قواعد وابستگی (هم‌باشی)
Association Rules
(Part 1)

M.M. Pedram
pedram@khu.ac.ir
Kharazmi University
(Fall 2007)

1

تحلیل سبد بازار (Market Basket Analysis)

- یکی از استراتژیهای داده کاوی
- هدف، یافتن ارتباطات جالب بین محصولات خرده‌فروشی (retail products) است، و تجزیه و تحلیل اقلامی که محتمل است با هم خریداری شـوند؛ و با کاهش هزینه ها، سود افزایش یابد.
- کاربردهای نتایج تحلیل توسط خرده فروشان
  - در ارتقا طراحی،
  - مرتب سازی قفسه ها (shelf)،
  - مواردی که باید در کاتالوگ ذکر شود،
  - استراتژی های cross-marketing و micro-marketing

✓ **Cross-marketing** : suggesting related products or services to a customer who is considering buying something.
- If you're buying a book on Amazon.com, for example, you may be shown a list of books similar to the one you've chosen or books purchased by other customers that bought the same book you did.
- A search on a company's Web site for bed linens might also bring up listings of matching draperies. The most ubiquitous example of cross-sell is likely the oft-spoken fast food phrase: "Would you like fries with that?"

✓ **Micro–marketing:** The study of activities of an organization, i.e. The activities a firm practices in order to react controllably to external forces, e.g., setting objectives and selecting target markets.

2

قواعد هم‌باشی (Association Rules)

- کشف وابستگیهای جالب بین attributeهایی که در پایگاه داده وجود دارند،
- مشهورترین روش برای تحلیل سبد بازار،
- در نظر گرفتن تمام ترکیبات گروه محصولاتی که به طور بالقوه جالبند،
- قابلیت تولید هزاران قاعده هم‌باشی با تعداد اندکی attribute،

3

تجزیه و تحلیل وابستگی (Affinity Analysis)

- فرآیند کلی تعیین اینکه چه چیزهایی با هم هستند و جلو می روند.
- این تجزیه و تحلیل در تجزیه وتحلیل سبد بازار به کار می رود.
- خروجی تجزیه و تحلیل سبد بازار مجموعه‌ای از هم‌باشی های مربوط به رفتار خرید مشتری است.
- **قواعد هم‌باشی (association rules)**: اگر هم‌باش‌ها توسط قواعد توصیف شوند به آنها قواعد هم‌باشی گویند.
  - کاربرد این قواعد در **تعیین استراتژیهای مناسب برای خرید محصولات** است.

4

تفاوت قواعد همباشی با قواعد عادی کلاس‌بندی

1. در قواعد هم‌باشی، هر **attribute** در یک قاعده می تواند به عنوان یک پیش شرط (یا مقدم) استفاده شود و در یک قاعده دیگر در بخش "نتیجه"(تالی) ظاهر شود در حالی که در قواعد عادی اینگونه نیست.

2. در قواعد عادی، در بخش "نتیجه" (تالی) فقط از یک **attribute** استفاده می شود، اما تولید کننده های قواعد هم‌باشی این امکان را می دهند که در نتیجه‌ی (تالی) قواعد چندین **attribute** را استفاده کرد.

5

مثالی از قواعد هم باشی

هدف آن است که تعیین شود در خرید اقلام زیر توسط مشتری، چه ارتباطاتی وجود دارد:

- Milk
- Cheese
- Bread
- Eggs

قواعد همباش زیر می تواند مطرح باشد:

1. اگر مشتری **Milk** بخرد، آنگاه **Bread** هم خواهد خرید.
2. اگر مشتری **Bread** بخرد، آنگاه **Milk** هم خواهد خرید.
3. اگر مشتری **Milk** و**Eggs** بخرد، آنگاه **Cheese** و **Bread** هم خواهد خرید.
4. اگر مشتری **Milk** و **Cheese** و **Eggs** بخرد، آنگاه **Bread** هم خواهد خرید.

توضیح: برای قاعده اول (و هر قاعده دیگر) سوالی به صورت زیر می‌تواند مطرح شود:
"چقدر محتمل است که خرید Milk منجر به خرید Bread شود؟"

6

"دامنه‌ی پشتیبانی قاعده" و "میزان اعتماد به قاعده"

- **درجه پشتیبانی قاعده (support for a rule)**

بیانگر درصدی از تمام معاملات(خریدها) است که شامل attributeهای آن قاعده است.

فرض کنید در مثال قبل، در مجموع 14000 خرید صورت گرفته است و در 5000 مورد خرید Bread و Milk با هم بوده است لذا دامنه کاربرد قاعده ۱(وهمین طور۲) که این دو صفت را شامل می‌شود، به صورت زیر است:

$$Supp(A \Rightarrow B) = Supp(B \Rightarrow A) = Supp(A, B)$$
$$= \frac{|A \cap B|}{|X|} = P(A \cap B) = \frac{5000}{14000} = 0.357 \equiv 35.7\%$$

- **درجه اعتماد به قاعده (rule confidence value)**

میزان اعتمادبه هر قاعده، یعنی احتمال ظاهر شدنattribute تالی به شرط وجود attribute مقدم، بیان می گردد.
مثلا برای قاعده اول اگر 10000 خرید مشتریها شامل Milk باشد و از آنها 5000 خرید شامل خرید Bread هم باشد در آن صورت میزان اعتماد به آن، به صورت زیر است:

$$Conf\ (A \Rightarrow B) = P(B \mid A) = \frac{|A \cap B|}{|A|} = \frac{Supp(A \rightarrow B)}{supp(A)}; \qquad supp(A) = \frac{|A|}{|X|}$$

$$P(\text{Bread خرید} \mid \text{Milk خرید}) = \frac{5000}{10000} = 0.5 \equiv 50\%$$

7

- **قاعده قوی (Strong Rule)** : قاعده ای است که دامنه پشتیبانی و میزان اعتماد آن بالا باشد.
  - هدف استخراج قواعد هم‌باشی، استخراج قواعد قوی است.

- **خاصیت Apriori** : که اگر مجموعه ای از صفت–مقدارها، پرتکرار(frequent) باشد، هر زیر مجمـــوعه‌ای از آن نیز پرتکرار است.

- اگر $AB=\{x_1, x_2, x_3, x_4\}$ و $A=\{x_1, x_2, x_3\}$ پرتکرار باشد، لذا Supp آنها باید بزرگ است، بنابراین نسبت آنها به عنوان Conf تعریف می‌شود، و مطلوب است به یک نزدیک باشد.

8

مثال

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

A⇒ B	Supp	Conf
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%

9

استخراج قواعد هم‌باش(Mining Association Rules)

- از آنجا که در مسائل عملی، تعداد ترکیبهای ممکن برای تالی قواعد هم‌باشی می تواند بسیار زیاد باشد(از دید ریاضی)، لذا روشهای عادی تولید قواعد برای تولید این قواعد استفاده نمی شود!
- الگوریتم Apriori، برای استخراج قواعد هم‌باشی استفاده می شود.

10

الگوریتم Apriori

- توسط **Agrawal** و همکارانش در ۱۹۹۳ مطرح شد.

R. Agrawal, T. Imielinski and A. Swami, "Mining association rules between sets of items in massive databases", *Proc. ACM SIGMOD Conf. on Management of Data, Washington DC, USA, 1993, pp.* 207-216.

- این الگوریتم **Item Sets** را تولید می‌کند.
  - Item Sets(مجموعه اقلام): ترکیباتی از صفت–مقدارها که شرط شمول (coverage) موردنظر را برآورده کنند.
  - شرط شمول (coverage): دامنه پشتیبانی(Support) توسط یک مجموعه اقلام (پشتیبانی کمینه یا min_supp).
  - در این الگوریتم ترکیب‌های مختلف صفت–مقدار تولید می‌شود و آنهایی که شرط شمول را برآورده نمی‌سازند، کنار گذارده می‌شوند. لذا این الگوریتم در زمانی معقول قواعد همباشی را تولید می‌کند.
  - خاصیت Aperiori استفاده می‌شود.

گام‌های الگوریتم Apriori

1. **تولید مجموعه های اقلام(Item Sets)**: تولید ترکیباتی که شرط شمول را برآورده می سازند. (**support** برای **Item Sets**)

2. **تولید قواعد هم‌باش به کمک مجموعه اقلام تولیدشده در گام I**: در این مرحله نه تنها قواعدی که میزان اعتماد بالایی دارند، مورد توجه قرار می‌گیرند بلکه یافتار زیر نیز برای آنها برقرار باشد:

$$Conf(A \Rightarrow B) - Supp(B) > d$$

or

$$Supp(A \rightarrow B) - Supp(A) \cdot Supp(B) > k$$

که d و k ثابت های مناسبی هستند.

مثال

❖ **مثال مربوط به پذیرش آگهی بیمه عمر را در نظر بگیرید:**

فرض: یک مجموعه داده از رفتار مشتریان در ارتباط با خرید چند Item در اختیار داریم.

هدف: بر اساس رفتار خرید مشتریها قواعدی بیان شود که نحوه ی ارتباط اقلام خریداری شده با هم مشخص گردد.

- **A Subset of the Credit Card Promotion Database**

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	No	Yes	Yes	Male
No	Yes	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female

گام 1 : تولید Item Sets

a. ابتدا **Single Item Sets** (یعنی انواع ترکیب یک صفت–مقدار) را مدنظر قرار می دهیم. اولین attribute عبارت است از Magazine Promotion که دارای دو مقدار Yes ,No است. بر اساس این صفت دو Item Set می توان تصور کرد که تعداد تکرار هر یک به قرار ذیل است:

Single Item Sets	No. of Item	accept
Magazine pro = Yes	**7**	✓
Magazine pro = No	**3**	x

چون مورد دوم کم است (یعنی شرط شمول حداقل 40٪ را نمی پوشاند)، لـذا نادیده گرفته می شود.

جدول Single Item Set به صورت زیر به دست خواهد آمد:

- **Single-Item Sets**

Single-Item Sets	Number of Items
Magazine Promotion = Yes	7
Watch Promotion = Yes	4
Watch Promotion = No	6
Life Insurance Promotion = Yes	5
Life Insurance Promotion = No	5
Credit Card Insurance = No	8
Sex = Male	6
Sex = Female	4

15

b. ترکیب Single Item Set برای به دست آوردن Two Item Set :

از دو مورد اول جدول صفحه قبل شروع می کنیم:

Magazine promotion = Yes ,Watch Promotion=Yes

3 مورد از 10 مورد dataset

چون زیر 40٪ است، آنرا نمی پذیریم.

دو مورد بعدی را بررسی می کنیم:

Magazine Promotion = Yes ,Watch Promotion = No

4 مورد از 10 مورد dataset

شرط شمول (40٪)را براورده می کند، لذا پذیرفته می شود.

بنابراین:

16

ادامه روش فوق منجر به تولید 11 مورد two Item Set زیر می شود:

- **Two-Item Sets**

Two-Item Sets	Number of Items
Magazine Promotion = Yes & Watch Promotion = No	4
Magazine Promotion = Yes & Life Insurance Promotion = Yes	5
Magazine Promotion = Yes & Credit Card Insurance = No	5
Magazine Promotion = Yes & Sex = Male	4
Watch Promotion = No & Life Insurance Promotion = No	4
Watch Promotion = No & Credit Card Insurance = No	5
Watch Promotion = No & Sex = Male	4
Life Insurance Promotion = No & Credit Card Insurance = No	5
Life Insurance Promotion = No & Sex = Male	4
Credit Card Insurance = No & Sex = Male	4
Credit Card Insurance = No & Sex = Female	4

17

c. ترکیب two Item Set برای به دست آوردن tree Item Set :

از جـــدول two Item Sets هر دو درایه که یک صفت مشـــترک دارند را با هم در نظـــر می گیریم پس دو تای اول را در نظر می گیریم که ترکیب آنها یک three Item Set خواهد بود:

Magazine Promotion=Yes & Watch Promotion=No &
Life Insurance Promotion=Yes

1 مورد از data set

آن را نمی پذیریم.

18

تنها three Item Set هایی که حداقل 4 مورد از موارد data set اولیه را بپوشاند، از جدول 2-Item Set اخذ شده باشد، مورد زیر است:

Watch Promotion =No & Life Insurance Promotion =No & Credit Card Insurance = No

4 مورد از data set

این روند را برای مسائلی که می توانند 4-Item Set, 5-Item Set,… داشته باشند، ادامه می دهیم.

19

گام2 : تولید قواعد هم‌باش

a. از جدول 2-Item Set قوانینی که شرط حداقل اعتماد (مثلا 80٪) را برآورد می کنند، استخراج می کنیم.

مثلا اولین مورد جدول 2-Item Set را در نظر می گیریم:

IF Magazine Promotion=Yes

THEN Watch Promotion =No

P(Watch Pro. | Magazine Pro.)=3÷7 < 80%

پس آنرا نادیده می گیریم!

البته عکس قاعده بالا را هم بررسی می کنیم.

20

با بررسی سایر موارد می بینیم که دو قاعده زیر از روی 2-ItemSets قابل بیان هستندکه حداقل confidence ٪80 را برآورده می کنند.

IF Magazine Promotion =Yes
THEN Life Insurance Promotion =Yes
(5/7)=> ~~confidence~~

IF Life Insurance Promotion =Yes
THEN Magazine Promotion =Yes
(5/5) => confidence

21

b. از جدول 3-ItemSets قواعدی که ترکیب سه تایی از صفت–مقدار را می توانند شامل شوند، به دست می آوریم و آنهایی که شرط حداقل confidence را برآورد می سازند، نگاه می داریم.
در این مثال فقط یک 3-Item Set داریم. از روی آن برخی قاعده ها که سه تای آنها در ذیل ذکر می شود به دست می آوریم.

IF Watch Promotion = No & Life Insurance= No
THEN Credit Card Ins. = No **(4/4)=> confidence**

IF Watch Promotion = No
THEN Life Insurance= No & Credit Card Ins. = No **(4/6)**

IF Credit Card Ins. = No
THEN Watch Promotion = No & Life Insurance= No **(4/8)**

22

مثال

Transaction	Items
t_1	Blouse
t_2	Shoes,Skirt,TShirt
t_3	Jeans,TShirt
t_4	Jeans,Shoes,TShirt
t_5	Jeans,Shorts
t_6	Shoes,TShirt
t_7	Jeans,Skirt
t_8	Jeans,Shoes,Shorts,TShirt
t_9	Jeans
t_{10}	Jeans,Shoes,TShirt
t_{11}	TShirt
t_{12}	Blouse,Jeans,Shoes,Skirt,TShirt
t_{13}	Jeans,Shoes,Shorts,TShirt
t_{14}	Shoes,Skirt,TShirt
t_{15}	Jeans,TShirt
t_{16}	Skirt,TShirt
t_{17}	Blouse,Jeans,Skirt
t_{18}	Jeans,Shoes,Shorts,TShirt
t_{19}	Jeans
t_{20}	Jeans,Shoes,Shorts,TShirt

23

Scan	Candidates	Large Itemsets
1	{Blouse},{Jeans},{Shoes}, {Shorts},{Skirt},{TShirt}	{Jeans},{Shoes},{Shorts} {Skirt},{Tshirt}
2	{Jeans,Shoes},{Jeans,Shorts},{Jeans,Skirt}, {Jeans,TShirt},{Shoes,Shorts},{Shoes,Skirt}, {Shoes,TShirt},{Shorts,Skirt},{Shorts,TShirt}, {Skirt,TShirt}	{Jeans,Shoes},{Jeans,Shorts}, {Jeans,TShirt},{Shoes,Shorts}, {Shoes,TShirt},{Shorts,TShirt}, {Skirt,TShirt}
3	{Jeans,Shoes,Shorts},{Jeans,Shoes,TShirt}, {Jeans,Shorts,TShirt},{Jeans,Skirt,TShirt}, {Shoes,Shorts,TShirt},{Shoes,Skirt,TShirt}, {Shorts,Skirt,TShirt}	{Jeans,Shoes,Shorts}, {Jeans,Shoes,TShirt}, {Jeans,Shorts,TShirt}, {Shoes,Shorts,TShirt}
4	{Jeans,Shoes,Shorts,TShirt}	{Jeans,Shoes,Shorts,TShirt}
5	$\emptyset$	$\emptyset$

24

مزایا و معایب الگوریتم Apriori

- مزایا
  - اجرای آن راحت است.
  - به طور موازی قابل اجراست.
- معایب
  - زمانبر است.
  - اگر attributeها، از نوع غیرباینری باشند، کارایی زمانی آن کمتر است.

ملاحظات کلی در مورد قواعد هم‌باشی

- در تعبیر قواعد هم‌باش باید دقت کرد زیرا بسیاری از روابطی که این قوانین کشف می کنند، بدیهی (trivial) هستند.

مثال: از ۱۰۰۰۰ خرید، ۷۰٪ شامل خرید شیر و نیز ۵۰٪ شامل نان می شوند. پس قاعده زیر محتملا بدست می آید:

IF customers purchase Milk **THEN** they also Purchase Bread.

فرض کنید confidence این قاعده هم‌باشی بالای ۴۰٪ باشد. اما این قاعده برای ما ارزش زیادی ندارد زیرا بیشتر مشتریها هر دو را می‌خرند. بنابراین، این قاعده اطلاع جدیدی به ما نمی‌دهد. (در مورد اینکه خرید نان با شیر را اگر آگهی کنیم به نفع ماست، زیرا این را می دانیم.)

دو نوع ارتباط که در قواعد هم باشی مورد توجه هستند

1. به آن قواعد هم‌باش علاقه‌مندیم که ارتقا فروش یک محصول خاص را ناشی از هم باشی با یک یا چند محصول دیگر بیان دارد.

   به این ترتیب، با این اطلاعات می‌توانیم فروش آنرا به عنوان نتیجه هم‌باشی بالا ببریم.

2. به آن قواعد هم‌باشی علاقه‌مندیم که برای یک هم‌باشی خاص مقدار **confidence** کمتری نسبت به آنچه انتظار داشتیم را بیان کنند.

   به این ترتیب می‌توانیم نتیجه بگیریم که محصولاتی که در آن قاعده هم‌باش به کار رفته اند با هم رقابت می کنند.

یک نکته

هر قاعده قوی، برای نمایش لزوما مناسب نیست. بلکه قواعدی که در یافتار ذکر شده هم صدق کنند، مورد توجه خواهند بود.

مثال:

دبیرستانی با **5000** دانش آموز و یک فروشنده صبحانه را در نظر بگیرید.

٦٠٪ دانش آموزان بسکتبال بازی می‌کنند.

٧٥٪ دانش آموزان صبحانه می‌خورند.

٤٠٪ دانش آموزان بسکتبال بازی می‌کنند و صبحانه می‌خورند.

شرط شمول =٤٠٪، شرط حداقل میزان اعتماد = ٦٠٪

قاعده روبرو تولید می شود: **(play basketball) → (eat cereal)**

چون شرط شمول را برآورده می‌کند، ونیز **Conf= 2000/3000 = 0.66**

اما:

Supp(A,B) – Supp(A).Supp(B) = 0.4 - 0.6*0.75 = -0.05 <0 !!