Chapter 13 – Association Rules

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

What are Association Rules?

- Study of “what goes with what”
  - “Customers who bought X also bought Y”
  - What symptoms go with what diagnosis
- Transaction-based or event-based
- Also called “market basket analysis” and “affinity analysis”
- Originated with study of customer transactions databases to determine associations among items purchased

Used in many recommender systems

Bound Away
Last Train Home

Share your own customer images

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

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

Terms

"IF" part = **antecedent**

"THEN" part = **consequent**

"Item set" = the items (e.g., products) comprising the antecedent or consequent

- Antecedent and consequent are *disjoint* (i.e., have no items in common)

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

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

- Ideally, we want to create all possible combinations of items
- **Problem:** computation time grows exponentially as # items increases
- **Solution:** consider only “frequent item sets”
- Criterion for frequent: *support*

Support

Support = # (or percent) of transactions that include both the antecedent and the consequent

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

Apriori Algorithm

Generating Frequent Item Sets

For *k* products…

1. User sets a minimum support criterion
2. Next, generate list of one-item sets that meet the support criterion
3. Use the list of one-item sets to generate list of two-item sets that meet the support criterion
4. Use list of two-item sets to generate list of three-item sets
5. Continue up through *k*-item sets

Measures of Performance

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

Lift = *confidence*/(*benchmark confidence*)

Benchmark confidence = transactions with consequent as % of all transactions

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

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

Generate all rules that meet specified support & confidence

- Find frequent item sets (those with sufficient support – see above)
- From these item sets, generate rules with sufficient confidence

Example: Rules from {red, white, green}

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

- [(support {red, white, green})/(support {red, white})]

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

- [(support {red, white, green})/(support {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)
- *Confidence* shows the rate at which consequents will be found (useful in learning costs of promotion)
- *Support* measures overall impact

Caution: The Role of Chance

Random data can generate apparently interesting association rules

The more rules you produce, the greater this danger

Rules based on large numbers of records are less subject to this danger

Example: Charles Book Club

ChildBks	YouthBks	CookBks	DoItYBks	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, DoItYBks=>	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=>	DoItYBks	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, DoItYBks=>	GeogBks	375	552	217	2.096618
12	56.79	ChildBks, DoItYBks=>	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

Practical Tips

- Choice of Proper Level of Abstraction
  - Beverage: Cola : Coca Cola : Coca Cola 250ml: Diet Coca Cola 250ml
  - 롯데리아 vs 카페베네
- Virtual Items
  - Day of the week, Time of the day, Season, Region, Shopper's Gender/age
    - "IF shopper in 20's AND night THEN cup ramen"
  - Membership ID
    - Other behavior

Summary

- Association rules (or *affinity analysis,* or *market basket analysis*) produce rules on associations between items from a database of transactions
- Widely used in **recommender systems**
- Most popular method is **Apriori algorithm**
- To reduce computation, we consider only “frequent” item sets (=support)
- Performance is measured by *confidence* and *lift*
- Can produce a profusion of rules; review is required to identify useful rules and to reduce redundancy