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

By

EMMA M. COLLINS

B.A., Mathematics, Humboldt State University, 2018

A report submitted to the

Department of Mathematical and Statistical Sciences

at the University of Colorado Denver in partial fulfillment

of the requirements for the degree of

Master of Science

Statistics

Summer 2020

This project is for the Master of Science degree by

Emma M. Collins

has been approved for the

Statistics Program

by

Dr. Stephanie Santorico, chair

Dr. Erin Austin

Dr. Adam Spiegler

Abstract

The apriori algorithm was originally created for use on transaction data, large datasets that record the items bought together in individual purchases over a set length of time. Since its development in the mid 90's, the apriori algorithm has expanded into various areas of research. The algorithm finds groupings of items frequently purchased together through use of joint and conditional probabilities and pruning. The pruning method removes sets of items that do not reach user-picked thresholds to avoid longer computation times. The apriori algorithm has been used in commerce, medical, and environmental fields, where items sold in a store are replaced by purchasing habits, patient demographics and symptoms, or geographical features, respectively. Each topic has unique context with varying user-picked thresholds and different dimensions: we expect more observations of rainfall in a specific tropic climate than the number of patients diagnosed with a disease at a specific hospital. Implications of incorrect analysis vary from inaccurate weather forecasts to life or death scenarios. These parameters and frameworks are explored and compared for each area of study after introducing and explaining the apriori algorithm. Lastly, a brief introduction to improvements and hybrid methods of the algorithm is provided.

Table of Contents

1. Introduction

In 1994, growth in bar-code technology allowed for store transactions to be collected and formed into very large databases. This led to the question, "what items in a store are frequently bought together?" Consider all the items in a grocery store and all the transactions in any given day, week, month, or year; the number of calculations needed to answer this question is very large. As a result, Rakash Agrawal and Ramakrishnan Srikant, both IBM employees at the time, introduced the apriori algorithm in 1994.

The apriori algorithm is an unsupervised, clustering method that uses user-selected thresholds of joint and conditional probabilities to return sets of items often purchased together. The phrase "*a priori*" is Latin for "from the former;" thus, the apriori algorithm reduces the number of calculations by using a looped process that reduces the dataset size based on information in the previous iteration (Miriam-Webster, n.d.). As a result, the algorithm has faster computation time than its predecessors, the AIS algorithm and SETM algorithm (Agrawal & Srikant, 1994).

Over 25 years later, the apriori algorithm has expanded to various areas of research. While the algorithm remains the same, the context and usage of the algorithm can vary greatly. Argawal and Srikant even stated in their original publication, "what is useful or interesting is often application dependent." For example, consider the use of the algorithm to determine what dishes and drinks in a restaurant are bought together and also to ascertain what adverse effects of medicine appear together in lung cancer patients. While there may be only 50 menu items in a restaurant there may be hundreds of drugs and harmful medical conditions when studying cancer patients. When using the apriori algorithm, the level of desired confidence in the results will also differ. While a restaurant owner wants to be sure of a promotion before spending money on marketing, a decision a doctor or clinician makes regarding medication and potential side effects can have life or death consequences; thus, a higher level of confidence is mandated.

In this paper, the apriori algorithm is introduced and explained before exploring applications in commercial, medical, and environmental fields to demonstrate the algorithms flexibility in diverse disciplines. The parameters and context of each area are compared and discussed. We conclude the paper with a brief introduction to hybrids and extensions of the apriori algorithm.

2. Methodology

The apriori algorithm has a very simple requirement; the data needs to be binary or be able to be transformed into binary data. For example, a categorical variable with three responses, such as party affiliation, are split into three new dummy variables; Democrat, Republican, and Independent. If a variable is continuous, such as age, it must be split into categories and then a dummy variable is created for each category (e.g., under 30 years old, between 30 and 60 years old, and above 60 years old). Observations either have an attribute (1) or they do not have an attribute (0). If data is missing, e.g., a person were to leave a question blank in a survey, the observation will have a 0 in place of all responses. For example, if someone does not disclose their party affiliation, there will be a 0 for Democrat, a 0 for Republican, and a 0 for Independent for the single observation. Besides the binary data requirement, the apriori algorithm has no assumptions or pre-filtering steps.

2.1 Association Rules

The apriori algorithm mines association rules found in datasets. Association rules are joint values of two sets of attributes, A and B, characterized by probabilities referred to as support and confidence (Hastie, Tibshirani, & Friedman, 2009). These disjoint sets are called itemsets. An association rule between itemsets, A and B, is denoted as

$$A \rightarrow B[support, confidence],$$

where A is referred to as the antecedent and B is referred to as the consequent.

Let X and Y be two indicator variables for A and B, respectively, and let n be a single observation out of N total observations. Then X and Y can be expressed as,

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

and,

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

For example, let $A = \{milk, bread\}$ and $B = \{chocolate\}$. If chocolate and bread are bought in the tenth transaction of the day but milk is not bought in the same transaction, then $X = 0$ and $Y = 1$.

Support is how often A and B appear together in N observations. That is,

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

Support is sometimes defined as a count, simply $frequency(X, Y)$, but support as a proportion is more common. Confidence is how often an itemset, B, appears given another itemset, A, is already observed. That is,

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

Note,

$$support(A \rightarrow B) = support(B \rightarrow A),$$

but

$$confidence(A \rightarrow B) \neq confidence(B \rightarrow A).$$

When calculating support and confidence, two random indicator variables, X and Y, are used to determine the frequency in which the entirety of the sets A and B appear together. Since these variables are taken from a sample of data, the exact proportions for support and confidence are unknown and the joint and conditional probabilities are estimates.

In addition, depending on the software and packages used, it is possible for the antecedent to be the empty set, as seen in the R package, *arules* (Hahsler, Buchta, Gruen, & Hornik, 2019). In this instance,

$$support(\emptyset \rightarrow A) = \hat{P}(X = 1) = confidence(\emptyset \rightarrow A).$$

However, in practice, it is possible to set parameters for the algorithm so only sets of certain sizes are considered.

Another measure frequently calculated after finding all association rules is called lift. Lift can be thought of as a type of correlation between itemsets or the importance of the association rule (IBM, n.d.). Lift for an association rule of two itemsets can be calculated by

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

If lift is equal to 1, there is no correlation between itemsets; the attributes appeared together by chance. If lift is greater than one, there is positive correlation between itemsets; the appearance of one attribute encourages the appearance of the other(s). If lift is less than one, there is negative correlation between itemsets; the appearance of one attribute discourages the other(s). While lift can technically be any number greater than zero, in practice it is generally between 0 and 4. Now there is an understanding of support, confidence, and lift we can detail how the apriori algorithm works and its benefits.

2.2 Apriori Algorithm

Consider a dataset with N observations and K attributes. If we wanted to consider all possible combinations of association rules, of any size, calculating the support and confidence for each rule would be a massive undertaking. All association rules with only two attributes would be over 2^K calculations. Not only is this a big number, it is computationally expensive and would take a lot of time.

The apriori algorithm uses thresholds for support, confidence, and pruning to drastically minimize the number of calculations and computation time. The user sets minimum levels for support, s_m, and confidence, c_m. Itemsets that meet the minimum support level are called large itemsets, L_k, and new potentially large itemsets are called candidate itemsets, C_k, where k indicates the number of attributes in the itemset (Agrawal & Srikant, 1994). The algorithm executes the following steps:

1. Generate all large itemsets of size 1, L_1.
2. For $k \geq 2$, generate C_k by finding all possible combinations of itemsets of size k from itemsets in L_{k-1}.
3. Generate L_k by calculating the support for the itemsets in C_k and removing those with support less than s_m.
4. Repeat steps 2 and 3, increasing k by one for each pass until the maximum size of the itemsets is reached or until the minimum support is no longer reachable.
5. Generate potential association rules from all L_k and calculate confidence for all large itemsets and return those greater than or equal to c_m.

The large itemsets returned after step five are the association rules of the dataset, given s_m and c_m. An additional step can be added after step five where the lift of the association rules is calculated. This iterative process removes unnecessary computations by calculating one probability for one itemset and removing anything below the minimum support before calculating probabilities for bigger itemsets.

2.3 Example

Consider a grocery store with the following five transactions, seen in Table 1:

Table 1 A simple example of grocery store transactions

Transaction	**Items**
001	milk, bread, cheese
002	milk, butter
003	butter, cheese
004	milk, bread, butter, cheese, chocolate
005	milk, bread

Table 2 displays how this data would look converted into the binary format required by the apriori algorithm. Let $s_m = 0.25$ and $c_m = 0.8$. To perform the apriori algorithm, first find L_1, the set of single items with support greater than or equal to $s_m = 0.25$ (step 1):

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

Table 2 The transaction data in binary format.

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

We then find C_2, seen in Table 3, by taking all possible combinations of items found in L_1.

Table 3 Candidate itemsets of size 2, found from large itemsets of size 1.

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

L_2 is found by removing itemsets from C_2 less than $s_m = 0.25$. Thus:

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

We then find C_3, displayed in Table 4, by finding all combinations of items from L_2, excluding any itemset containing both bread and butter (recall, $P(A \cap B) \leq P(A)$.)

Table 4 Candidate itemsets of size 3, found from large items of size 2.

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

Thus, $L_3 = \{milk, bread, cheese\}$. There are no itemsets of size 4 that meet the minimum support, since we cannot include milk, cheese, and butter in an itemset. We then generate all association rules and calculate the confidence, seen below in Table 5.

Table 5 Potential association rules generated from all large itemset combinations.

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

Then lastly, we discard any potential association rules with confidence less than $c_m = 0.8$, with the remaining rules listed in Table 6. For $s_m = 0.25$ and $c_m = 0.8$, the resulting association rules are:

Table 6 Association rules of with s_m=0.25 and c_m=0.8 from the transaction dataset from Table 1 and Table 2.

Antecedent	**Consequent**	**Support**	**Confidence**	**Lift**
{}	{milk}	0.80	0.80	1.00
{bread}	{milk}	0.60	1.00	1.25
{bread, cheese}	{milk}	0.40	1.00	1.25
{milk, cheese}	{bread}	0.40	1.00	1.67

In practice, with larger datasets and more association rules returned, the apriori algorithm can have additional parameters besides s_m and c_m. Common parameters are the minimum and maximum number of items in an itemset and how to display the resulting association rules (i.e., descending confidence). Although the algorithm may return outwardly similar rules, seen in the last two association rules above, dependent on the application, these redundancies can have nuances that may change a user’s decision on what to do with the algorithm results. These contextual differences are explored in the various applications of the apriori algorithm.

3. Applications

As stated earlier and seen in the example, the original use for the apriori algorithm was to determine what items were frequently purchased together. However, the algorithm can also be applied to commercial, medical, and environmental research.

3.1 Commercial Applications

Use of the apriori algorithm for commercial reasons is closely related to its original purpose to learn consumer purchasing habits. The commercial use does not have to be strictly limited to items in a store. Studies summarized below analyze consumer demographics and smartphone purchasing habits and explore relationships between food and drinks ordered at a restaurant.

Published in September 2019, Mortale and Darak combined over 500 transactions from 100 individuals in India to determine smartphone purchasing habits. Their goal was to compare customer smartphone brand loyalties and determine how to divide customers into specific customer groups to increase product sales. While the total number of attributes is not disclosed, there are at least 15 attributes involving respondent age, respondent sex, smart phone usage (i.e., business or social), current smartphone brand, next smartphone brand, current operating system, and next operating system. The minimum support and minimum confidence used is also not shared, but given the resulting association rules, we assume $s_m \approx 0.20$ and $c_m \approx 0.45$.

Mortale and Darak used different limitations on what association rules were returned to gain insight on smartphone consumers, as well as calculating lift for returned association rules. They filtered results twice based on association rules related to smartphone brands and association rules related to operating system, with ten rules for each limitation. All twenty rules had only two items, one as the antecedent and one as the consequent. Mortale and Darak concluded Samsung, Redmi, and Apple branded smartphones are more attractive to male customers, Redmi is more desirable to consumers between 23 and 32 years old, and current Android users are likely to purchase another Android as their next smartphone. These conclusions are based off lift; each association rule related to their conclusions had a lift over 2.5, while other rules had lift values ranging from 1.58 to 2.36. However, of the 100 respondents, 56 identified as men and 44 identified as women, and the age group 23-32 years old contained 47% of consumers. While sex is split relatively evenly, the domination of one age group may have influenced what association rules are returned.

Another study of the apriori algorithm in a commercial field was published in 2019 by Kurnia, Isharianto, Giap, Hermawan, and Riki. Their goal was to determine sales patterns of the O! Fish restaurant in Jakarta, Indonesia to determine promotional strategies for upselling items. They randomly selected 150 transactions across 38 menu items consisting of 23 food dishes and 15 beverages with $s_m = 0.04$ and $c_m = 0.60$. Lift was not calculated.

Five association rules were returned in total, with each antecedent and each consequent consisting of a single menu item, organized by confidence ranging from 0.60 to 1.00. These rules aligned with previous promotional strategies applied manually, but with less time and computation.

3.2 Medical Applications

The use of the apriori algorithm in the medical field is a large departure from its original intention, with more complex results and careful conclusions. Some studies involve using the apriori algorithm to predict heart disease and discover relationships between drugs used to treat lung cancer and potential adverse events.

In 2017, Mirmozaffari, Alinezhad, and Gilanpour used the apriori algorithm for heart disease prediction. The original purpose of the study was to compare the performance of the apriori algorithm in MATLAB and WEKA (Waikato Environment for Knowledge Analysis) to help medical analysts conclude which software to use for better prediction accuracy. It contains 209 patient records, used under supervision from the National Health Ministry in Iran, and 8 variables divided into over 20 attributes. These variables include age, chest pain type, resting blood pressure, blood sugar, resting electrocardiographic, maximum heart rate, exercise angina, and heart disease diagnosis.

The apriori algorithm was completed three times with varying levels of s_m; $s_m = 0.5, 0.6,$ and 0.7 while the minimum confidence remained constant at $c_m = 0.9$. Lift was also calculated. The smallest minimum support returned four association rules, the middle minimum support returned two association rules, and the largest minimum support returned one association rule. All returned rules had a single attribute in both the antecedent and consequent. Given all returned association rules, lift was relatively close to 1, with a maximum deviation of 0.03. Mirmozaffari, et al. (2017), concluded MATLAB had slightly faster computing times and was therefore better in predicting heart disease in patients.

A large study published in April 2018 from Chen, Yang, Wang, Shi, Tang, and Li reported adverse events from various drug treatments in patients with Non-Small Cell Lung Cancer (NSCLC). In 2012, China reported over 700,000 new cases of lung cancer, with a five-year survival rate at 15% (Chen, et al., 2018). Since cancer patients are typically treated with multiple drugs, including chemotherapy, Chen, et al., aimed to detect adverse drug reactions based on combinations of different treatments prescribed to patients. Using medical records from 2010 to 2016, there were 16,527 patients with over one million combined prescriptions. The prescriptions were combined per patient and ingredients along with adverse events, such as high cholesterol and anemia, to 502 attributes. Minimum support and confidence levels of $s_m = 0.01$ and $c_m = 0.10$ were used and lift was not calculated.

A total of 558 association rules were obtained, with results filtered and observed based on the number of drugs related to an adverse event. 177 association rules related to a single drug and adverse events while 381 association rules were related to two drugs and adverse events.

Chen, et al. (2018), also developed and used a hybrid apriori algorithm to conclude twenty-two strongly correlated association rules with high confidence; the hybrid algorithm is discussed later. These top rules consisted of one adverse effect as the antecedent and one drug as the consequent for sixteen association rules and one adverse effect as the antecedent and two drugs as the consequent for six association rules. The researchers advised physicians cautiously prescribe drugs for treatment and suggested retrospective or prospective clinical investigations to further study causality.

3.3 Environmental Applications

Studies using the apriori algorithm in the environmental fields revolve around natural disasters: flooding and forest fires. Each study focuses on what attributes lead to their respective disaster with the goal of implementing preventative protocols.

In 2017, Harun, Makhtar, Aziz, Zakaria, Abdullah, and Jusoh published research that used the apriori algorithm to predict flood areas in Malaysia. With annual rainfall averaging 3000 meters per year and over 189 river basins, flooding is the most common and destructive natural disaster for the nation (Harun, et al., 2017). Their goal was to identify associations between water levels in basins with flood areas to help allocate resources for flood management and create an early warning system for residents. Harun, et al. (2017), completed the apriori algorithm for 7 different districts, each with varying numbers of observations and attributes seen in Table 7, with collected data from November 2009 to January 2015.

Table 7 The number of observations and attributes for each of the seven districts studied.

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

The attributes are villages that experienced flooding during an observation. The lowest minimum support and minimum confidence used was $s_m = 0.25$ and $c_m = 0.67$, respectively, with lift calculated as well. Each district returned four to eight association rules, with forty-five association rules in total, each with only one attribute in both the antecedent and the consequent. The confidence of all but seven association rules had confidence of 1, and lift ranged from 1.5 to 4.25, indicative of strong, highly correlated association rules. Harun, et al. (2017), were able to conclude that the likelihood of two villages flooding is low, because of low support, but when one floods, it is very likely the other village will flood because of high confidence and lift.

In 2017, Jafarzadeh, Mahdavi, and Jafarzadeh evaluated forest fire risk in the Ilam Province in Iran, using the apriori algorithm. The Ilam province is on the western side of Iran, home to many forests and settlements susceptible to many fires. Using data from 2010 to 2015 from 53 fires in the province, Jafarzadeh, et al. (2017), aimed to discover what land attributes and population demographics appeared most often to implement protective and preventative measures to save property and lives. There were 12 variables consisting of elevation, slope, aspect, distance from settlements, roads, farmland, and rivers, population density, temperature, precipitation, land use, and standing dead oak trees, split into 52 attributes.

Minimum support and minimum confidence were $s_m = 0.08$ and $c_m = .80$, lift was not calculated. Thirteen association rules were returned, seven of which had confidence of 1.0. The association rules had differing antecedent and consequent sizes, ranging from one to four attributes and one to seven attributes, respectively. One association rule with 0.8 support and 1.0 confidence showed strong relationship between wildfire occurrence with close distance to settlements, high population density, close distance to roads, steep slope, many standing dead oak trees, high temperature, land coverage, and close distance to farmlands. Jafarzadeh, et al. (2017), concluded natural factors in forest wildfires are less significant than human activity, largely due to dependence on the forests as a resource and recreational area in addition to lack of fire-safety awareness.

4. Results

After reviewing six studies in three different fields, the apriori algorithm can be applied to many different datasets, with differing dimensions and parameters. These differing parameters are summarized in Table 8. The number of observations and attributes varies from 53 to over 16,000 and 9 to 502, respectively. Minimum support values range from 0.01 to 0.7 while minimum confidence values range from 0.10 to 0.90. Each area had one study that calculated lift and used the statistic to make conclusions and one study that did not calculate lift. The only shared feature is five out of the six studies focus on itemsets of size two, with one attribute in both the consequent and the antecedent.

Table 8 Dimensions, parameters, the number of association rules, size of returned association rules, and if lift was used for each study observed.

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

*Information not disclosed but estimated from other information in the study.
**Researchers limited results, full number of rules returned is unknown.

5. Discussion

5.1 Conclusions

The apriori algorithm is a straightforward yet robust method for data mining and analysis. Comparing the algorithms use in commercial, medical, and environmental disciplines, the only clear pattern of the apriori algorithm focuses on association rules involving only two attributes, one attribute in the antecedent and one attribute in the consequent likely for simple interpretability. However, one thing remains clear: context behind the apriori algorithm is very important.

In commercial applications studied, the number of observations and attributes of the datasets used are surprisingly low. In the creation of the apriori algorithm, it was tested on datasets ranging from 100,000 to one million observations (Agrawal & Srikant, 1994). It is likely, most large datasets are considered proprietary information which most corporations are unlikely to share, much less publish their findings. While the number of observations and attributes studied may not reflect most commercial applications, the goals and consequences do. Both studies made definitive conclusions based on the resulting association rules. The study involving smartphone consumers concluded various smartphone brands were more appealing to men and younger age groups in addition to brand loyalty among Android users. The study of the Indonesian restaurant determined six association rules to help implement promotions in the restaurant. Acting on these conclusions have similar results; improved marketing to a customer base constructed from consumer habits.

The use of the apriori algorithm in the medical field is more conservative when making conclusions. Neither study made direct conclusions about causality nor recommended immediate action. The study on heart disease prediction concluded there are better software programs for running the algorithm to help physicians make diagnoses while the study on adverse events in lung cancer treatment suggested their results were a good starting place but clinical trials were needed for further evidence. Since lives are potentially at risk with one false association rule, the algorithm is often used as a preliminary step to conduct much larger, intensive studies.

The environmental studies lie between commercial and medical applications in terms of consequences. The studies evaluating forest fire risk and predicting flood areas had very similar goals; find association rules linking attributes to help mitigate natural disasters and implement preventative planning. Each study made clear conclusions and suggest implementing protective measures but also suggest more research.

The apriori algorithm is a simple, yet effective method to find commonly occurring attributes. It can be used in a wide range of disciplines with the only requirement being binary data. It can be used as a precursory method to a deeper analysis, as seen in medical fields, or used alone as observed in commercial fields for marketing strategies.

The apriori algorithm is not a complicated method; it involves probabilities taught in an introductory-level college statistics course (including MATH 2830 at CU Denver). However, its simplicity also necessitates caution. It is susceptible to multiple testing bias because it is easy to rerun with different parameters to find “interesting” results. Conversely, users may also have to use the algorithm to fine tune parameters. Having minimum support and minimum confidence too low can result in being inundated with association rules, but if the minimum values are too high, the user could miss potentially relevant and important association rules. Like all statistical methods, it is crucial to understand the context of the data and make responsible conclusions.

5.2 Extensions and Future Work

The simplicity of the method has allowed for the apriori algorithm to have extensions and hybrid methods. There are many modifications to increase computation speed, with minor adjustments to the apriori algorithm, and while these modifications are helpful, they are not very interesting. However, there are three extensions to the apriori algorithm worth noting to suit different research needs.

The first method is the aprioriTID algorithm, also introduced by Agrawal and Srikant in 1994 in the same publication as the original apriori algorithm. It is very similar to the original; however, it creates another set between generating C_k and L_k. The set, $\overline{C_k}$, is the set of candidate itemsets of size k where a transaction identification (TID) is used to reduce the number of calculations performed for each pass of the algorithm. The aprioriTID algorithm tends to work faster for greater values of k, whereas the original apriori algorithm performs faster for lower values of k (Agrawal & Srikant, 1994). Therefore, Agrawal and Srikant (1994) also created the aprioriHybrid algorithm. The hybrid algorithm switches from the original apriori algorithm to the aprioriTID algorithm for a specified value of k. The hybrid algorithm can improve computation time but requires more computer memory to run.

Another extension is the apriori optimization algorithm. Introduced by Chang and Lui in 2011, the improved apriori algorithm uses a hash structure in two-dimensional space and calculates the inverse distance between two attributes to save computation time and storage space. It has been used with wireless sensor network datasets to increase network coverage as well as other applications in electronics (Qiang & Zhang, 2018).

The last apriori extension was created by Chen, et al. (2018), in the same study of discovering adverse events with drugs in patients with lung cancer. They introduce a χ^2 test into the algorithm to create a new parameter, called the minimum test value, determined by significance level and degrees of freedom. The minimum test value determines whether there is positive or negative correlation between the antecedent and the consequent of the association rules, similar in concept to lift.

Future work could feature more analysis on relationships between database dimensions and minimum support and minimum confidence values. From the six publications examined, there is no obvious pattern between apriori parameters and the dimensions of the data, much less any trend in their respective fields. Earlier work, prior to this iteration of the project, focused on simulating correlated binary data to examine this relationship, but due to time restrictions and computing power was unfeasible here. The next step, given more time, would be to determine how different parameter sizes and different dimensions change the resulting association rules to establish ideal minimum support and minimum confidence based on database dimensions to diminish false positive association rules.

6. References

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Very Large Data Base.* Santiago, Chile.

Chang, R., & Lui, Z. (2011). An Improved Apriori Algorithm. *Proceedings of 2011 International Conference on Electronics and Optoelectronics.* Dalian.

Chen, W., Yang, J., Wang, H.-L., Shi, Y.-F., Tang, H., & Li, G.-H. (2018). Discovering Associations of Adverse Events with Pharmacotherapy in Patiends with Non-Small Cell Lung Cancer Using Modified Apriori Algorithm. *BioMed Research International*.

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

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*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.* New York: Springer.

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

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*.

KP, M., Singh, R. K., & Kumar, S. S. (2012). Apriori-Hybrid Algorithm as a Tool for Colon Cancer Microarray Data Classification. *International Journal of Engineering Research and Development*.

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.

Miriam-Webster. (n.d.). *A priori*. Retrieved from Miriam-Webster.com Dictionary: https://www.merriam-webster.com/dictionary/a%20priori

Mirmozaffari, M., Alinezhad, A., & Gilanpour, A. (2017). Data Mining Apriori Algorithm for Heart Disease Prediction. *International Journal of Computing, Communications & Instrumentation Engineering*, 20-23.

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*.

Qiang, J., & Zhang, S. (2018). Research on Sensor Network Optimization Based on Improved Apriori Algorithm. *Journal on Wireless Communications and Networking*.