Association Rules
(Aka Affinity Analysis or Market Basket Analysis)
Prof. Tom Fomby
Department of Economics
Southern Methodist University
Dallas, Texas 75275
July 2013

Association Rules represent an **unsupervised learning method** that attempts to capture associations between groups of items. Association Rules have also been referred to in the literature as Market Basket analysis or Affinity analysis. For example, consider the EXCEL file **Associations.xls** that details the transactions of individuals in a bookstore. The first 12 cases (out of 2000) taken from the Associations file are reproduced below in Table 1.

Table 1

12 Bookstore Transactions
In Binary Format

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

The above reporting style is called **Binary Format** as compared to **Item List Format** as represented in Table 2 below where the first 5 transactions from a grocery store are recorded.

Table 2
5 Grocery Store Transactions
In Item List Format

hering	corned_b	olives	Ham	turkey	bourbon	ice_crea
baguette	soda	hering	Cracker	heineken	olives	corned_b
avocado	cracker	artichok	Heineken	ham	turkey	sardines
olives	bourbon	coke	Turkey	ice_crea	ham	peppers
hering	corned_b	apples	Olives	steak	avocado	turkey

More formally an **item list** is a set of items chosen during a given trip to a store while a **transactions file** is a data file that details the items purchased by customers during trips to the store. Therefore, the Associations.xls file in XLMINER is a transactions file, the first 12 transactions of which are listed in Table 1 above. The item list of the first transaction consists of {YouthBks, DoItYBks, GeogBks}.

The question of interest here is the following: If items in group A appear in a market basket, what is the probability that items in group C will also be purchased? In probability terms, what is the conditional probability of C, given A, i.e. Pr(C|A)? Also we might ask, "If the fact that group A items are in the market basket, does that condition at all the probability that the group C items will also be in the market basket?" That is, is simply Pr(C|A) = Pr(C) and thus knowing that group A items are in the market has no effect on group C items being in the market basket? Suppose, for example, that a person in the bookstore has purchased an Italian Cookbook. What is the probability that the same person would also purchase a General Cookbook given that the Italian Cookbook is in the market basket? Is the purchase of the General Cookbook independent of the purchase of an Italian Cookbook? Which items in the bookstore are associated with or have an affinity for each other? Answering questions like the above is the task of Association Rules.

Everyone should be familiar by now with the "sidebar" advertisements that go along with websites like Amazon.com. When looking at a given book on that website, one will often see some "recommended" books in the sidebar with a caption that reads, "Customers who bought this item also bought …" These sidebar books have some "affinity" in previous consumer purchases on Amazon to the book that you are currently viewing on your computer screen. More than likely these sidebar books have been determined by the use of association rules. Naturally the sidebar books help Amazon sell more books and increases its revenue from book sales. Also if one were designing the arrangement of items in a store, one would surely like to convenience the customer so that goods that have affinities to each other would be located close to each other so that consumer search time would be minimized and store revenues would be maximized!

The Basic Terminology and Notation of Association Rules

Let **antecedent** items be those items that "cause" the purchase of "consequent" goods whereas the **consequent** items are those that are purchased because antecedent items have been purchased. Which comes first, the milk or the cookies? In association rule analysis the timing of the purchases is not the crucial matter, but the association is. Finally, let the **support** of a set of items be the number of transactions in which that set of items occurs in the transactions file.

Let all of the items in the antecedent set be denoted by A. In contrast, let all items in the consequent set be denoted by C. Now consider that we have a transactions list like in Table 1. What we want to do is determine some associations that represent rules that have high confidence. By an association rule we mean "If A then C with a given probability." The **confidence of a rule** is calculated as follows:

Confidence = (number of transactions containing all of the items in A **and** C)/
(number of transactions containing the items in A)

= (support of $A \cup C$) /support of A)

$$= \Pr(C|A) = \frac{\Pr(A \cup C)}{\Pr(A)} \quad . \tag{1}$$

The latter equality above denotes the law of conditional probability. The higher the confidence (conditional probability) for an association rule is, the better the rule.

Another important concept in association rules is that of the "Lift" of the rule. The **Lift Ratio of an Association Rule** is defined as follows:

$$\text{Lift} = \text{Confidence/Expected Confidence} \tag{2}$$

where

Expected Confidence = (number of transactions having the consequent items)/
(total number of transactions)

$$= \Pr(C).$$

Another way at looking at Lift is

$$\text{Lift} = \frac{\Pr(C \mid A)}{\Pr(C)}. \tag{3}$$

When Lift is equal to 1 it must be that A and C are independent because $\Pr(C \mid A) = \Pr(C)$. Only when the probability of C is affected by the occurrence of A is the Lift of the rule greater than one. Of course, the larger the Lift of a rule, the better the rule is.

Comparing Lift with Confidence, the Lift of a rule is a **relative measure** in the sense that it compares the degree of dependence in a rule versus independence between the consequent items and the antecedent items. The rules that have higher Lift will have higher dependence in them. In contrast, the percentage confidence is an **absolute measure**. Given the antecedent items, the percentage confidence is the probability that the consequent items will be purchased.

Searching for Good Association Rules: The A priori Algorithm

Searching randomly through a transaction file for meaningful rules is an impossible task. As in the recursive binary algorithm for building regression and

classification trees there are methods for making the search for good Association Rules feasible. The search method used for finding good association rules is called the **A priori Algorithm**. This algorithm is due to R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487 – 499, Santiago, Chile, September 1994.

The general procedure for executing the A priori Algorithm is as follows: Let a **k-item set** be a set of k items of interest. The **support** of that specific k-item set is the number of transactions in which the specific k items occur in the transactions file. For example, if 3 specific items occur in 200 transactions in a transactions file of 2000 transactions, the support of that 3-item set is 200. (Equivalently, the support of a k-item set could be defined as the percentage of the transactions in the transactions file where the k-items appear. In the above example that would be 10%.) Now the A priori algorithm starts out with determining the support of all 1-item sets. If any 1-item set does not meet a given threshold (t) pre-determined by the analyst, the item is dropped from further consideration in making association rules and the other 1-item sets retained. The next step consists of using the retained 1-item sets to form 2-item sets that have minimum support of t. All 2-item sets that do not have **minimum support** are dropped from further consideration while the other 2-items sets are retained. After the requisite 2-item sets have been retained, they, along with the retained 1-item sets, can be used to build 3-items sets that meet the minimum support criterion. It follows that the candidate 4-item sets are built from the previous minimum support 3-items and 1-item sets and are examined for minimum support. This process continues until at the K-item set level there are no K-item sets that meet the minimum support criterion. We then have 1 through (K-1)-item sets that have minimum support.

At the next stage all possible Association Rules are considered using the retained 1 through (K-1) item sets. The Rules are parsed by requiring that retained rules meet a minimum confidence requirement, for example 50%. Minimum confidence is then the second tuning parameter in the A priori algorithm with the first being the minimum support parameter. It should be noted, however, that in forming Association Rules, the antecedent item set has to be **disjoint** from the consequent item set in each Association Rule. In the end then, only Association Rules that have minimum confidence and whose antecedent sets and consequent sets are disjoint and have minimum support are chosen for examination and possible implementation.

Obviously, the larger (smaller) the minimum support and confidence requirements, the fewer (more) rules will be determined. Unfortunately, the rules determined from any one setting of the minimum support and confidence requirements will often be different from the rules determined from other settings of the minimum requirements. Therefore, in order to have a "robust" view of the Association Rules that a given transactions file can generate, we should try several minimal support and confidence requirements when building a meaningful set of Association Rules. Of course, no matter what the A priori algorithm might produce in terms of "strong" association rules, **if a rule is not interpretable or intuitive in the face of domain-**

specific knowledge, it probably should not be adopted and used for decision-making purposes.

A Simple Example to Illustrate the A priori Algorithm

Suppose we have the following transactions from which we want to build some association rules that have good confidence and lift.

{1,2,3,4}, {1,2}, {2,3,4}, {2,3}, (1,2,4}, {3,4}, and {2,4}.

The numbers represent specific items like bread, milk, and the like.

One way to build such a set of association rules is to apply the A priori algorithm. The A priori algorithm proceeds first by identifying one-item sets that have minimum support. Assume that we choose 3 as the minimum support (or a threshold t = 3/7 = 0.42851). The one-item sets that meet this support requirement are

1 – Item Sets	Support
{1}	3
{2}	6
{3}	4
{4}	5

Now using these frequent 1 – Item sets we can form candidate 2 – Item sets, {1,2}, {1,3}, {1,4}, {2,3}, {2,4}, and {3,4}. Which of these candidates meet the minimum support requirement? The minimum support 2 – Item sets are the first four sets reported below.

2 – Item Sets	Support	
{1,2}	3	
{2,3}	3	
{2,4}	4	
{3,4}	3	
---	---	---
{1,3}	1	eliminated
{1,4}	2	eliminated

From the "qualifying" 2 – Item sets and "qualifying" 1 – Item sets we can build candidate 3 – Item sets. The candidate 3 – Item sets and their supports are as follows:

3 – Item Sets	Support
{1,2,3}	1
{1,2,4}	2

{2,3,4}	2
{1,3,4}	1

As none of the 3 – Item sets meet the minimum support requirement, we are finished with the first phase of the A priori algorithm.

The second phase of the algorithm involves the building of Association Rules $A \Rightarrow C$ that satisfy a minimum confidence requirement. Recall that the confidence of a rule is defined as P(C|A) where C represents a consequent item set and A denotes a **disjoint** antecedent item set (i.e. the C and A sets have no items in common).

One can start building candidate association rules by starting out with rules that have a single item antecedent (from the qualifying 1 – Item sets) and a single item consequent (from the qualifying 1 – Item sets). These rules with their corresponding confidences are as follows:

Rule	Confidence
{2}|{1}	support{1,2}/support{1} = 3/3
{3}|{1}	support{1,3}/support{1} = 1/3
{4}|{1}	support{1,4}/support{1} = 2/3
{3}|{2}	support{2,3}/support{2} = 3/6
{4}|{2}	support{2,4}/support{2} = 4/6
{4}|{3}	support{3,4}/support{3} = 3/4

and **reversing** the antecedents and the consequents of the above candidate rules provide the rest of the single item antecedent and consequent candidate rules

Rule	Confidence
{1}|{2}	support{1,2}/support{2} = 3/6
{1}|{3}	support{1,3}/support{3} = 1/4
{1}|{4}	support{1,4}/support{4} = 2/5
{2}|{3}	support{2,3}/support{3} = 3/4
{2}|{4}	support{2,4}/support{4} = 4/5
{3}|{4}	support{3,4}/support{4} = 3/5 .

Now moving to the higher-order candidate Association Rules, we have the following candidate Association Rules. (Remember the antecedent and consequent item sets must be disjoint).

Rule	Confidence
{2,3}|{1}	support{1,2,3}/support{1} = 1/3
{2,4}|{1}	support{1,2,4}/support{1} = 2/3
{3,4}|{1}	support{1,3,4}/support{1} = 1/3

{3,4}|{2}	support{2,3,4}/support{2} = 2/6
{1,2}|{3}	support{1,2,3}/support{3} = 1/4
{2,4}|{3}	support {2,3,4}/support{3} = 2/4
{1,2}|{4}	support{1,2,4}/support{4} = 2/5
{2,3}|{4}	support{2,3,4}/support{4} = 2/5

and, of course, reversing the antecedents and consequents of the above rules we obtain the additional candidate rules

Rule	Confidence
{1}|{2,3}	support{1,2,3}/support{2,3} = 1/3
{1}|{2,4}	support{1,2,4}/support{2,4} = 2/4
{1}|{3,4}	support{1,3,4}/support{3,4} = 1/3
{2}|{3,4}	support{2,3,4}/support{3,4} = 2/3
{3}|{1,2}	support{1,2,3}/support{1,2} = 1/3
{3}|{2,4}	support {2,3,4}/support{2,4} = 2/4
{4}|{1,2}	support{1,2,4}/support{1,2} = 2/3
{4}|{2,3}	support{2,3,4}/support{2,3} = 2/3 .

Then using a minimum 50% confidence for qualifying Association Rules we would accept the following rules with their confidences and lift ratios displayed sorted from highest confidence to lowest confidence.

Rule	Confidence	Lift Ratio
{2}|{1}	3/3	conf/(support{2}/7) = (3/3)/(6/7) = 7/6
{2,4}|{1}	2/3	conf/(support{2,4}/7) = (2/3)/(4/7) = 7/6
{1}|{2}	3/6	conf/(support{1}/7) = (3/6)/(3/7) = 7/6
{1}|{2,4}	2/4	conf/(support{1}/7) = (2/4)/(3/7) = 7/6
{4}|{3}	3/4	conf/(support{4}/7) = (3/4)/(5/7) = 21/20
{3}|{4}	3/5	conf/(support{3}/7) = (3/5)/(4/7) = 21/20
{2}|{4}	4/5	conf/(support{2}/7) = (4/5)/(6/7) = 14/15
{4}|{1}	2/3	conf/(support{4}/7) = (2/3)/(5/7) = 14/15
{4}|{2}	4/6	conf/(support{4}/7) = (4/6)/(5/7) = 14/15
{4}|{1,2}	2/3	conf/(support{4}/7) = (2/3)/(5/7) = 14/15
{4}|{2,3}	2/3	conf/(support{4}/7) = (2/3)/(5/7) = 14/15
{2}|{3}	3/4	conf/(support{2}/7) = (3/4)/(6/7) = 7/8
{3}|{2}	3/6	conf/(support{3}/7) = (3/6)/(4/7) = 7/8
{3}|{2,4}	2/4	conf/(support{3}/7) = (2/4)/(4/7) = 7/8
{2}|{3,4}	2/3	conf/(support{2}/7) = (2/3)/(6/7) = 7/9

When ordering the above Association Rules by highest lift ratio to lowest lift ratio we have the rules {1} ⇒ {2}, {1} ⇒ {2,4}, {2} ⇒ {1}, and {2,4} ⇒ {1} having the highest lift ratios of 7/6 and the rule {3,4} ⇒ {2} having the lowest lift ratio of 7/9. Obviously, the Association Rules that have lift ratios of 1 or less are not helpful.

A More Detailed Example

To illustrate the application of the A priori algorithm to the Associates.xls transactions file containing book purchases by type, consult Table 3 below. Here the determined association rules have been sorted from the rule with the highest Lift Ratio to the lowest, all of which satisfy the minimal support requirement of t = 250 and minimum confidence requirement of 50%. Of course, clicking on the column for confidence would allow one to sort the rules according to highest to lowest confidence if that were the order of priority of the user. (Note: In the below XLMINER output the column heading that reads "Support ($A \cup C$)" really means the support of sets that have both A **and** C in them.)

Table 3

Association Rules
Determined by XLMINER
With Minimum Support = 250
And Minimum Confidence = 50%

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio ↓
1	58.4	ChildBks, CookBks=>	GeogBks	512	552	299	2.115885
2	54.17	GeogBks=>	ChildBks, CookBks	552	512	299	2.115885
3	52.49	ArtBks=>	ChildBks, CookBks	482	512	253	2.050376
5	50.39	ChildBks, CookBks=>	YouthBks	512	495	258	2.035985
4	52.12	YouthBks=>	ChildBks, CookBks	495	512	258	2.035985
6	57.03	ChildBks, CookBks=>	DoItYBks	512	564	292	2.022385
7	51.77	DoItYBks=>	ChildBks, CookBks	564	512	292	2.022385
8	52.9	ArtBks=>	GeogBks	482	552	255	1.916832
9	79.63	CookBks, YouthBks=>	ChildBks	324	846	258	1.882497
10	79.35	ChildBks, DoItYBks=>	CookBks	368	862	292	1.841017
11	77.87	CookBks, DoItYBks=>	ChildBks	375	846	292	1.84082
12	77.66	CookBks, GeogBks=>	ChildBks	385	846	299	1.835989
13	78.18	ChildBks, YouthBks=>	CookBks	330	862	258	1.813963
14	77.85	ArtBks, ChildBks=>	CookBks	325	862	253	1.806175
15	75.75	ArtBks, CookBks=>	ChildBks	334	846	253	1.790745
16	76.67	ChildBks, GeogBks=>	CookBks	390	862	299	1.778809
17	70.65	GeogBks=>	ChildBks	552	846	390	1.670264
18	70.63	RefBks=>	ChildBks	429	846	303	1.669725
19	71.1	RefBks=>	CookBks	429	862	305	1.649549
20	69.75	GeogBks=>	CookBks	552	862	385	1.618245
21	69.29	ArtBks=>	CookBks	482	862	334	1.607763
22	67.43	ArtBks=>	ChildBks	482	846	325	1.594028

23	66.67	YouthBks=>	ChildBks	495	846	330	1.576044
24	66.49	DoItYBks=>	CookBks	564	862	375	1.542677
25	65.25	DoItYBks=>	ChildBks	564	846	368	1.542511
26	65.45	YouthBks=>	CookBks	495	862	324	1.518667
27	60.52	ChildBks=>	CookBks	846	862	512	1.404179
28	59.4	CookBks=>	ChildBks	862	846	512	1.404179

To consider the computations involving the number 1 rule above, {ChildBlks, CookBks} $\Rightarrow$ {GeogBks}, let us first compute the rule's confidence. Since there are 2000 transactions in this file, the confidence in this rule is calculated as

Confidence = (support of $A \cup C$) / support of A) = (299/512)

$$= \frac{\Pr(A \cup C)}{\Pr(A)} = \frac{299/2000}{512/2000} = 0.584 \quad .$$

The calculation of the Lift Ratio for the first rule proceeds as follows:

$$\text{Lift} = (\text{confidence}/\Pr(C)) = \frac{0.584}{(552/2000)} = 2.115885.$$

Obviously, there is substantial dependence (affinity) between the consequent {GeogBks} and the antecedent {ChildBlks, CookBks}.