}, \{2\}, \{3\}, \{4\}\}$. Z tejto množiny je možné spájaním vygenerovať nasledovných kandidátov veľkosti 2, $C2 = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}\}$. Minimálnu stanovenú hodnotu podpory spĺňajú nasledovné dvojprvkové množiny $L2 = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}\}$. Z uvedenej množiny dvojprvkových frekventovaných množín položiek možno spájaním vygenerovať nasledovných trojprvkových kandidátov $C3 = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 3, 4\}, \{2, 3, 4\}\}$. Z tejto množiny orezávaním odstránime množiny $\{1, 3, 4\}$, $\{2, 3, 4\}$, vzhľadom na to, že ich podmnožina $\{3, 4\}$ sa nenachádza v $L2$, t. j. nie je frekventovaná. Minimálnu stanovenú hodnotu podpory spĺňa iba nasledovná trojprvková množina $L3 = \{\{1, 2, 4\}\}$. Z tejto množiny trojprvkových frekventovaných množín, už ale nie je možné

vygenerovať žiadneho štvorprvkového kandidáta, a tak algoritmus končí nájdením všetkých frekventovaných množín {*L*1, *L*2, *L*3} v *D*.



Obr. 1 *Hľadanie frekventovaných množín* [7]

Vo všeobecnosti by sme postup pri hľadaní frekventovaných množín mohli rozdeliť do troch krokov:
1. Vygenerovanie množiny kandidátov C_k na základe L_{k-1} – spájanie.
2. Odstránenie takých množín z C_k, ktorých podmnožina sa nenachádza v L_{k-1} – orezávanie.
3. Zaradenie takých množín z C_k do L_k, ktoré spĺňajú stanovenú minimálnu hodnotu podpory.

Po nájdení frekventovaných položkových množín, t .j. kombinácií, ktoré vyhovujú svojou početnosťou, sa vytvárajú pravidlá spĺňajúce podmienku minimálnej spoľahlivosti.
Každá kombinácia *T* sa rozdelí na všetky možné dvojice podkombinácií *X* a *Y* také, že $Y = T - X$. Na základe tejto skutočnosti podpora pravidla bude rovnaká ako podpora kombinácie, t. j. $support(X \Rightarrow Y) = support(T)$.
Spoľahlivosť vypočítame ako podiel početnosti kombinácie *T* a predpokladu *X*, $confidence(X \Rightarrow Y) = support(T)/support(X)$, pričom $support(X) \geq support(T)$, vzhľadom na to, že predpoklad *X* je skôr vygenerovaná frekventovaná položková množina, t. j. *X* je kratšia menej obmedzujúca kombinácia.
Ak *X'* je nadkombináciou kombinácie *X*, je *X'* prísnejšie, t. j. splní ho menej príkladov. Vzhľadom na to, že pri výpočte spoľahlivosti je početnosť *X'* v menovateli zlomku, pričom čitateľ zostáva rovnaký, je $confidence(X' \Rightarrow T - X') \geq confidence(X \Rightarrow T - X)$, t. j. ak nevyhovuje zadanej minimálnej spoľahlivosti pravidlo $X' \Rightarrow T - X'$, nevyhovuje ani žiadne pravidlo $X \Rightarrow T - X$.
Ak napr. pre kombináciu {1, 2, 4} pravidlo $\{1, 2\} \Rightarrow \{4\}$ nespĺňa podmienku minimálnej spoľahlivosti, potom ju nemôžu spĺňať ani pravidlá $\{1\} \Rightarrow \{2, 4\}$, $\{2\} \Rightarrow \{1, 4\}$ a nemusia sa teda vôbec uvažovať.

Z hore uvedeného vyplýva, že kombinácie *T* začíname rozdeľovať tak, že najskôr záver *Y* je tvorený kombináciou dĺžky 1 (položková množina *Y* je jednoprvková). Následne u tých pravidiel, ktoré dosiahnu minimálnu spoľahlivosť sa do *Y* pridá ďalšia položka z *X* atd.
Pre úplnosť uveďme, že existujú aj ďalšie postupy ako získať pravidlá.