Лекция 12. Извличане на многонивови и многомерни асоциативни правила

12.1. Извличане на многонивови асоциативни правила от транзакционни бази данни

Съществуват множество приложения, в които е трудно да бъдат намерени убедителни асоциативни правила между данни, съставени от артикули, описани на едно низко, примитивно ниво на абстракцията. Причината е разпръснатостта на подобни данни в общото многомерно пространство от възможни артикули. Обаче, от тези данни могат да бъдат извлечени интересните правила, но на друго по-високо ниво на абстракция, използвайки йерархии от понятия (виж Лекция 4), съществуващи за анализираните артикули. Да разгледаме един конкретен пример.

Пример 12.1. Да предположим, че разполагаме с база данни от транзакции за продажби, извършени в компютърния отдел на някоя фирма, специализирана в продажба на електроника. Една част от тази база е показа на Фиг. 12.1. Понятийната йерархия за съответните артикули, показана на Фиг. 12.2, определя последователността от отображения на множеството от понятия от по-низко ниво на абстракцията, към по-високо, съдържащо по-общи понятия. Показаната йерархия се състои от 4 нива, означени като 0, 1, 2 и 3, като по съглашение, те са номерирани отгоре надолу, започвайки с нивото 0 за корневият възел с етикет **всичко** (най-абстрактното ниво).

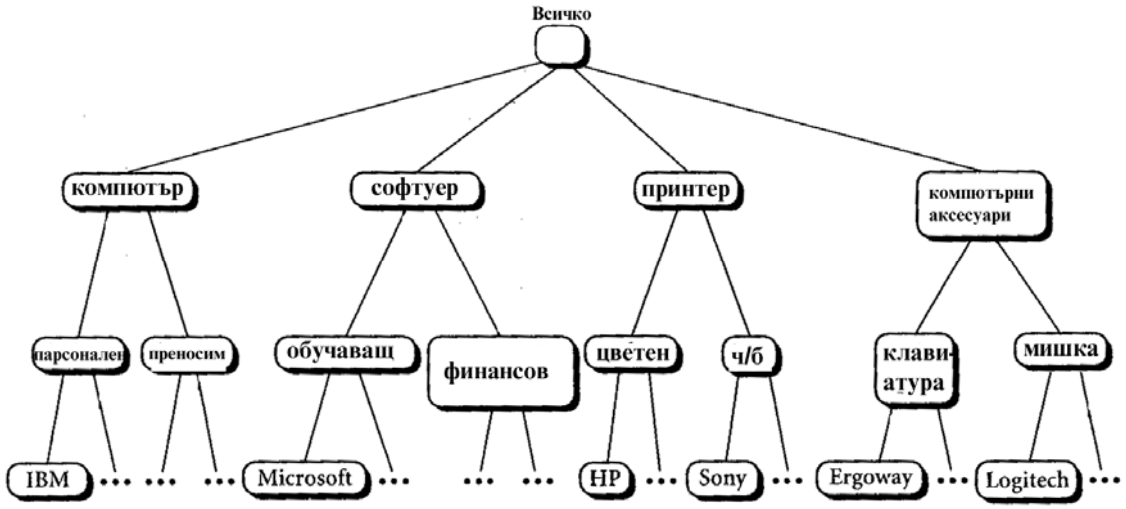
ТИД	Продадени артикули
T1	Персонален компютър на IBM, Ч/Б принтер на Sony
T2	Обучаващия софтуер на Microsoft, Финансовия софтуер на Microsoft
T3	Компютърна мишка на Logitech, Компютърна клавиатура на Ergoway
T4	Персонален компютър на IBM, Финансовия софтуер на Microsoft
T5	Персонален компютър на IBM
...	...
...	...
...	...

Фиг. 12.1. Част от примерната база данни от транзакции

Показаните в таблица артикули са от най-низкото ниво на йерархия, затова е доста тежко да бъдат намерени в тях интересни шаблони. Например, ако всеки от артикули “Персонален компютър на IBM” и ”Ч/Б принтер на Sony” участват в много малка част от транзакции, то може да се окаже доста затруднително да намерим някоя убедителна асоциация между тях – вероятността хората да закупят тези артикули заедно е малка, следователно, е малко вероятно, че артикулното множество {“Персонален компютър на IBM” , ”Ч/Б принтер на Sony”} да удовлетворява прага за минималната поддръжка. Обаче, ако разгледаме обобщението на ”Ч/Б принтер на Sony” към “Ч/Б принтер”, то очевидно е, че вероятността да намерим убедителните асоциации между “Персонален компютър на IBM” и ”Ч/Б принтер” ще бъде по-голяма, отколко между “Персонален компютър на IBM” и ”Ч/Б принтер на Sony”. Същото важи и за още по-общи понятия, като “компютър” и “принтер” – вероятността те да бъдат поръчани заедно е по-голяма от съвместната покупка

на “Персонален компютър на IBM” и ”Ч/Б принтер на Sony”. Следователно, е доста по-вероятно намирането на убедителни асоциативни правила между понятията от различни нива на понятийната йерархия, отколкото между примитивни понятия.

Асоциативните правила, извлечени с помощта на понятийни йерархии, се наричат *многонивови асоциативни правила*, тъй като те разглеждат понятия от повече от едно ниво на абстракцията.



Фиг. 12.2. Понятийна йерархия за електронни артикули

12.1.1. Подходи за извличане на многонивови асоциативни правила

Да разгледаме някои подходи към извличане на многонивови асоциативни правила, използващи за оценката на правила схемата “поддръжка – доверие”. В общия случай се използва стратегията “отгоре-надолу”, когато поддържащите суми се изчисляват за определяне на често срещани артикулни множества на всяко понятийно ниво, започвайки с нивото 1, и продължавайки в посоката на по-специфични понятия, докато вече не могат да бъдат намерени никакви други често срещани артикулни множества. Т.е. след намирането на често срещани артикулни множества на ниво 1, се търсят такива на ниво 2 и т.н. За търсенето на всяко от нивата може да се използва произволен алгоритъм за извличане на често срещани артикулни множества от типа на Apriori или неговите варианти.

Ще разгледаме по-долу различни варианти на този общ подход, като ги илюстрираме с фигури от 12.3 до 12.7. На тях възлите указват на артикул или артикулното множество, което е било проверявано, а възлите с удебелени контури отбелязват, че проверяваният артикул или множество са често срещани.

- *Използване на един и същ праг за минималната поддръжка (постоянен праг за поддръжка) за всички понятийни нива*: при търсене на често срещани артикулни множества се използва един и същ постоянен праг за минималната поддръжка не зависимо от нивото на абстракцията на артикулите. Например, на

Фиг. 12.3 на всички нива е използван минималният праг от 5%. Установено е, че и двата артикула “компютър” и “преносим компютър” са често срещани, докато “персонален компютър” – не е.

Фиг. 12.3. Многонивов анализ с постоянен праг за поддръжка.

Използването на постоянен праг за поддръжка опростява процедурата за търсене. Методът също така е прост в този смисъл, че от потребителите се изисква да бъде указан само един праг за минималната поддръжка. Може да бъде използвана и техниката за оптимизация, базирана на факта, че един предшественик е надмножеството на своите наследници – следователно, търсенето може да избегне проверката на такива артикулни множества, които съдържат някой артикул, предшествениците на който нямат необходимата минимална поддръжка.

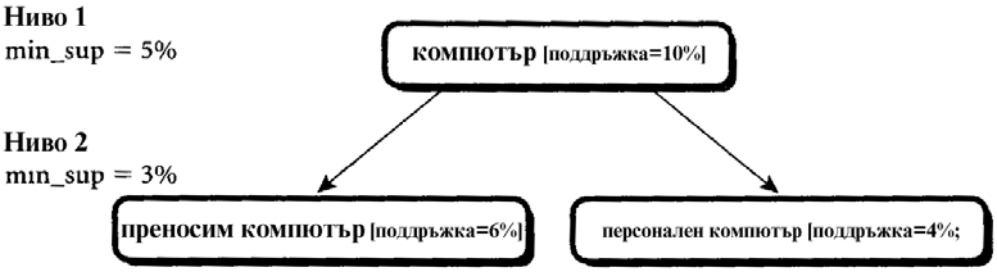
Обаче, подходът с използването на постоянния праг за поддръжка има определени недостатъци. Малко вероятно е, артикулите от по-низките нива на абстракции да присъстват толкова често в базата данни, като тези от по-високи нива в понятийната йерархия. Ако прагът за минималната поддръжка е установен прекалено високо, то възможно е изпускане на няколко важни асоциации, съществуващи на низкото ниво на абстракцията. Ако прагът е прекалено нисък, то възможно е генерирането на много неинтересни асоциации, възникващи на високите нива на абстракцията. Тези недостатъци мотивират възникване на следващия подход.

- *Използване на намален праг за минималната поддръжка на по-низки нива на абстракцията*: Всяко ниво на абстракцията има свой минимален праг за поддръжка. Колкото по-низко е нивото на абстракцията, толкова е по-малък съответния праг. Например, на Фиг. 12.4. прагът за минималната поддръжка на ниво 1 е 5%, а на ниво 2 – 3%. По този начин “компютър”, “преносим компютър” и “персонален компютър” се разглеждат всички като често срещани артикулни множества.

Съществуват няколко алтернативни стратегии за търсене, прилагани за извличане на многонивови асоциативни правила с намален праг за поддръжка:

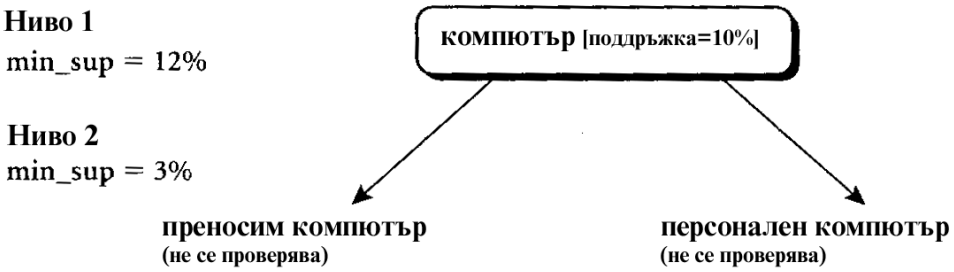
- *Независимо търсене ниво по ниво:* това е пълното търсене в ширина, при което не се използват никакви основни знания за често срещани атрибутни

множества. Всеки възел се проверява, не зависимо от това, дали неговият възел предшественик е често срещан или не.



Фиг. 12.4. Многонивов анализ с намален праг за поддръжка.

- *През-нивово филтриране чрез единичен артикул:* един артикул на i-то ниво се проверява тогава и само тогава, когато неговият възел предшественик на (i-1)-во ниво е често срещан. С други думи, по-специфични асоциации се изследват само, ако те са произлизат от по-общи. Ако възелът е често срещан, то се изследват неговите преки наследници ("деца"); в противен случай всички неговите наследници се изхвърлят от по-нататъшното търсене. Например, на Фиг. 12.5. възлите наследници на "компютър" (т.е. "преносим компютър" и "персонален компютър") не се проверяват, тъй като "компютър" не е често срещано артикулното множество.



Фиг. 12.5. Търсене с намален праг чрез през-нивово филтриране чрез единичен артикул

- *През-нивово филтриране чрез k-артикулно множество:* Едно k-артикулно множество на i-то ниво се проверява тогава и само тогава, когато неговото съответно k-артикулно множество-предшественик на (i-1)-во ниво е често срещано. На пример, на Фиг. 12.6 2-артикулното множество {компютър, принтер} е често срещано, по тази причина възлите {преносим компютър, ч/б принтер}, {преносим компютър, цветен принтер}, {персонален компютър, ч/б принтер} и {персонален компютър, цветен принтер} ще се проверяват.

Стратегията за независимото търсене ниво по ниво е много хлабава в смисъл, че може да доведе до проверка на голям брой на често срещани артикули от по-низките нива, намирайки по този начин асоциациите между артикули с малка важност. Например, ако артикулът "компютърна мебел" се поръчва рядко, то няма особен смисъл да се проверява, дали един такъв по-специфичен артикул като

“компютърен стол” има асоциацията с “преносим компютър”. Обаче, ако артикулът “компютърни принадлежности” се продава често, то може да бъде изгодно да се види, дали съществува шаблонът между “преносим компютър” и “мишка”.

Фиг. 12.6. През-нивово филтриране чрез k-артикулно множество (k=2)

Стратегията за през-нивовото филтриране чрез k-артикулно множество позволява на системата да проверява само децата на често срещани артикулни множества. Това е едно много строго ограничение, тъй като обикновено не съществуват много k-артикулни множества, които при своето комбиниране остават също често срещани (особено ако k > 2). При този подход могат да бъдат отхвърлени доста потенциално интересни шаблони.

Стратегията за през-нивовото търсене чрез единичен артикул представлява един компромис между двете описани по-горе крайности. Обаче, този метод може да изпуска асоциациите между артикулите от низкото ниво, които са често срещани на базата на намаления праг за поддръжка, но чиито предшественици не удовлетворяват минималния праг за поддръжка (тъй като праговете за поддръжка на всяко ниво са различни). Например, ако “цветен монитор” присъства на понятийно ниво i като често срещнат артикул на базата на прага за минималната поддръжка на ниво i, но неговия родител “монитор” не е често срещан съгласно прага на ниво (i –1), то ще бъде изпусната една такава често срещана асоциация като *персонален компютър* $\Rightarrow$ *цветен монитор* .

Една модифицирана версия на описаната стратегия, адресираща тези проблеми, е известна под името *управлявано през-нивово филтриране чрез единичен артикул*. При нея се използва така наречения “*праг за преминаване между нивата*” (*level passage support)*, който се установява, за да позволи “преминаване надолу” (т.е. към по-низки нива) на относително често срещани артикулни множества (наричани *почти често срещани – sub-frequent*). С други думи, методът позволява да бъдат проверявани децата на артикули, които не удовлетворяват прага за минималната поддръжка, ако тези артикули удовлетворяват прага за преминаване между нивата. Всяко понятийно ниво може да има свой собствен праг за преминаване между нивата. Прагът за преминаване между нивата за определеното ниво е обикновено се установява равен на някоя стойност между прага за минималната поддръжка на следващото по-низко ниво и прага за минималната поддръжка на даденото ниво. Потребителят може да избере да намали прага за прехода между нивата за по-високи понятийни нива, за да позволи проверката на наследниците на почти често

срещани артикулни множества на по-низки нива. Намаляването на прага за преминаване между нивата до нивото на прага за минималната поддръжка на най-низкото ниво ще позволи да бъдат проверявани наследниците на всички артикули. Например, на Фиг. 12.7 установяването на прага за преминаване между нивата (level_passage_sup) на нивото 1 в 8%, позволява възлите "преносим компютър" и "персонален компютър" от ниво 2 да бъдат проверявани и определени като често срещани, дори при условие, че техният възел-предшественик "компютър" не е често срещан. Чрез добавяне на този механизъм потребителите придобиват гъвкавостта за по-голям контрол над процеса на извличане на правила на различни нива на абстракцията, както и да намаляват броя на безсмислени асоциации, които в противен случай би били проверявани и генерирани.

Фиг. 12.7. Многонивовия анализ чрез управлявано през-нивово филтриране чрез единичен артикул

До сега сме разглеждали методи за намиране на често срещани множества от артикули, принадлежащи към едно и също понятийно ниво. Правилата, които включват артикули, принадлежащи към различни нива на абстракцията, се наричат *през-нивови асоциативни правила.* Един образец на такова правило е, например, $компютър \Rightarrow ч/б\ принтер$. Как се намират подобни асоциативни правила? При търсене на асоциации между понятийните нивата i и j, където j е по-низкото ниво на абстракцията от i, за да могат да бъдат включени в анализ артикулите от ниво j трябва навсякъде да бъде използван намален праг за минималната поддръжка на ниво j.

12.1.2. Проверка за излишни многонивови асоциативни правила

Понятийните йерархии са полезни за ИЗД, тъй като те позволяват намирането на зависимостите между понятия от различни нива на абстракцията от типа на многонивови асоциативни правила. Обаче, някои от намерените правила могат да се окажат излишни поради релацията "предшественик", съществуваща между участващи в тях артикули. Например, да разгледаме следните два правила, в които участва артикули "персонален компютър на IBM" и "персонален компютър", който е неговия предшественик съгласно понятийната йерархия от Фиг. 12.2:

$$персонален\ компютър \Rightarrow ч/б\ принтер\ [поддръжка = 8\%, доверие = 70\%] \quad (12.1)$$

$$\textit{персонален компютър на IBM} \Rightarrow \textit{ч / б принтер}\,[\textit{поддръжка} = 2\%, \textit{доверие} = 72\%] \quad (12.2)$$

Дали второто от тези намерени правила предоставя някаква по-нова информация от първото? Ако не, то трябва да бъде изтрито като излишното. Да разгледаме, как можем да намерим отговор на този въпрос. Ще казваме, че едно правило R1 е *предшественик* на правило R2, ако R1 може да бъде получено чрез заместване на артикулите от R2 с техните предшественици в зададена понятийна йерархия. Например, правилото (12.1) е предшественик на правилото (12.2), тъй като "персонален компютър" е предшественик на "персонален компютър на IBM". Използвайки това определение, едно правило може да бъде разгледано като излишно, ако неговите поддръжка и доверие *са близки до техните "очаквани" стойности,* изчислени на базата на правилото – предшественик на анализираното правило. Например, да предположим, че правило (12.1) има степен на доверие 70% и степен на поддръжка – 8%, и около една четврт от продажби на всички артикули "персонален компютър" пада на "персонален компютър на IBM". Може да се очаква, че правилото (12.2) ще има степен на доверие около 70% (тъй като всички членове на "персонален компютър на IBM" са също и членове на "персонален компютър") и степен на поддръжка около 2% (т.е. 8% x ¼). Ако точно това се наблюдава на практика, то правилото (12.2) не е интересно, тъй като то не предлага никаква нова информация и е по-малко общо от правилото (12.1). Следователно, то е излишно и трябва да бъде изтрито.

12.2. Извличане на многомерни асоциативни правила от релационни бази данни

До сега сме разглеждали начини за научаване само на булеви асоциативни правила, т.е. правила, които описват асоциации между стойностите на една единствена размерност (атрибут). Например, булевото правилото (12.1) може да бъде преписано чрез използване на атрибут *купува*:

$$\textit{купува}(X, \textit{персонален компютър}) \Rightarrow \textit{купува}(X, \textit{ч / б принтер})\,[\textit{поддръжка} = 8\%, \textit{доверие} = 70\%]$$

където променливата X представя купувачи на съответните артикули.

Правилата, които описват асоциациите между повече от един атрибут, се наричат *многомерни*. Пример на едно многомерно асоциативно правило е:

$$\textit{възраст}(X, \text{"30..90"}) \wedge \textit{женен}(X, \textit{да}) \Rightarrow \textit{брой\_коли}(X, 2)$$

Многомерните асоциативни правила могат да съдържат както номинални атрибути, (т.е. приемащи фиксирания брой дискретни стойности), така и непрекъснати атрибути. Един директен начин да сведем задачата за намиране на многомерни асоциативни правила към тази за намирането на булеви асоциативни правила е да заменим всеки номинален атрибут, които има n възможни стойности, с нови n

булеви атрибута, като стойността на един от тези атрибут ще бъде 1 (истина), ако в дадената транзакция (запис в релационна база данни) оригиналния номинален атрибут приема съответната стойност. Непрекъснати атрибути, които имат малък брой числови стойности, могат да бъдат бинаризирани по същия начин или към тях да бъде приложена някаква процедура за дискретизация, превръщаща непрекъснатия атрибут в номинален, а след това, по вече описания по-горе начин, и в множество от булеви атрибути. След подобно преобразование за намиране на многомерни асоциативните правила може да бъде приложена всякаква процедура за намиране на булеви асоциативни правила.

Пример 12.2. Да разгледаме приложение на описаната процедура за таблицата от Фиг. 12.8. Атрибутът "възраст" се дискретизира на два интервала "20..29" и "30..39", номинален атрибут "женен" се превръща в два булеви атрибута: "женен = да" и "женен = не". Тъй като брой стойности на непрекъснатия атрибут "брой_коли" е малък, то той не се дискретизира, а всяка от числовите му стойности се превръща в булевия атрибут, т.е. "брой_коли = 0", "брой_коли = 1" и "брой_коли = 2". Преобразуваната едномерна таблица е представена на Фиг. 12.9.

ID на записа	Възраст	Женен	Брой_коли
100	24	Не	1
200	25	Да	0
300	29	Не	0
400	34	Да	2
500	38	Да	2

Фиг. 12.8. Пример на релационна таблица с данни

ID	възраст=20..29	възраст=30..39	женен=да	женен=не	брой_коли=0	брой_коли=1	брой_коли=2
100	1	0	0	1	0	1	0
200	1	0	1	0	0	1	0
300	1	0	0	1	1	0	0
400	0	1	1	0	0	0	1
500	0	1	1	0	0	0	1

Фиг. 12.9. Преобразувана таблица с булеви атрибути

Обаче, съществуват два проблема, свързани с приложение на този прост подход към непрекъснати (числови) атрибути:

- *Минимална поддръжка*. Ако броят на дискретизационни интервали на някой числов атрибут (или броят на неговите стойности, ако той не е дискретизиран) е голям, степента на поддръжката за всеки един интервал може да бъде низка. Следователно, без да бъдат използвани по-големи интервали, някои от правилата, включващи този атрибут, могат и да не бъдат намерени, тъй като те няма до минат прага за минималната поддръжка.
- *Минималното доверие*. Дискретизацията на непрекъснатите стойности по интервали води до определена загуба на информация. Някои правила могат да удовлетворят прага за минималното доверие само, когато един от артикулите,

включени в лявата част на правилото, се състои от единичната стойност или малък интервал. Тази загуба на информацията нараства с нарастването на размера на интервалите.

Например, от таблицата на Фиг. 12.9 следва, че правилото: "брой_коли=0" $\Rightarrow$ "женен=не" има 100% доверие. Обаче, ако дискретизираме атрибут "брой_коли" на интервали, така че 0 и 1 попадат в един интервал, най-близкото правило : "брой_коли=0..1" $\Rightarrow$ "женен=не" ще има само 66.6% доверие.

Налице е една "зациклена" ситуация – ако интервалите са прекалено големи, то някои от правила няма да надминат прага за минималното доверие; ако те са много малки, то някои правила могат да не минат прага за минималната поддръжка.

12.2.1. Извличане на количествени асоциативни правила

Един възможен изход от този "цикъл" е предложен в (R. Srikant and R. Agrawal 1996). Те предлагат решението на проблема с "минималната поддръжка" чрез разглеждането на всички възможни интервали, получавани чрез сливането на съседни "базови" дискретизационни интервали. Проблемът за "минималното доверие" остава, но може да бъде облекчен чрез увеличаване на броя на базовите интервали.

Многомерните асоциативни правила, в които интервалите за дискретизация на непрекъснати (количествени) атрибути се определят *динамично*, се наричат *количествени асоциативни правила*.

За съжаление, увеличаването на броя на базовите интервали при едновременното комбиниране на съседните интервали води до възникване на два нови проблема:

- *"Време за изчисление"*. Ако един количествен атрибут има *n* стойности (или базови интервала), то съществуват в средно $O(n^2)$ диапазона или интервала, в които може да бъда включена една конкретна стойност или базов интервал. Следователно, броя на възможни артикули, създавани за един запис, нараства експоненциално, което води до експоненциалната експлозия във времето за изпълнение.
- *"Много правила"*. Ако една стойност (или базов интервал) на някой количествен атрибут удовлетворява прага за минималната поддръжка, то същото ще важи и за всеки диапазон (или интервал), съдържащ тази стойност (или базов интервал). Следователно, броят на правила нараства експоненциално, като много от тях могат и да не бъдат интересни (убедителни).

За облекчаване на проблема "време за изчисление" авторите ограничават начина, по който могат да се комбинират съседните стойности/базови интервали чрез въвеждане на един нов, задаван от потребителя параметър – *праг за максимална поддръжка*. Те прекратяват комбиниране на интервали, ако тяхна комбинирана степен за поддръжка надхвърля тази стойност. Обаче, ако някоя единична

стойност/базов интервал надхвърля прага за максималната поддръжка, тя продължава да участва в създаването на асоциативни правила.

За да проработи на практиката предложената схема, трябва да бъде решен въпросът, дали един количествен атрибут трябва да бъде дискретизиран и ако да, то на колко базови интервала. За да отговорят на този въпрос авторите въвеждат *мярка за частична пълнота*, която оценява количествено загубата на информация, причинявана в резултат от дискретизацията. Основната идея е следната. Нека R е множеството от правила, получено чрез разглеждане на всички възможни диапазони върху всички базови стойности на количествени атрибути. Нека R' е множеството от правила, получено чрез разглеждане на всички възможни диапазони върху дискретизационни интервали на количествени атрибути. Един начин за измерване на загубата на информация, получена от преминаването от R към R', е да се види за всяко правило в R, колко "далече" от него е "най-близкото" му правило от R'. Колкото по-далече е най-близкото правило, толкова по-голяма ще е загубата на информацията. Мярката за частична пълнота се извежда от дефинирането на "близките" правила като обобщения и от използването на отношение на степените на поддръжка на тези правила като мярка за разстояние между тях.

Доказано е, че използването на дискретизация с еднаква дълбочина, при която атрибутът се разбива на базови интервали с *приблизително еднакъв брой* атрибутни стойности във всеки интервал, минимизира броя на интервали, удовлетворяващи зададеното от потребителя ниво на частична пълнота K $(K > 1)$. Броят на базовите интервали k се определя по следната формула:

$$k = \frac{2 \times n}{m \times (K - 1)} \quad (12.3)$$

където

n = брой на количествени атрибути
m = праг за минималната поддръжка (като отношение)
K = ниво на частична пълнота

Ако броят на стойностите на някой количествен атрибут е по-малък или равен на изчисления брой дискретизационни интервали, то такъв количествен атрибут не се подлага на дискретизация.

Проблемът "много правила" се решава чрез въвеждане на една специална "мярка за интересност" на правилата, която се базира върху отклонение от очакваното и може да се разглежда като едно обобщение на начина за определяне на "излишни" многонивови асоциативни правила, разгледан в т. 12.1.2. Тази мярка свързва понятието "интересно" с правила, чиято степен на поддръжка и/или (в зависимост от желания на потребителя) на доверие е по-голяма от очакваното. Например, ако имаме два правила:

"възраст=20..25" $\Rightarrow$ "брой_коли=1..2 [поддръжка=2%, доверие=70%] и
"възраст=20..30" $\Rightarrow$ "брой_коли=1..2 [поддръжка=8%, доверие=70%]

то първото правило може да се смята за излишно, тъй като не носи никаква допълнителна информация и е по-малко общо от второто правило.

И така, проблемът за намиране на количествени асоциативни правила може да бъде решен чрез следните пет стъпки:

1. Определяне на броя на базови интервала за всеки количествен атрибут.
2. Унифициране на представянето на номинални и количествени атрибути. За номинални атрибути атрибутните стойности се отобразяват в последователни цели числа. За количествени атрибути, които не се дискретизират, атрибутните стойности се отобразяват в последователни цели числа по такъв начин, за да бъде запазен реда на стойностите. Ако атрибутът е дискретизиран на базови интервали, то те са отобразяват в последователни цели числа по начина, който запазва реда на интервалите. От тук нататък, алгоритмът използва унифицираното представяне на всички атрибути, обаче знае, кои от тях са номинални, а кои – количествени.
3. Определят се степени на поддръжка за всяка стойност както на номинални, така и на количествени атрибути. Освен това, за количествени атрибути съседните стойности се комбинират, докато степента на тяхната поддръжка е по-малка от указания от потребителя *праг за максималната поддръжка*. Всички стойности и получени диапазони от стойности над прага за минималната поддръжка за количествени атрибути, както и стойностите над този праг за номинални атрибути, образуват множество от *често срещани артикули*.
4. По начина, близък до този на Apriori, се намира множеството от всички често срещани k-артикулни множества (т.е. тези, които над прага за минималната поддръжка).
5. Получените често срещани k-артикулните множества се използват за генериране на асоциативни правила (пак чрез прилагане на механизма на Apriori). Тези от тях, които са над прага за минималното доверие, остават.
6. Изтриват се излишните правила чрез прилагането на мярка за интересност.

12.2.2. Извличане на асоциативни правила, базирани на разстояние

Разгледаните до сега методи за извличане на количествени асоциативни правила се базираха върху прости методи за дискретизация на числови атрибути. Обаче, подобни методи не отразяват семантиката на непрекъснати (интервални) атрибути, тъй като не отчитат разстояние между точки или интервали. Да разгледаме, например, данните за атрибута *цена*, дискретизирани на интервали с еднаква ширина, еднаква дълбочина и на базата на разстояние (Фиг. 12.10).

Очевидно е, че дискретизацията, базирана на разстояние, изглежда най-интуитивна, тъй като групира в един и същ интервал стойностите, които са близки, докато дискретизацията на интервали с еднаква дълбочина групира заедно много различни

стойности (напр. [22, 50]), а тази с еднаква ширина – създава интервали, които не съдържат никакви стойности.

Цена (лв)	*Еднаква ширина*	*Еднаква дълбочина*	*Базирани на разстояние*
7	[0, 10]	[7, 20]	[7,7]
20	[11, 20]	[22, 50]	[20, 22]
22	[21, 30]	[51, 53]	[50, 53]
50	[31, 40]		
51	[41, 50]		
53	[51, 60]		

Фиг. 12.10. Разбиване на интервали чрез различни дискретизационни методи

Класическите асоциативни правила не позволяват заключения за приближени значения на атрибутните стойности. Например, да разгледаме следното правило:

тип_артикула(X, електроника) $\wedge$ *производител(X, чуждестранна фирма)* $\Rightarrow$ *цена(X, 200$)*

В действителност, по реално е цените на електронни артикули, произведени от някоя чуждестранна фирма, да бъдат *близки* или *приблизително равни* на 200$, от колко са точно 200$. Би било желателно асоциативните правила да могат да отразяват тази степен на близост. Обърнете внимание, че класическите мерки за степента на поддръжка и доверие не разглеждат близостта на стойностите на атрибута.

Всички тези факти стимулираха разработка на методи за извличане на *асоциативни правила, базирани на разстояние*, които отчитат семантиката на интервални данни и близостта на атрибутните стойности. R. Miller и Y. Yang (1997) предложиха алгоритъм, състоящ от две фази: първата фаза използва модифицирания метод за агломеративна йерархична клъстеризация BIRCH (виж за подробности Лекция 6) за разбиването на непрекъснати атрибути на интервали или клъстери, а втората – създава асоциативни правила, базирани на разстояние, чрез търсене на групи от клъстери, които често се срещат заедно.

Да предположим, че данните се описват чрез някоя релационна схема $R = \{A_1, ..., A_m\}$, съдържаща m непрекъснати атрибута. Нека изследваната база данни съдържа r m-мерни обекта (записи) $T = \{t_1, ..., t_r\}$. За всяко подмножество на атрибути $X \subseteq R$ ще означаваме с $t_i[X]$ проекцията на запис t_i върху X, а чрез $S[X] = \{t_1[X], ...t_N[X], N \leq r\} \subseteq T[X]$ – множество от записи, проектирани върху X. *Диаметър* $d(S[X])$ на множеството $S[X]$ се дефинира като средно разстояние между всички двойки от записите в това множество:

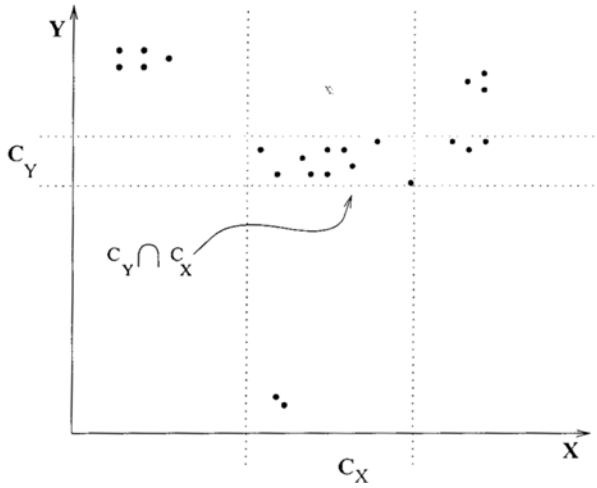
$$d(S[X]) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_X(t_i[X], t_j[X])}{N(N-1)} \quad (12.4)$$

където δ_X е избраната мярка за разстояние (например Евклидовото или абсолютното), дефинирана върху X атрибута. Ясно е, че диаметърът оценява

плътността (или гъстота) на едно множество от многомерни точки (записи), при тяхното проектиране на избраното подмножество от атрибути. Използвайки тези понятия, можем да дефинираме понятие *клъстер* $C_X[X]$ като множество от проектирани върху X записи, които удовлетворяват определен *праг за плътност* d_0^X и *праг за честота на срещане* s_0 (т.е. минимален брой записи в клъстера):

$$\begin{aligned} &d(C_X[X]) \le d_0^X \\ &|C_X[X]| \ge s_0 \end{aligned} \quad (12.5)$$

Едно просто асоциативно правило, базирано на разстояние, може да се представи във вида: $C_X[X] \Rightarrow C_Y[Y]$, където X и Y са не пресичащи се подмножества от атрибути. Искаме това правило да бъде убедително – това означава, че точки, групирани в клъстер C_X при проектиране на атрибути Y ще лежат в или близо до клъстера C_Y (виж фиг. 12.11 – случай, когато X и Y са едноелементни множества).



Фиг. 12.11. C_X *е клъстер, проектиран на X, а* C_Y *– на Y.*

Ако означим чрез $C_X[Y]$ клъстерът C_X, проектиран върху Y, то колко по-малко е разстоянието между $C_X[Y]$ и $C_Y[Y]$, толкова по-силна е *степента на асоциация* между клъстери $C_X[X]$ и $C_Y[Y]$ (т.е. толкова по-убедително е правилото $C_X[X] \Rightarrow C_Y[Y]$). Разстоянието между два множества от многомерни точки може да бъде измерено чрез такива стандартни статистически мерки, като, например, средното разстояние между клъстери *D1* или абсолютното разстояние между центровете (центроиди) на клъстери *D2*:

$$D1(C^1[X], C^2[X]) = \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\delta_X(t_j^1[X], t_j^2[X])}{N_1 N_2}, \quad D2(C^1[X], C^2[X]) = \delta_X(\vec{X}_0^1, \vec{X}_0^2), \quad (12.6)$$

където $\vec{X}_0^i = \frac{\sum_{j=1}^{N_i} t_j^i[X]}{N_i}, i = 1,2.$

И така, ще казваме, че едно просто асоциативно правило, базирано на разстояние, $C_X[X] \Rightarrow C_Y[Y]$ е вярно със степен на асоциация D_0, ако:

$$D(C_Y[Y], C_X[Y]) \leq D_0, \quad (12.7)$$

където D е избраната мярка за разстояние между клъстерите.

В общия случай, клъстерите могат да бъдат комбинирани в едно общо асоциативно правило, базирано на разстояние:

$$C_{X_1} \wedge C_{X_2} \wedge ... \wedge C_{X_x} \Rightarrow C_{Y_1} \wedge C_{Y_2} \wedge ...C_{Y_y} \quad (12.8)$$

където X_i и Y_j са непресичащи се по двойки подмножества от атрибути и са изпълнени следните три условия:

1. Всеки от клъстерите в условието на правило е убедително асоцииран със всеки от клъстерите в заключението на правило:
   $$D(C_{Y_j}[Y_j], C_{X_i}[Y_j]) \leq D_0, \ \ 1 \leq i \leq x, 1 \leq j \leq y$$
2. Всички клъстерите в условието на правило се срещат заедно:
   $$D(C_{X_j}[X_i], C_{X_i}[X_i]) \leq d_0^{X_i}, \ \ \forall \mathrm{i} \neq \mathrm{j}$$
3. Всички клъстерите в заключението на правило се срещат заедно:
   $$D(C_{Y_j}[Y_i], C_{Y_i}[Y_i]) \leq d_0^{Y_i}, \ \ \forall \mathrm{i} \neq \mathrm{j}$$

Както виждате, при асоциативни правила, базирани на разстояние, степента на асоциация замества степен на доверие в класически (и количествени) асоциативни правила, докато прагът на плътността замества понятие за прага на поддръжката.

12.3. Мерки за интересност

Алгоритмите за асоциативния анализ обикновено извличат голямо количество от правила, удовлетворяващи избраните прагове за минималната поддръжка и доверие. Тъй като обемът и размерността на реални търговски бази данни могат да бъдат много големи, е нормално да се извеждат хиляди или дори милиони правила, повечето от които могат и да не бъдат интересни. Пресяването на правилата с цел идентификация на интересните правила е далеч не тривиална задача тъй като „това, което за един е боклук, за друг може да е съкровище”. По тази причина е много важно да бъде установено определено множество от общопризнати критерии за оценяване на качество на асоциативните правила.

Първият набор от критерии може да бъде установен с помощта на статистически аргументи. Правилата, включващи набор от взаимно независими артикули или покриващи твърде малко транзакции се разглеждат като неинтересни, тъй като те

могат да отразяват някои фалшиви тенденции в данни. Подобни правила могат да бъдат елиминирани чрез прилагане на една *обективна мярка за интересност*, която използва статистики, изведени от данни, за да определи дали правилото е интересно или не. Примерите на обективните мерки за интересност са степен на поддръжка, доверие, корелация и др.

Вторият набор от критерии може да бъде получен чрез използване на субективни аргументи. Едно правило се разглежда като субективно неинтересно, ако то не разкрива някоя неочаквана информация за данни или не предоставя определени полезни знания, които могат да доведат до извличане на определена полза. Например правилото {масло} → {хляб} не може да се разглежда като интересно, независимо от високите степени на доверие и поддръжка, тъй като отношението, изразено чрез това правило, е изглежда прекалено тривиално. От другата страна, правилото {пелени} → {бира} е интересно, тъй като връзката е доста неочаквана и може да предложи на търговци интересни възможности за увеличаване на продажби. Включването на субективните знания в процеса на оценяване на правила е една много тежка задача, тъй като изисква от експертите в проблемната област да предоставят значителния обем от предварителната информация.

Дору са класифицирани някои от подходите за включване на субективните знания в задачата по намиране на интересни правила:

- *Визуализация*. Този подход изисква наличие на една дружелюбна среда, позволяваща включване в процеса на оценяване на човека-експерт. Тя позволява на експертите да взаимодействат със ИЗД система чрез интерпретиране и намиране на интересни закономерности.
- *Подход, базиран на шаблони*. Този подход позволява на потребителите да ограничават типове правила, извличани от ИЗД алгоритма. Вместо да предоставя на експерта всички извлечени правила, системата връща само правила, отговарящи на предоставени от потребителя изисквания – шаблони.
- *Субективна мярка за интересност*. Една субективна мярка може да бъде дефинирана на базата на такава информация за областта, като понятийната йерархия или възможностите за печалби от артикули. Мярката може след това да се ползва за филтриране на правила, които са очевидни и неизползваеми.

12.3.1. Обективни мерки за интересност

Една обективна мярка представлява водения от данни подход към оценяване на качеството на асоциативни правила. Подходът е проблемно независим и изисква минимални входни данни от потребителите, необходими за определяне на прага за филтриране на низко качествени правила. Обективната мярка обикновено се изчислява на база на броя на срещания, записани в една *таблица на съответствия (contingency table)*. В таблица 12-1 е представен

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Таблица 12-1. Двумерна таблица на съответствия за променливи A и B

един пример на такава таблица за двойка двоични атрибути (променливи) A и B. Нотацията $\overline{A}\,(\overline{B})$ се използва за указание, че A (B) липсват в една транзакция. Всеки елемент f_{ij} в тази 2 x 2 таблица означава броя на срещане. Например, f_{11} е броя, колко A и B се срещат заедно в една и съща транзакция, докато f_{01} е броят на транзакции, които съдържат B, но не съдържат A. Сумата по реда f_{1+} представлява поддържащата сума на A, докато сумата по колонка f_{+1} представлява поддържащата сума за B.

И накрая, макар, че нашата дискусия се фокусира основно на асиметрични двоични променливи (атрибути), таблиците на съответствия са приложими и за други типове атрибути, например симетрични двоични, номинални и наредени атрибути.

12.3.1.1. Ограничения на подхода „поддръжка – доверие"

Съществуващата формулировка на извличането на асоциативни правила се основава на мерките за поддръжка и доверие, целящи елиминиране на неинтересни правила. Недостатъкът на мярката за поддръжка е, че тя позволява отхвърляне на много потенциално интересни шаблона, съдържащи артикули с поддържащата сума под зададения праг. Проблемите с мярка за доверие са по-тънки, и най-добре могат да бъдат илюстрирани чрез примера.

Пример 12.3. Да предположим, че ние анализираме връзката между хора, които пият чай и кафе. Можем да съберем информация за предпочитания към тези напитки сред една група хора и да обобщим техните отговори в таблицата от типа на Таблица 12-2.

	$Coffee$	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

Таблица 12-2. Предпочитания към напитки в една група от 1000 души

Информацията от таблицата може да се използва за оценяване на асоциативното правило {чай} → {кафе}. На пръв поглед изглежда, че хората, които пият чай също имат тенденция да пият и кафе, тъй като поддръжката на правилото (15%) и доверието (75%) са достатъчно високи. Този аргумент би бил приемлив, ако не съществуваше фактът, че процентът на хора, които пият кафе, не зависимо от това, дали те пият чай или не, е 80%, докато процентът на пиещи чай хора, които също така пият и кафе, е само 75%. По тази начин, фактът, че човекът е любител на чай, в действителност *намалява* вероятността той да пие и кафе от 80% на 75%. Следователно, правилото {чай} → {кафе} е подвеждащо, не зависимо от високата му степен на доверие.

Проблемите с използването на мярката за доверие са свързани с факта, че тази мярка игнорира степента на поддръжката на артикулното множество, намиращо се в заключението на правило. Действително, ако вземем предвид степента на поддръжката на хора, пиещи кафе, ние няма да се изненадаме да намерим, че много от хората, пиещи чай, пият и кафе. По-учудващо ще е, че частта на любителите на чая, които пият и кафе, е в действителност по-малка от тези, които

просто обичат кафе, което сочи на съществуване на обратната връзка между пиещи чай и пиещи кафе.

Поради описаните ограничения на подхода поддръжка – доверие, за оценяване на качеството на асоциативни правила са били разработени множество от други обективни мерки, част от които ще разгледаме в тази лекция.

12.3.1.2. Фактор на интересност

Примерът с чай-кафе показва, че правилата с висока степен на доверие могат понякога да бъдат подвеждащи, тъй като мярката за доверие игнорира степента на поддръжка на артикулното множество в заключението на правило. Един от начините да се справим с този проблем е чрез прилагане на метриката, известна като *напредък (lift):*

$$Lift(A \to B) = \frac{c(A \to B)}{s(B)} \qquad (12.9)$$

която изчислява отношението между доверието на правилото и поддръжката на артикулното множество, формиращо заключението на правилото. За двоични атрибути мярката за напредък е еквивалентна на друга обективна мярка, наречена *фактор на интересност*, която се дефинира по следния начин:

$$I(A,B) = \frac{s(A,B)}{s(A) \times s(B)} = \frac{Nf_{11}}{f_{1+}f_{+1}} \qquad (12.10)$$

Факторът на интересност сравнява честотата на срещане на един шаблон срещу базовата честота на срещане, изчислена при предположението за статистическата независимост. Тази базовата честота на срещане за една двойка взаимно независими атрибути е:

$$\frac{f_{11}}{N} = \frac{f_{1+}}{N} \times \frac{f_{+1}}{N} \text{ или, еквивалентно, } f_{11} = \frac{f_{1+}f_{+1}}{N} \qquad (12.11)$$

Това уравнение следва от стандартния подход за използване на пропорция като приближение за стойността на вероятност. Пропорцията f_{11}/N е едно приближение за съвместната вероятност *P(A, B),* докато f_{1+}/N и f_{+1}/N са приближенията за *P(A)* и *P(B)* съответно. Ако *A* и *B* са статистически независими, то *P(A, B) = P(A) * P(B),* което води до формулата (12.11). Използвайки формулите (12.10) и (12.11) мярката за интересност може да се интерпретира по следния начин:

$$I(A,B) = \begin{cases} 1, \text{ако A и B са независими;} \\ > 1, \text{ако A и B са положително корелират} \\ < 1, \text{ако A и B са отрицателно корелират} \end{cases} \qquad (12.12)$$

За примера с чай-кафе, показан на Таблица 12-2. $I = \frac{0.15}{0{,}2 \times 0.8} = 0.9375$, което показва наличието на слаба отрицателна корелация между пиячите на чая и пиячите на кафе.

Ограничения на фактора на интересност

Ще илюстрираме ограниченията на фактора на интересност с един пример от областта на извличане на закономерности от текста. В тази област е разумно да смятаме, че асоциациите между една двойка думи зависят от броя на документи, които съдържат тези думи. На пример, поради наличието на силна зависимост между думите *data* и *mining* очакваме те да се срещат заедно по-често, от колко думите *compiler* и *mining* в една колекция от статии в областта на компютърни науки.

В таблица 12-3 са показани честоти на срещане на две двойки думи *{p, q}* и *{r, s}*. Използвайки формулата (12.10) факторът на интересност за *{p, q}* е 1.02, а за *{r, s}* е 4.08. Тези резултати са донякъде обезсърчаващи по следните причини. Макар че *p* и *q* присъстват заедно в 88% от всички документи, техния фактор за интересност е близък към 1, което е знак за статистическата независимост на *p* и *q*. От другата страна, фактор на интересност на *{r, s}* по-висок от този на *{p, q},* макар че *r* и *s* рядко се срещат заедно в един и същ документ. Мярката за доверие, по всяка вероятност, ще е по-добър избор в тази ситуация, тъй като разглежда асоциацията между *p* и *q* (94.6%) като много по-силна от тази между *r* и *s* (26.6%).

	p	$\bar{p}$	
q	880	50	930
$\bar{q}$	50	20	70
	930	70	1000

	r	$\bar{r}$	
s	20	50	70
$\bar{s}$	50	880	930
	70	930	1000

Таблица 12-3. Таблици на съответствия на двойки думи {p, q} и {r,s}

12.3.1.3. Корелационен анализ

Корелационният анализ е една статистически базирана техника за анализ на връзки между двойки атрибути (променливи). За непрекъснати атрибути корелацията се определя чрез вече познатия ни коефициент на корелацията (на Пиарсон). За двоични атрибути корелацията може да се измерва чрез така наречения φ-коефициент:

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}} \qquad (12.14)$$

Стойността на корелация варира от -1 (пълна отрицателна корелация) до + 1(пълна положителна корелация). Ако атрибутите са статистически независими, то φ = 0. Например, корелацията между пиячите на чай и кафе от примера 12.1 е -0.0625.

Ограничения на корелационен анализ

Проблемите с използване на коефициента на корелация могат да се видят на примера с асоциациите между думите, представен чрез таблица 12-3. Макар че думите *p* и *q* се срещат заедно по-често от *r* и *s,* техните φ-коефициенти са еднакви: *φ(p, q) = φ(r, s) = 0.232*. Причината за това е, че φ-коефициентът придава еднаква важност както на взаимното присъствие, така и на взаимното отсъствие на

артикули в една транзакция. По тази причина е по-подходящо тази мярка да се използва за анализа на симетрични двоични атрибути.

12.3.1.4. IS мярка

IS е една алтернативна мярка, предложена за работа с асиметрични двоични атрибути. Мярката се определя по следния начин:

$$IS(A,B) = \sqrt{I(A,B) \times s(A,B)} = \frac{s(A,B)}{\sqrt{s(A)s(B)}} \quad (12.15)$$

Обърнете внимание, че IS е голяма, когато факторът на интересност и поддръжка на шаблона са големи. Например, стойността на IS за двойката думи *{p, q}* и *{r, s}* са 0.946 и 0.286 съответно. Обратно на резултатите, получавани от фактора на интересност и φ-коефициента, мярката IS предполага, че асоциацията между *{p, q}* е по-силна от тази между *{r, s},* което се съгласува с това, което ние очакваме от асоциациите между думи в документи.

Възможно е да се покаже, че за двоични атрибути IS математически еквивалентна на мярката на косинуса. Във връзка с това да разгледаме **A** и **B** като една двойка от двоични вектори, **A** • **B** = *s(A, B)* – скаларното произведение на вектори и $|\mathbf{A}| = \sqrt{s(A)}$ е размерът на вектора **A**. Следователно:

$$IS(A,B) = \frac{s(A,B)}{\sqrt{s(A)s(B)}} = \frac{\mathbf{A} \bullet \mathbf{B}}{|\mathbf{A}| \times |\mathbf{B}|} = \cos(\mathbf{A},\mathbf{B}) \quad (12.16)$$

Мярката IS може да бъде изразена и като геометричното средно между факторите за доверие на асоциативните правила, извлечени от една двойка двоични атрибути:

$$IS(A,B) = \sqrt{\frac{s(A,B)}{s(A)} \times \frac{s(A,B)}{s(B)}} = \sqrt{c(A \to B) \times c(B \to A)} \quad (12.17)$$

Тъй като геометричното средно между две числа винаги е по-близо до по-малкото число, марката IS на едно атрибутно множество *{p, q}* е по-малка, когато едно от правилата $p \to q$ или $q \to p$ има низка степен на доверие.

Ограничения на IS мярка

IS мярката за двойка независими атрибутни множества *A* и *B* е равна на:

$$IS(A,B) = \frac{s(A,B)}{\sqrt{s(A)s(B)}} = \frac{s(A) \times s(B)}{\sqrt{s(A) \times s(B)}} = \sqrt{s(A) \times s(B)}\,.$$

Тъй като тази стойност зависи от *s(A)* и *s(B),* IS мярка има същия проблем като и мярката за доверие, а именно, че стойността на мярка може да бъде доста голяма дори за некорелиращи или отрицателно корелиращи шаблони. Например, макар че стойността на IS между *p* и *q* от таблица 12-4 е голяма (0.889), тя все пак по-малка от очакваната стойност, когато тези артикули са статистически независими ($IS_{независ} = 0.9$).

	q	$\bar{q}$	
p	800	100	900
$\bar{p}$	100	0	100
	900	100	1000

Таблица 12-4. Пример на таблицата на съответствия за артикули p и q.

12.3.2. Алтернативни обективни мерки за интересност

Освен разгледаните по-горе мерки, съществуват и няколко алтернативни обективни мерки за интересност, предложени за анализ на връзки между двойки двоични атрибути. Тези мерки могат да бъдат разделени на две категории – *симетрични* и *асиметрични* мерки. Една мярка M е симетрична, ако $M(A \rightarrow B) = M(B \rightarrow A)$. Например, факторът на интересност е симетричната мярка, тъй като нейната стойност е идентична за правила $A \rightarrow B$ и $B \rightarrow A$. Обратно, доверието е асиметричната мярка, тъй като доверието на $A \rightarrow B$ и $B \rightarrow A$ може и да не бъде едно и също. Симетричните мерки обикновено се използват за оценяване на атрибутните множества, докато асиметричните – са по-подходящи за анализ на асоциативните правила. В таблици 12-5 и 12-6 са представени дефинициите на някои от такива алтернативни мерки в термини на честотите на срещания за една 2 по 2 таблица на съответствия.

Мярка (Символ)	Дефиниция
Корелация (ϕ)	$\frac{Nf_{11}-f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$
Шансове да успех (α)	$(f_{11}f_{00})/(f_{10}f_{01})$
Капа (κ)	$\frac{Nf_{11}+Nf_{00}-f_{1+}f_{+1}-f_{0+}f_{+0}}{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}$
Интересност (I)	$(Nf_{11})/(f_{1+}f_{+1})$
Косинус (IS)	$(f_{11})/(\sqrt{f_{1+}f_{+1}})$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2}$
Обща сила (S)	$\frac{f_{11}+f_{00}}{f_{1+}f_{+1}+f_{0+}f_{+0}} \times \frac{N-f_{1+}f_{+1}-f_{0+}f_{+0}}{N-f_{11}-f_{00}}$
Jaccard (ζ)	$f_{11}/(f_{1+}+f_{+1}-f_{11})$
Общо доверие (h)	$\min\left[\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}}\right]$

Таблица 12-5. Примери на симетрични обективни мерки за артикулното множество {A, B}

Мярка (символ)	Дефиниция
Goodman-Kruskal (λ)	$\left(\sum_j \max_k f_{jk} - max_k f_{+k}\right)/\left(N - \max_k f_{+k}\right)$
Взаимна информация (M)	$\left(\sum_i \sum_j \frac{f_{ij}}{N} \log \frac{Nf_{ij}}{f_{i+}f_{+j}}\right)/\left(-\sum_i \frac{f_{i+}}{N} \log \frac{f_{i+}}{N}\right)$
J-мярка (J)	$\frac{f_{11}}{N} \log \frac{Nf_{11}}{f_{1+}f_{+1}} + \frac{f_{10}}{N} \log \frac{Nf_{10}}{f_{1+}f_{+0}}$
Gini индекс(G)	$\frac{f_{1+}}{N} \times \left[\left(\frac{f_{11}}{f_{1+}}\right)^2 + \left(\frac{f_{10}}{f_{1+}}\right)^2\right] - \left(\frac{f_{+1}}{N}\right)^2$ $+ \frac{f_{0+}}{N} \times \left[\left(\frac{f_{01}}{f_{0+}}\right)^2 + \left(\frac{f_{00}}{f_{0+}}\right)^2\right] - \left(\frac{f_{+0}}{N}\right)^2$
Laplace (L)	$(f_{11}+1)/(f_{1+}+2)$
Убеждение (V)	$(f_{1+}f_{+0})/(Nf_{10})$
Фактор на сигурност (F)	$\left(\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}\right)/\left(1 - \frac{f_{+1}}{N}\right)$
Добавена стойност (AV)	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

Таблица 12.6. Примери на асиметрични обективни мерки за правилото A → B

12.3.3. Съгласуваност между обективните мерки за интересност

При наличието на разнообразни мерки за интересност естествено възниква въпросът, дали тези мерки произвеждат едно сходно подреждане на резултати, когато са приложени към едно и също множество от асоциативни шаблони. Ако мерките са съгласувани, то ние можем да изберем една от тях като мярката за оценяване. Ако не, важно е да се разбере, каква е разликата между тях, за да се избере тази мярка, която е най-подходяща за анализа на конкретен тип шаблони.

Пример	f_{11}	f_{10}	f_{01}	f_{00}
E_1	8123	83	424	1370
E_2	8330	2	622	1046
E_3	3954	3080	5	2961
E_4	2886	1363	1320	4431
E_5	1500	2000	500	6000
E_6	4000	2000	1000	3000
E_7	9481	298	127	94
E_8	4000	2000	2000	2000
E_9	7450	2483	4	63
E_{10}	61	2483	4	7452

Таблица 12-7. Пример на таблици на съответствия

Да предположим, че симетричните и асиметричните мерки се прилагат за ранжиране на 10 таблици на съответствия, показани в Таблица 12-7. Тези таблици са избрани да илюстрират разликите между сравняваните мерки. Ранжирането, направено от тези мерки, е показано в таблици 12-8 и 12-9 съответно (като 1 означава най-интересна таблица на срещания, а 10 – най-малко интересна). Макар че някои от мерките изглеждат съгласувани помежду си, има и определени мерки, което произвеждат доста различни резултати при ранжирането. Например, ранжирането, направено от φ-коефициента е в съгласие с това, направено от κ и S, но е различно от ранжирането, направено от фактора на интересност и от шансове за успех α. Повече от това, една от таблиците на съответствия – E_{10} е ранжирана най-низко от φ-коефициента, но и най-високо от фактора на интересност.

	ϕ	α	κ	I	IS	PS	S	ζ	h
E_1	1	3	1	6	2	2	1	2	2
E_2	2	1	2	7	3	5	2	3	3
E_3	3	2	4	4	5	1	3	6	8
E_4	4	8	3	3	7	3	4	7	5
E_5	5	7	6	2	9	6	6	9	9
E_6	6	9	5	5	6	4	5	5	7
E_7	7	6	7	9	1	8	7	1	1
E_8	8	10	8	8	8	7	8	8	7
E_9	9	4	9	10	4	9	9	4	4
E_{10}	10	5	10	1	10	10	10	10	10

Таблица 12-8. Ранжиране на таблиците на съответствия от симетричните мерки

	λ	M	J	G	L	V	F	AV
E_1	1	1	1	1	4	2	2	5
E_2	2	2	2	3	5	1	1	6
E_3	5	3	5	2	2	6	6	4
E_4	4	6	3	4	9	3	3	1
E_5	9	7	4	6	8	5	5	2
E_6	3	8	6	5	7	4	4	3
E_7	7	5	9	8	3	7	7	9
E_8	8	9	7	7	10	8	8	7
E_9	6	4	10	9	1	9	9	10
E_{10}	10	10	8	10	6	10	10	8

Таблица 12-9. Ранжиране на таблиците на съответствия от асиметричните мерки

Приведените в таблиците резултати показват, че един значителен брой на мерките предоставят противоречаща се информация за качеството на един и същ шаблон.

Както виждате, съществуват множество различни обективни мерки за интересност, обаче диапазон на тяхното използване е доста ограничен. Основната причина за това е, че в повечето случаи оценката за интересност остава прекалено субективен и зависещ от конкретно приложение проблем. Поради това, използването на експерти за крайно оценяване на извлечените правила остава най-често използван подход в практиката.