



Chapter 13 – Association Rules 关联规则分析

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

© Galit Shmueli and Peter Bruce 2010

What are Association Rules? 什么是关联规则



- Study of “what goes with what” 学习某事跟着某事发生
 - “Customers who bought X also bought Y” 购买X的客户同时也购买了Y
 - What symptoms go with what diagnosis 某种症状跟着某种诊断发生。
- Transaction-based or event-based 基于交易或者基于事件。
- Also called “market basket analysis” and “affinity analysis” 也被称为购物篮分析或者亲和度分析
- Originated with study of customer transactions databases to determine associations among items purchased 源于对消费交易数据库的研究，该研究目标是以判断消费者所购买的物品之间的关联。

Used in many recommender systems 用在很多推荐系统中



Bound Away [Last Train Home](#)



List Price: \$16.98

Price: **\$16.98** and eligible for **FREE Super Saver Shipping** on orders over \$25. [See details.](#)


Availability: Usually ships within 24 hours

Want it delivered Tomorrow? Order it in the next 4 hours and 9 minutes, and choose **One-Day S** checkout. [See details.](#)

[41 used & new](#) from \$6.99

▶ [See more product details](#)

[Share your own customer images](#)

 Based on customer purchases, this is the #82 [Early Adopter Product in Alternative Rock](#).

801x612

Buy this title for only \$.01 when you get a new Amazon Visa® Card

Apply now and if you're approved instantly, **save \$30** off your first purchase, earn **3% rewards**, get a **0% APR,*** and pay no



Amazon Visa discount: \$30.00

Applied to this item: - \$16.97

Discount remaining: \$13.03

[Find out how](#)

[\(Don't show again\)](#)

Customers who bought this title also bought:

- [Time and Water](#) ~ Last Train Home ([Why?](#))
- [Cold Roses](#) ~ Ryan Adams & the Cardinals ([Why?](#))
- [Tambourine](#) ~ Tift Merritt ([Why?](#))
- [Last Train Home](#) ~ Last Train Home ([Why?](#))
- [True North](#) ~ Last Train Home ([Why?](#))
- [Universal United House of Prayer](#) ~ Buddy Miller ([Why?](#))
- [Wicked Twisted Road \[ENHANCED\]](#) ~ Reckless Kelly ([Why?](#))
- [Hacienda Brothers](#) ~ Hacienda Brothers ([Why?](#))



Generating Rules 通常的规则



Terms 术语

“IF” part = antecedent “如果” 段 = （规则的）前件

“THEN” part = consequent “那么” 段 = （规则的）后件

“Item set” = the items (e.g., products) comprising the antecedent or consequent 项集 = 项（例如产品）组成的前件或者后件。

- Antecedent and consequent are *disjoint* (i.e., have no items in common) 前件和后件之间是分开的（也就是说没有相同的项）。



Tiny Example: Phone Faceplates

Transaction	Faceplate Colors Purchased				
1	red	white	green		
2	white	orange			
3	white	blue			
4	red	white	orange		
5	red	blue			
6	white	blue			
7	white	orange			
8	red	white	blue	green	
9	red	white	blue		
10	yellow				



Many Rules are Possible 有许多可能的规则



For example: Transaction 1 supports several rules, such as

- “If red, then white” (“If a red faceplate is purchased, then so is a white one”)
- “If white, then red”
- “If red and white, then green”
- + several more



Frequent Item Sets 频繁项集

- Ideally, we want to create all possible combinations of items 理想状况下，我们希望能生成所有可能的项组合。
- Problem: computation time grows exponentially as # items increases 问题：计算时间随着项的个数呈指数增长。
- Solution: consider only “frequent item sets” 解决方法：只考虑频繁项集。
- Criterion for frequent: *support* “频繁”的判断标准：支持度



Support 支持度

Support = # (or percent) of transactions that include both the antecedent and the consequent 支持度 = 同时包括前件和后件的交易数量。

Example: support for the item set {red, white} is 4 out of 10 transactions, or 40%



Apriori Algorithm (Apriori算法)

Generating Frequent Item Sets

生成频繁项集



For k products... 对于 k 个产品

1. User sets a minimum support criterion 用户设置一个最小支持度标准。
2. Next, generate list of one-item sets that meet the support criterion 接着生成满足支持度标准的单项集列表。
3. Use the list of one-item sets to generate list of two-item sets that meet the support criterion 使用单项集的列表生成满足支持度标准的2项集列表。
4. Use list of two-item sets to generate list of three-item sets 使用2项集列表生成3项集列表。
5. Continue up through k -item sets 继续直到生成 k 项集列表。



Measures of Performance 评估指标

- *Confidence*: the % of antecedent transactions that also have the consequent item set

置信度：包含前件的交易当中同时包含后件的交易百分比。

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

- $\text{Lift} = \text{confidence} / (\text{benchmark confidence})$

提升水平 = 置信度 / 基准置信度。

- *Benchmark confidence* = transactions with consequent as % of all transactions

基准置信度 = 所有交易中包含后件的交易百分比。

- $\text{Lift} > 1$ indicates a rule that is useful in finding consequent items sets (i.e., more useful than just selecting transactions randomly)

提升水平大于1说明关联规则对于找到后件交易项有用（这也就是说，比通过随机挑选交易找到后件交易项有用）。

Alternate Data Format: Binary Matrix



Transaction	Red	White	Blue	Orange	Green	Yellow
1	1	1	0	0	1	0
2	0	1	0	1	0	0
3	0	1	1	0	0	0
4	1	1	0	1	0	0
5	1	0	1	0	0	0
6	0	1	1	0	0	0
7	1	0	1	0	0	0
8	1	1	1	0	1	0
9	1	1	1	0	0	0
10	0	0	0	0	0	1



Process of Rule Selection 规则选择流程

Generate all rules that meet specified support & confidence 生成满足特定支持度和置信度的所有规则。

- Find frequent item sets (those with sufficient support – see above) 找到频繁项集（那些有足够大支持度的项集）
- From these item sets, generate rules with sufficient confidence 从这些项集中，生成有足够大置信度的规则。

A Tiny Example



Transaction	Faceplate Colors Purchased			
1	red	white	green	
2	white	orange		
3	white	blue		
4	red	white	orange	
5	red	blue		
6	white	blue		
7	white	orange		
8	red	white	blue	green
9	red	white	blue	
10	yellow			



Example: Rules from {red, white, green}

{red, white} > {green} with confidence = $2/4 = 50\%$

$$\blacksquare [(\text{support } \{\text{red, white, green}\}) / (\text{support } \{\text{red, white}\})]$$

{red, green} > {white} with confidence = $2/2 = 100\%$

$$\blacksquare [(\text{support } \{\text{red, white, green}\}) / (\text{support } \{\text{red, green}\})]$$

Plus 4 more with confidence of 100%, 33%, 29% & 100%

If confidence criterion is 70%, report only rules 2, 3 and 6

All Rules (XLMiner Output)



Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	Green=>	Red, White	2	4	2	2.5
2	100	Green=>	Red	2	6	2	1.666667
3	100	Green, White=>	Red	2	6	2	1.666667
4	100	Green=>	White	2	7	2	1.428571
5	100	Green, Red=>	White	2	7	2	1.428571
6	100	Orange=>	White	2	7	2	1.428571



Interpretation

□ *Lift ratio* shows how effective the rule is in finding consequents (useful if finding particular consequents is important) 提升比率显示用规则找到后件的有效性如何（在找到特定后件时特别有用）

□ *Confidence* shows the rate at which consequents will be found (useful in learning costs of promotion) 置信度显示（给定前件情况下）找到后件的比率是多少（在了解销售推广成本时有用）。

Caution: The Role of Chance 小心巧合



Random data can generate apparently interesting association rules 随机数据可以生成明显有趣的关联规则。

The more rules you produce, the greater this danger 你制造的规则越多，就越可能是巧合的结果。

Rules based on large numbers of records are less subject to this danger 如果在大量数据基础上生成规则，那么巧合的危险就小了。

Example: Charles Book Club



ChildBks	YouthBks	CookBks	DoltYBks	RefBks	ArtBks	GeogBks	ItalCook	ItalAtlas	ItalArt	Florence
0	1	0	1	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	1	0	1	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0
1	0	0	0	0	1	0	0	0	0	1
0	1	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0
1	1	1	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	1	0	0	0	0

Row 1, e.g., is a transaction in which books were bought in the following categories: Youth, Do it Yourself, Geography

XLMiner Output



Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	ItalCook=>	CookBks	227	862	227	2.320186
2	62.77	ArtBks, ChildBks=>	GeogBks	325	552	204	2.274247
3	54.13	CookBks, DoltYBks=>	ArtBks	375	482	203	2.246196
4	61.98	ArtBks, CookBks=>	GeogBks	334	552	207	2.245509
5	53.77	CookBks, GeogBks=>	ArtBks	385	482	207	2.230964
6	57.11	RefBks=>	ChildBks, CookBks	429	512	245	2.230842
7	52.31	ChildBks, GeogBks=>	ArtBks	390	482	204	2.170444
8	60.78	ArtBks, CookBks=>	DoltYBks	334	564	203	2.155264
9	58.4	ChildBks, CookBks=>	GeogBks	512	552	299	2.115885
10	54.17	GeogBks=>	ChildBks, CookBks	552	512	299	2.115885
11	57.87	CookBks, DoltYBks=>	GeogBks	375	552	217	2.096618
12	56.79	ChildBks, DoltYBks=>	GeogBks	368	552	209	2.057735

□ Rules arrayed in order of lift

□ Information can be compressed

e.g., rules 2 and 7 have same trio of books

Summary 总结



- Association rules (or *affinity analysis*, or *market basket analysis*) produce rules on associations between items from a database of transactions 关联规则（或者亲和度分析，购物篮分析）从交易数据库生成（购物）项之间的关联规则
- Widely used in recommender systems 广泛应用于推荐系统。
- Most popular method is Apriori algorithm 最广泛使用的方法是Apriori算法。
- To reduce computation, we consider only “frequent” item sets (=support) 为了减少计算量，我们仅考虑频繁项集。
- Performance is measured by *confidence* and *lift* 关联规则的评估指标是置信度和提升度。
- Can produce a profusion of rules; review is required to identify useful rules and to reduce redundancy 可能会生成大量规则，需要人工检视找到有用的规则并去掉冗余规则。