



AN INTRODUCTION TO THE APRIORI ALGORITHM FOR MINING ASSOCIATION RULES IN COMMERCIAL, MEDICAL, AND ENVIRONMENTAL FIELDS


PRESENTATION FOR M.S. OF STATISTICS, UNIVERSITY OF COLORADO, DENVER

COMMITTEE: DR. STEPHANIE SANTORICO (CHAIR), DR. ERIN AUSTIN, DR. ADAM SPIEGLER

EMMA COLLINS, SUMMER 2020




Outline

1. Introduction
 2. Methodology
 3. Applications
 4. Results
 5. Discussion
 6. References
- 




Introduction

- Rakash Argawal, Ramakrishnan Srikant in 1994
 - Originally used to determine what items in stores were frequently bought together [1]
 - Unsupervised, clustering method – clusters variables, not observations
 - Uses joint, conditional probabilities, user-chosen thresholds
- 



Introduction

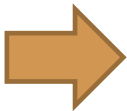
- “What is useful or different is often application dependent.”
 - Agrawal & Srikant [1]
 - Compare and contrast different context, number of observations and variables, and parameters
 - Commercial, medical, and environmental applications
- 



Methodology – Data

- Data needs to be binary, or transformed into binary

Obsv.	Seconds	Sex
034	33.6	Male
035	21.5	Male
036	26.8	Female
037	19.2	Male
038	23.7	Female



Obsv.	≤ 20 sec.	20-30 sec.	≥ 30 sec.	Male	Female
034	0	0	1	1	0
035	0	1	0	1	0
036	0	1	0	0	1
037	1	0	0	1	0
038	0	1	0	0	1



Methodology – Association Rules

- Association rules are joint values of attributes characterized by probabilities called *support* and *confidence* [2]
- Let A and B be sets of attributes, called *itemsets*

$$A \rightarrow B[\textit{support}, \textit{confidence}]$$

- A is called the antecedent, B is called the consequent

$$A = \{\textit{milk}, \textit{bread}\} \quad B = \{\textit{chocolate}\}$$

Methodology – Association Rules

- Let X and Y be indicator variables for A and B , respectively

$X = I(\text{all attributes of } A \text{ appear together in observation } n)$

$Y = I(\text{all attributes of } B \text{ appear together in observation } n)$

$$A = \{\text{milk}, \text{bread}\} \quad B = \{\text{chocolate}\}$$

- If milk and chocolate are bought together, but bread is not, $X = 0$ and $Y = 1$

Methodology – Association Rules

- *Support* is how often a set of attributes appear together in N total observations

$$\text{support}(A \rightarrow B) = \frac{\text{frequency}(X, Y)}{N} = \hat{P}(X = 1 \cap Y = 1)$$

- Note, $\text{support}(A \rightarrow B) = \text{support}(B \rightarrow A)$
- Sometimes just count, $\text{frequency}(X, Y)$

Methodology – Association Rules

- *Confidence* is how often an itemset, B , appears given another itemset, A

$$\text{confidence}(A \rightarrow B) = \frac{\text{frequency}(X, Y)}{\text{frequency}(X)} = \hat{P}(Y = 1 | X = 1)$$

- Note, $\text{confidence}(A \rightarrow B)$ does not necessarily equal $\text{confidence}(B \rightarrow A)$

Methodology – Association Rules

- Antecedent can be the empty set, then

$$\text{support}(\emptyset \rightarrow A) = \hat{P}(X = 1) = \text{confidence}(\emptyset \rightarrow A) \quad [3]$$

- The probabilities are estimates

Methodology – Association Rules

- *Lift* is a type of correlation measure, importance of the rule [4]

$$lift(A \rightarrow B) = \frac{support(X \rightarrow Y)}{support(X) * support(Y)} = \frac{\hat{P}(X = 1 \cap Y = 1)}{\hat{P}(X = 1) * \hat{P}(Y = 1)}$$

- Lift = 1, itemsets frequently appear together by chance
Lift > 1, one itemset encourages the other itemset
Lift < 1, one itemset discourages the other itemset

Methodology – Association Rules

- *Support* is how often itemsets A and B appear together
- *Confidence* is how often itemset B appears, given itemset A has appeared
- *Lift* is how important the rule is, if A and B appear together by chance

Methodology – Apriori Algorithm

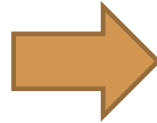
- Let s_m and c_m be user-selected minimums for support and confidence, respectively
- L_k are *large itemsets* that meet the required minimum support, of size k
- C_k are *candidate itemsets*, potentially large itemsets, of size k

Methodology – Apriori Algorithm

1. Generate L_1
2. For $k \geq 2$, generate C_k from combinations of L_{k-1}
3. Generate L_k by calculating support for the sets in C_k and selecting those with support $\geq s_m$
4. Repeat 2, 3 increasing k each time until dataset is exhausted or s_m is no longer reached
5. Generate potential association rules from all L_k , calculate confidence, return rules with confidence $\geq c_m$
6. (optional) Calculate lift of returned association rules

Apriori Algorithm Example

Transaction	Items
001	Milk, bread, cheese
002	Milk, butter
003	Butter, cheese
004	Milk, bread, butter, cheese, chocolate
005	Milk, bread



Transaction	Milk	Bread	Butter	Cheese	Chocolate
001	1	1	0	1	0
002	1	0	1	0	0
003	0	0	1	1	0
004	1	1	1	1	1
005	1	1	0	0	0

Let $s_m = 0.25$, $c_m = 0.80$

Step 1: Generate L_1

$$L_1 = \{\{milk\}, \{bread\}, \{butter\}, \{cheese\}\}$$

$$s_m = 0.25$$

$$c_m = 0.80$$

Step 2.1: Generate C_2

Candidate itemset, C_2	Support
{milk, bread}	0.6
{milk, butter}	0.4
{milk, cheese}	0.4
{bread, butter}	0.2
{bread, cheese}	0.4
{cheese, butter}	0.4

Step 2.1: Generate C_2

Candidate itemset, C_2	Support
{milk, bread}	0.6
{milk, butter}	0.4
{milk, cheese}	0.4
{bread, butter}	0.2
{bread, cheese}	0.4
{cheese, butter}	0.4

$$s_m = 0.25$$

$$c_m = 0.80$$

Step 3.1: Generate L_2

$$L_2 = \{\{milk, bread\}, \{milk, butter\}, \{milk, cheese\}, \{bread, cheese\}, \{cheese, butter\}\}$$

$$s_m = 0.25$$

$$c_m = 0.80$$

Step 3.1: Generate L_2

$$L_2 = \{\{milk, bread\}, \{milk, butter\}, \{milk, cheese\}, \{bread, cheese\}, \{cheese, butter\}\}$$

Step 2.2: Generate C_3 (cannot have bread, butter together - $P(A \cap B) \leq P(A)$)

Candidate itemset, C_3	Support
{milk, bread, cheese}	0.4
{milk, cheese, butter}	0.2

Step 3.2: Generate L_3

$$L_3 = \{\{milk, bread, cheese\}\}$$

Step 3.2: Generate L_3

$$L_3 = \{\{milk, bread, cheese\}\}$$

$$s_m = 0.25$$

$$c_m = 0.80$$

Cannot have milk, cheese, and butter in an itemset satisfy s_m , the loop is exhausted (step 4).

Step 5: Generate all combination of rules from L_1, L_2, L_3 .

$$L_1 = \{\{milk\}, \{bread\}, \{butter\}, \{cheese\}\}$$

$$L_2 = \{\{milk, bread\}, \{milk, butter\}, \{milk, cheese\}, \{bread, cheese\}, \{cheese, butter\}\}$$

$$L_3 = \{\{milk, bread, cheese\}\}$$

Step 5: Generate all combination of rules from L_1, L_2, L_3 .

$$s_m = 0.25$$

$$c_m = 0.80$$

$$L_1 = \{\{milk\}, \{bread\}, \{butter\}, \{cheese\}\}$$

$$L_2 = \{\{milk, bread\}, \{milk, butter\}, \{milk, cheese\}, \{bread, cheese\}, \{cheese, butter\}\}$$

$$L_3 = \{\{milk, bread, cheese\}\}$$

Antecedent	Consequent	Support	Confidence
{}	{milk}	0.80	0.80
{}	{bread}	0.60	0.60
{}	{butter}	0.60	0.60
{}	{cheese}	0.60	0.60
{milk}	{bread}	0.60	0.75
{milk}	{cheese}	0.40	0.75
{milk}	{butter}	0.40	0.50
{bread}	{milk}	0.60	1.00
:	(excess itemsets removed for space)	:	:
{bread, cheese}	{milk}	0.40	1.00
{milk, cheese}	{bread}	0.40	1.00
{milk, bread}	{cheese}	0.40	0.67

Step 5: Generate all combination of rules from L_1, L_2, L_3 .

$$L_1 = \{\{milk\}, \{bread\}, \{butter\}, \{cheese\}\}$$

$$L_2 = \{\{milk, bread\}, \{milk, butter\}, \{milk, cheese\}, \{bread, cheese\}, \{cheese, butter\}\}$$

$$L_3 = \{\{milk, bread, cheese\}\}$$

$$s_m = 0.25$$

$$c_m = \mathbf{0.80}$$

Antecedent	Consequent	Support	Confidence
{}	{milk}	0.80	0.80
{}	{bread}	0.60	0.60
{}	{butter}	0.60	0.60
{}	{cheese}	0.60	0.60
{milk}	{bread}	0.60	0.75
{milk}	{cheese}	0.40	0.75
{milk}	{butter}	0.40	0.50
{bread}	{milk}	0.60	1.00
:	(excess itemsets removed for space)		:
{bread, cheese}	{milk}	0.40	1.00
{milk, cheese}	{bread}	0.40	1.00
{milk, bread}	{cheese}	0.40	0.67


Discard rules with confidence $< c_m$, association rules for $s_m = 0.25$ and $c_m = 0.80$:

Antecedent	Consequent	Support	Confidence
{}	{milk}	0.80	0.80
{bread}	{milk}	0.60	1.00
{bread, cheese}	{milk}	0.40	1.00
{milk, cheese}	{bread}	0.40	1.00

Are the last two rules different?



Applications

- Commercial, medical, and environmental
 - How do minimum support and minimum confidence thresholds change?
 - What are the number of observation and attributes?
 - What are the size of the association rules returned?
 - **What is at risk if the association rules are false?**
- 

Applications – Commercial


- Kurnia, Isharianto, Giap, Hermawan, and Riki, 2019 [5]
- 150 transactions and 38 menu items from restaurant in Jakarta, Indonesia
- $s_m = 0.04$, $c_m = 0.60$
- Six returned association rules, two items per rule
- Helped create promotions, strategies to upsell menu items

Applications – Medical

- Chen, Yang, Wang, Shi, Tang, and Li, April 2018 [6]
- Find association between drugs used to treat non-small cell lung cancer and adverse events
- Studied 16,527 patient records, 502 attributes of drug combinations and adverse events
- $s_m = 0.01$ and $c_m = 0.10$





Applications – Medical


- 558 association rules obtained
 - 177 with two items (one drug, one adverse event)
 - 381 with three items (two drugs, one adverse event)
 - Twenty-two rules with high confidence deemed relevant
 - Suggested physicians cautiously prescribe drugs, clinical trials needed to conclude causality
- 

Applications – Environmental

- Jafarzadeh, Mahdavi, and Jafarzadeh in 2017 [7]
- Predict areas with high susceptibility of forest fires in the Ilam Province in Iran
- 53 observations, 52 attributes (geographic features, population densities)
- $s_m = 0.08$, $c_m = 0.80$



Applications – Environmental

- Thirteen association rules returned
 - Seven rules had confidence of 1.00
 - Antecedent size ranged from one to four, consequent ranged from one to seven
 - Strong relation between forest fires and human activity, not geographic features
 - Implement more fire safety campaigns to reduce the number of forest fires
- 

Results

Study	Menu Items [5]	Lung Cancer [6]	Forest Fires [7]	Smartphone Usage [8]	Heart Disease [9]	Flood Areas [10]
Observations	150	16,527	53	100	209	27-50
Attributes	38	502	52	15*	20*	9-40
s_m	0.04	0.01	0.08	0.20*	0.50, 0.60, 0.70	0.25
c_m	0.60	0.10	0.80	0.45*	0.90	0.67
Returned rules	6	558	13	20**	4, 2, 1	4-8
Antecedent size	1	1	1-4	1	1	1
Consequent size	1	1 or 2	1-7	1	1	1
Lift used	No	No	No	Yes	Yes	Yes

* Information not disclosed, estimated from other information in the study


** Researchers limited results, full numbers unknown

Conclusions

- Simple yet robust method, applicable in many areas
- No clear pattern regarding disciplines, s_m , c_m , and dataset dimensions
- Two item association rules seem most common/popular
- Commercial
 - Improved marketing to a customer base constructed by consumer habits
 - Risk of misplaced marketing money
 - Potentially not reflective of big data commercial use [1]




Conclusions

- Medical
 - Conservative when making conclusions
 - Life and death situations form poor medical advice
 - Good starting point for a larger study
 - Environmental
 - Link attributes to mitigate natural disasters, prevention
 - Strong conclusions but also recommend further study
 - Help avoid loss of property and potentially lives
- 



Conclusions

- Susceptible to multiple testing bias
 - Knowing data is very important
 - s_m, c_m too high and miss important association rules
 - s_m, c_m too low and be inundated with association rules
 - Context is important to make informed, sound conclusions
- 

Extensions


- AprioriTID, Agrawal & Srikant [1]
 - Creates another set between C_k and L_k using the transaction ID (TID) to generate itemsets
- Apriori-Hybrid, Agrawal & Srikant [1]
 - Starts with original algorithm, switches to aprioriTID for a specified value of k

Extensions

- Apriori Optimization Algorithm, Chang and Lui [11][12]
 - Calculates inverse distance between two attributes to save computation time and storage
- Apriori Hybrid Method, Chen, et al. [6]
 - Introduces a χ^2 test as a measure of correlation



Future Work

- Previous work focused on the relationship between dataset dimensions and s_m, c_m
 - Potentially find an ideal s_m, c_m given dataset dimensions, context
 - Diminish false-positive results
- 

References

- [1] Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Very Large Data Base*. Santiago, Chile.
- [2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.
- [3] Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2019). arules: Mining Association Rules and Frequent Itemsets. Retrieved from <https://CRAN.R-project.org/package=arules>
- [4] IBM. (n.d.). *Lift in an association rule*. Retrieved from IBM Knowledge Center: https://www.ibm.com/support/knowledgecenter/SSEPGG_10.5.0/com.ibm.im.model.doc/c_lift_in_an_association_rule.html

References


- [5] Kurnia, Y., Isharianto, Y., Giap, Y., Hermawan, A., & Riki. (2019). Study of Application of Data Mining Market Basket Analysis for Knowing Sales Patterns (Association of Items) at the O! Fish restaurant using Apriori Algorithm. *International Conference on Advance and Scientific Innovation*. IOP.
- [6] Chen, W., Yang, J., Wang, H.-L., Shi, Y.-F., Tang, H., & Li, G.-H. (2018). Discovering Associations of Adverse Events with Pharmacotherapy in Patients with Non-Small Cell Lung Cancer Using Modified Apriori Algorithm. *BioMed Research International*.
- [7] Jafarzadeh, A. A., Mahdavi, A., & Jafarzadeh, H. (2017). Evaluation of Forest Fire Risk Using the Apriori Algorithm and Fuzzy C-Means Clustering. *Journal of Forest Science*.
- [8] Mortale, S. L., & Darak, M. J. (2019). Clustering and Pattern Mining of Customer Transaction Data using Apriori Algorithm. *International Journal of Recent Technology and Engineering*.

References

- [9] Mirmozaffari, M., Alinezhad, A., & Gilanpour, A. (2017). Data Mining Apriori Algorithm for Heart Disease Prediction. *International Journal of Computing, Communications & Instrumentation Engineering*, 20-23.
- [10] Harun, N. A., Makhtar, M., Aziz, A. A., Zakaria, Z. A., Abdullah, F. S., & Jusoh, J. A. (2017). The Application of Apriori Algorithm in Predicting Flood Areas. *International Journal on Advanced Science Engineering Information Technology*.
- [11] Qiang, J., & Zhang, S. (2018). Research on Sensor Network Optimization Based on Improved Apriori Algorithm. *Journal on Wireless Communications and Networking*.
- [12] Chang, R., & Lui, Z. (2011). An Improved Apriori Algorithm. *Proceedings of 2011 International Conference on Electronics and Optoelectronics*. Dalian.



Thank You!

- Dr. Santorico, Dr. Austin, and Dr. Spiegler
 - Drew, Dani, and Rebecca
 - Mom and Dad
- 

Applications – Commercial II

- Mortale and Daruk, September 2019 [8]
- Surveyed 100 smartphone users in India, over 15 attributes
- $s_m \approx 0.20$, $c_m \approx 0.45$
- Returned 20 rules based on smartphone brand and smartphone operating system
- Create consumer groups for better marketing

Applications – Medical II

- Mirmozaffari, Alinezhad, and Gilanpour in 2017 [7]
- Find association between patient demographics, symptoms and heart disease
- 209 patient records and over 20 attributes, National Health Ministry in Iran
- Used algorithm three times; $s_m = 0.50, 0.60$, and 0.70 , $c_m = 0.90$
- Seven rules total, all with two attributes

Applications – Environmental II

- Harun, Makhtar, Aziz, Zakaria, Abdullah, and Jusoh in 2017 [10]
- Determine flooding areas and patterns in seven districts in Malaysia
- $s_m = 0.25$, $c_m = 0.65$, lift calculated
- 45 rules total, all size two
- Allocate resources and implement an early-alert system

District	Observations	Attributes
Setiu	36	26
Marang	27	9
Kemaman	38	34
Besut	50	40
Dungun	47	12
Hulu Terengganu	49	38
Kuala Terengganu	44	29



Previous Work

- Original plan: simulate correlated binary data, run algorithm multiple times with varying s_m, c_m
 - Find optimal s_m, c_m based on different store sizes
 - Creating correlated binary data is difficult
 - Computation time was very long
- 