Assignment 1: Association Rules Analysis Using R

Written by Daanish Ahmed

DATA 630 9041

Semester Summer 2017

Professor Edward Herranz

UMUC

June 7, 2017

Introduction

Association rules are an extremely valuable component to the data mining process because they provide useful information on the relationships between variables. These rules consist of patterns within datasets that reveal certain itemsets frequently appearing together (Han, Kamber, & Pei, 2011). As a result, it is important to be able to mine these patterns from the data and determine whether they indicate strong relationships between the variables. However, a major obstacle to this task is the amount of information to process and the resulting impact it can have on a computer's performance. Therefore, one useful method for association rules mining is the Apriori algorithm, which creates a list of frequently-occurring association rules. This algorithm checks to see if an itemset has a support (i.e. frequency) that is greater than the minimum support value, and it will only include items and their subsets if they meet this minimum value (Tan, Steinbach, & Kumar, 2015). Using the R programming language, I will examine a dataset containing information on burn victims, and I will use the Apriori algorithm to generate a list of association rules. I expect that the results of my analysis will provide vital information on the causes of certain burn cases as well as identifying factors which influence burn severity. Such findings may greatly contribute towards saving lives in the future.

Data Exploration

The first step of my analysis is to describe the exploration of the original dataset. This dataset contains 9 variables and 1000 observations (see Figure 1 in Appendix B). The variables include the record ID, treatment facility, whether the patient is alive or dead, the patient's age, gender, race, total burn surface area, whether the burn involved an inhalation injury, and whether

flame was involved in the injury ("Code Sheet for Variables," 2014). Of these variables, the ID and treatment facility are integers while the age and total burn surface area are numeric variables. The remaining five variables are factors, each with two levels (see Figure 1 in Appendix B). I then generated a set of descriptive statistics which provide some interesting facts regarding the variables (see Figure 2 in Appendix B). The data shows that over 70% of all victims are male, the average victim age is around 32, and only a slight majority of burn incidents involve direct contact with flames. Additionally, these statistics indicate that there are no missing values in any of the variables. These findings are helpful, but there are more valuable pieces of information within this dataset that can be found by mining the association rules.

Preprocessing

Prior to implementing the Apriori algorithm, it is necessary to perform preprocessing steps on the data to ensure that the dataset is optimized for analysis. Before I started exploring the dataset, I replaced some of the values to make the data easier to understand. Several variables, such as race and gender, initially had numeric values which represented categorical text values (such as 0 for female and 1 for male). Using the provided code sheet ("Code Sheet for Variables," 2014), I chose to replace such instances with their actual values to allow readers to easily understand my data. Next, I proceeded to remove unnecessary variables from the dataset. Certain variables such as unique identifiers (IDs) are not useful for the analysis, and removing them will improve the performance of the data mining process ("Data Preprocessing," 2017). Since the first column is an ID, I chose to remove it. The next step is to handle any missing values that appear within the data. However, I found that no missing values existed within any of the variables (see Figure 3 in Appendix B). The final preprocessing step involves the discretization of all numeric

variables. The Apriori method will not function correctly unless all variables in the dataset are discrete ("Association Rule Mining," 2017). Due to this, it is necessary to convert the three remaining numeric variables (facility, age, and total burn surface area) into factors. To do this, I utilized the "discretize" method in R which requires the "arules" package to be installed. I chose to discretize these variables into six categories based on equal frequency, which divided them into six intervals with approximately equal numbers of items in each bin (see Figure 4 in Appendix B). After completing this step, the data is ready for implementation with the Apriori algorithm.

Algorithm

The Apriori method in R allows for the input of several parameters to adjust the algorithm's output. Entering values for support and confidence sets the minimum threshold for these metrics, meaning that any rules with support or confidence values below the input will be omitted from the list. Adjusting the minimum length will set the lowest number of items that are allowed in each rule (consisting of the left and right-hand sides combined). Setting this value greater than or equal to 2 will also prevent rules with blank itemsets from appearing on the list. The input values that I used for this algorithm are a support of 0.15, a confidence of 0.85, and a minimum length of 2. Through testing, I found that choosing parameters close to the default values (0.1 support and 0.8 confidence) produced the best results with the dataset I was using. Choosing parameters that were much higher than the default rates resulted in the algorithm returning a much smaller set of rules. After removing redundant rules, I noticed that there would be too few rules to conduct a meaningful analysis. On the other hand, setting the parameters lower than the default values resulted in too many rules with low frequency or accuracy rates. But by setting the parameters to

be only slightly greater than the default values, I ended up with a reasonably-sized list of 86 association rules (see Figure 5 in Appendix B).

From here I sorted the rules by lift in descending order, meaning that the rules displayed at the top have the strongest positive correlations between their itemsets. To refine the list even further and improve system performance, I pruned the list of all redundant rules. I achieved this by searching the list for redundant rules and storing each of these rules in an array. I then created a new list containing every single rule except for those found in the array. The top ten rules in the new list (see Figure 8 in Appendix B) contain many of the rules that had existed in the original sorted list but without the redundant ones (see Figure 6 in Appendix B)—making the new list more concise and interesting. The descriptive statistics for this list offer some useful facts and reveal that there are 44 rules remaining (see Figure 7 in Appendix B). These statistics show that the median values for support and confidence are about 21% and 95% respectively, and that the median lift is around 1.12. Now that the list is ready for analysis, I will proceed to examine some of the relationships between the items in this dataset.

Initial Results

According to the top ten items in the list, the most likely age group to be involved in burn accidents are infants or toddlers under the age of 3 (see Figure 8 in Appendix B). Also, none of the top rules involve lethal burns, most do not involve direct contact with flames, and most do not feature injuries related to inhalation. These findings make sense logically, since they imply that most burn injuries are not life-threatening. According to Durtschi, Kohler, Finley, and Heimbach (1980), the majority of burn injuries on young children are small. This claim supports the fact that

many of these rules involve a relatively low total burn surface area between 0.1% and 2.1%. The top four rules in this list have lift values close to 2, indicating that they each contain strong positive correlations between their itemsets. However, the other rules in the list have lift values near 1.15—implying that their correlations are much weaker. The descriptive statistics for these rules indicate that the median lift value is only around 1.12 (see Figure 7 in Appendix B), which suggests that the majority of these rules have very weak correlations.

On the other hand, most of the top ten rules have confidence values between 94% and 100%, indicating that these rules are extremely accurate. Yet when examining the support values, it is evident that most of these rules have values that are only slightly higher than the minimum support threshold of 0.15. Though these rules are still considered "frequent" according to the Apriori algorithm, their strength is somewhat diminished due to low lift and support values. I created a scatterplot to further examine the relationship between confidence, support, and lift (see Figure 9 in Appendix B). The image shows that the highest lift values corresponded with the lowest allowed support and confidence values. I would still argue that the strongest rules in this list are the four rules with the highest lift values. Even though these rules have the lowest confidence and support rates, their confidence values are still extremely high (between 91% and 95%), and their support values are indeed higher than the minimum threshold. However, we are likely to discover more useful insights by mining rules that contain specific itemsets.

Analysis of Death Cases

One of the cases that I wish to focus on is the possibility of a burn injury resulting in death. My initial list was lacking in any rules resulting in dead victims—as evidenced by the fact that

none of the rules contained the item {DEATH=dead} on the right-hand side (see Figure 8 in Appendix B). This is because dead victims occur very rarely in this dataset, and thus death cases were omitted from the original list due to their low support values. Because of this, I had to create a new rules list with the Apriori method while using a lower minimum support threshold. I selected a support of 0.05, a confidence of 0.5, and a minimum length of 2. The consequence of choosing lower metric values is that the resulting rules are likely to be weaker due to lower frequency and accuracy rates. But if I do not perform this procedure, then I might miss some potentially useful information regarding the causes of death in burn cases. After creating this list, I sorted the rules by lift and removed all redundant rules through the same procedure that I had followed earlier. These steps resulted in a list of eight association rules.

By looking at this list, we can immediately see some of the potential causes of death (see Figure 10 in Appendix B). The first rule on the list indicates a death caused by direct contact with flames, an injury due to inhalation, and a total burn surface area between 22.7% and 98.0%. Most of the other rules in this list have at least one of these items in their itemsets, suggesting that these might be some of the most common conditions found on burn fatalities. These eight rules have lift values ranging between 3.50 and 4.94, indicating that there are very strong positive correlations between the itemsets in each rule. Based on this finding alone, these rules seem quite useful for analyzing the causes of death. However, the confidence values only range between 53% and 74%, meaning that the rules are not always accurate. Furthermore, the highest level of support is around 9.8%, indicating that most of these rules seldom occur. To further illustrate the difference between lift and support among these rules, I created a grouped plot that compares the two metrics (see Figure 11 in Appendix B). The image shows that support does not necessarily correspond with lift, since the rules with higher lift values tend to have lower support values and vice versa.

Altogether, these findings raise doubts as to whether these rules are reliable for analytical purposes. However, this does not necessarily mean that the rules should be disregarded. It is quite likely that the low support and confidence values are due to the relatively small sample size of data. The current dataset features 1000 records with only 150 cases of deaths (see Figure 2 in Appendix B). If the dataset had more observations and a larger collection of death instances, it is possible that at least the support values would be higher.

Analysis of Inhalation Injuries

In addition to examining the causes of death, I also looked at all rules concerning inhalation injuries. As was the case when analyzing victim deaths, I had to create a new rules list for this itemset due to the shortage of records involving inhalation injuries. To create this list, I included the item {INH_INJ=yes} on the left-hand side rather than on the right because I wanted to see the results of what an inhalation injury can lead to. Once again, I used low support and confidence values of 0.05 and 0.5 respectively in order to allow as many rules into the list as possible. After pruning this list, the code returned a total of five association rules (see Figure 12 in Appendix B). These rules suggest that inhalation injuries often imply a direct contact with flames, the victim's death, and or a total burn surface area between 22.7% and 98.0%. These results are supported by Lafferty (2017), who claims that inhaling smoke is the leading cause of death during fire incidents. Morrow et al. (1996) also suggest that inhalation injuries result in much larger burns, which supports my findings regarding the total burn surface area. The remaining two rules in the list imply that inhalation injuries are most associated with Caucasian males. However, this is most likely due to the fact that the grand majority of victims in this particular dataset are male, while a slight majority are Caucasian (see Figure 2 in Appendix B).

The metrics for these rules are relatively similar to the values found within the victim death cases (see Figure 12 in Appendix B). The confidence values vary from 59% to 95%, and only two rules have accuracy rates that are higher than 70%. The highest support value in this list is 11.6%, suggesting that all of these rules are relatively infrequent. The first two rules have high lift values near 4—indicating strong positive correlations between the itemsets—while the third rule has a lower lift value of 1.8. From here, I created a circle graph to see the relationships between the metrics for these rules (see Figure 13 in Appendix B). This graph once again indicates that support and lift are not necessarily related, since higher lift values often correspond with lower support values and vice versa. Altogether, some of these metrics may appear to make these rules appear weaker and less useful for analysis. However, many of these findings are supported by other sources, and therefore it is possible that the low confidence or support values are due to the small sample size. Inhalation injuries only make up 122 out of 1000 cases in this dataset (see Figure 2 in Appendix B), and thus it is likely that using a larger sample size of data can produce results which are more accurate.

Conclusion

The results of my analysis provide several insights that can be potentially useful towards identifying the causes of burns and determining burn severity. The initial rules suggest that infants and toddlers are the most likely age group to be involved in burn incidents. Furthermore, most burns are non-lethal, involve no inhalation injuries, and result in minimum burn surface areas. These initial rules are very accurate but do not always produce the strongest correlations between their itemsets. The next component of my analysis consisted of studying the factors that contributed to victim deaths. I found that deaths are usually caused by inhaling smoke, direct

contact with flames, and suffering a high total burn surface area. The final part of the study involved the analysis of inhalation injuries and the consequences of having such conditions. These rules suggest that inhaling smoke often corresponds with higher burn surface areas, the presence of flames, and a higher likelihood of death. The findings thus indicate that smoke inhalation is strongly connected to death due to burn injuries. For both cases, the rules tend to have high lift values—implying that these rules consist of strong positive correlations. However, both sets of rules suffer from lower support and confidence rates, suggesting that these rules are relatively weak and might not be ideal for analysis.

Nevertheless, I would still argue in favor of at least considering these rules for further analysis. I believe that one of the reasons for the low support or confidence values is that there are relatively few cases of death or inhalation injuries within my dataset. Due to the scarcity of such records, it is understandable that the support values would be low. One suggestion towards overcoming this limitation is to use a larger sample size of records, or even use a dataset that focuses predominantly on death cases or inhalation injuries. Featuring a larger number of relevant observations would certainly increase the frequency of the rules, and it may affect the accuracy as well. If the confidence values increase when using a larger dataset, then my association rules will be validated and will have more analytical importance. If the confidence or lift values decrease, then we can still use the Apriori algorithm to find stronger rules. This algorithm has proven to be extremely effective at mining useful information within this dataset. I expect that further implementation of this method can produce even stronger results in the future, allowing the mined information to save the lives of many burn victims.

References

Association Rule Mining with R - Week 3 Exercise. (2017, June 5). Retrieved June 5, 2017.

Code Sheet for Variables in the Burn Study. (2014, December 26). Retrieved June 6, 2017, from https://github.com/lbraglia/aplore3/blob/master/rawdata/BURN/BURN_Code_Sheet.pdf

Data Preprocessing in R - Week 2 Exercise. (2017, May 22). Retrieved May 31, 2017.

Durtschi, M. B., Kohler, T. R., Finley, A., & Heimbach, D. M. (1980). Burn injury in infants and young children [Abstract]. *Surg Gynecol Obstet, 150*(5), 651-656. Retrieved June 9, 2017, from https://www.ncbi.nlm.nih.gov/pubmed/7368048.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: concepts and techniques* (3rd ed.). Retrieved May 22, 2017.

Lafferty, K. A. (2017, May 24). Smoke Inhalation Injury. Retrieved June 09, 2017, from http://emedicine.medscape.com/article/771194-overview

Morrow, S. E., Smith, D. L., Cairns, B. A., Howell, P. D., Nakayama, D. K., & Peterson, H. (1996). Etiology and outcome of pediatric burns [Abstract]. *Journal of Pediatric Surgery, 31*(3), 329-333. doi:10.1016/s0022-3468(96)90732-0

Tan, P., Steinbach, M., & Kumar, V. (2015). *Introduction to data mining*. Dorling Kindersley: Pearson. Retrieved June 4, 2017.

Appendix A

R Script for Assignment 1

```
# DATA 630 Assignment 1

# Written by Daanish Ahmed

# Semester Summer 2017

# June 6, 2017

# Professor Edward Herranz


# This R script conducts an association rules analysis on a dataset containing

# statistical information on burn victims.  The purpose of this assignment is to

# generate a list of association rules and analyze the results to gain useful

# insights.  This script consists of several components, including opening and

# initializing the data, exploration and preprocessing, implementing the Apriori

# algorithm, pruning redundant rules, and the commands for analyzing the results.



# This section of code covers opening the dataset and initializing the packages

# that are used in this script.


# Sets the working directory for this assignment.  Please change this directory

# to whichever directory you are using, and make sure that all files are placed

# in that location.
```

```
setwd("~/Class Documents/2016-17 Summer/DATA 630/R/Assignment 1")

# In order to run the discretization commands, we need to use the arules and
# arulesViz packages.

# If you have not installed these packages yet, remove the two # symbols below.
# install.packages("arules")
# install.packages("arulesViz")

# Loads the arules and arulesViz packages into the system.
library("arules")
library("arulesViz")

# Opens the CSV file "burn_edit.csv".
burn <- read.csv(file="burn_edit.csv", head=TRUE, sep=",")

# End of opening the dataset.


# This section of code covers data preprocessing.  It includes exploration of
# the original dataset, removing unique identifiers, dealing with missing
# values, and discretization of numeric variables.
```

```
# Previews the burn dataset.
View(burn)

# Shows the descriptive statistics for all variables in the dataset.
summary(burn)

# Displays the structure of the burn data.  This is necessary to see if there
# are any unique identifiers (IDs) that can be removed.  Such variables are not
# useful for the analysis and should be removed.
str(burn)

# The first variable is an ID, and we remove it.
burn <- burn[, -1]

# Verifies that the ID variable has been removed.
str(burn)

# This function checks to see how many missing values are in each variable.
apply(burn, 2, function(credit) sum(is.na(credit)))

# Since there are no missing values in any of the variables, we do not need to
# replace any values.
```

```
# We need to discretize all of the variables that are not already factors.  This
# is required before running the Apriori rules method.
burn$FACILITY <- discretize(burn$FACILITY, "frequency", categories=6)
burn$AGE <- discretize(burn$AGE, "frequency", categories=6)
burn$TBSA <- discretize(burn$TBSA, "frequency", categories=6)

# Verifies that the facility, age, and TBSA variables have been successfully
# converted to factors.
summary(burn$FACILITY)
summary(burn$AGE)
summary(burn$TBSA)

# End of data preprocessing.


# This section of code covers the implementation of the Apriori algorithm.
# This method will create a list of decision rules using the data provided in
# the burn dataset.  Afterwards, redundant rules will be removed from the list.

# This generates a list of rules using the following parameters: 0.15 support,
# 0.85 confidence, and minimum length of 2.
rules <- apriori(burn, parameter= list(supp=0.15, conf=0.85, minlen=2))
```

```
# Creates a sorted list that sorts the rules by lift in descending order, and
# displays the top 10 rules in the list.
rules.sorted <- sort(rules, by="lift")
inspect(rules.sorted[1:10])

# This code looks for redundant rules in the current sorted list of rules and
# stores each redundant rule in an array.
subset.matrix <- is.subset(rules.sorted, rules.sorted)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- F
redundant <- colSums(subset.matrix, na.rm=T) >= 1

# Returns each redundant rule and its corresponding index.
which(redundant)

# This code uses the array of redundant rules and creates a new rules list
# that excludes the redundant rules.
rules.pruned <- rules.sorted[!redundant]

# Shows the descriptive statistics for the new list of rules, including the
# number of rules, length of rules, and statistical information for support,
# confidence, and lift.
summary(rules.pruned)
```

```
# Examines the top 10 rules in our pruned rules list.
inspect(rules.pruned[1:10])

# End of Apriori algorithm implementation.


# The next few sections of code include all commands that were used in the
# results analysis portion of the assignment.  These include visualizations
# and generating additional lists of rules for specific itemsets.

# Creates a scatterplot of the top 10 pruned rules with the support on the
# x-axis, confidence on the y-axis, and lift as color-coded dots.
plot(rules.pruned[1:10])

# End of analyzing initial pruned rules.  The next sections discuss additional
# rules for specific itemsets.


# This code finds the association rules that result in victim's deaths.  The
# original list did not have any rules with the condition {DEATH=dead} on the
# RHS, thus I created a new rules list using the same procedure as before.

# This command creates a new rules list containing {DEATH=dead} on the RHS.
```

```
# Since dead victims rarely occur in the dataset, it is necessary to set a
# lower minimum support threshold.  Likewise, I set a lower confidence value.
rules_dead <- apriori(burn, parameter= list(supp=0.05, conf=0.5, minlen=2),
                      appearance=list(rhs=c("DEATH=dead"), default="lhs"))

# Sorts the list by lift in descending order.
rules_dead <- sort(rules_dead, by="lift")

# Creates an array to store all redundant rules.
subset.matrix <- is.subset(rules_dead, rules_dead)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- F
redundant <- colSums(subset.matrix, na.rm=T) >= 1

# Creates a new rules list that excludes the redundant rules.
rules_dead.p <- rules_dead[!redundant]

# Displays all rules in the new rules list.
inspect(rules_dead.p)

# Creates a grouped plot for this rules list.  The size of each circle
# represents the support value, while the color indicates the lift value.
plot(rules_dead.p, method = "grouped")
```

```
# End of analyzing victim death cases.



# This code finds the association rules which involve inhalation injuries.
# This process once again involves creating a new list, but this time
# {INH_INJ=yes} is on the LHS to show what inhalation injuries cause.

# This command creates a new rules list containing {INH_INJ=yes} on the LHS.
# Due to the low frequency of inhalation injuries, it is necessary to set
# lower thresholds for support and confidence.
rules_inh <- apriori(burn, parameter= list(supp=0.05, conf=0.5, minlen=2),
                     appearance=list(lhs=c("INH_INJ=yes"), default="rhs"))

# Sorts the list by lift in descending order.
rules_inh <- sort(rules_inh, by="lift")

# Creates an array to store all redundant rules.
subset.matrix <- is.subset(rules_inh, rules_inh)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- F
redundant <- colSums(subset.matrix, na.rm=T) >= 1

# Creates a new rules list that excludes the redundant rules.
rules_inh.p <- rules_inh[!redundant]
```

```
# Displays all rules in the new rules list.
inspect(rules_inh.p)

# Creates a circle graph for this rules list.  The size of each circle
# represents the support value, while the color indicates the lift value.
plot(rules_inh.p, method="graph", control=list(type="items"))

# End of analyzing inhalation injuries.


# End of results analysis.

# End of script.
```

Appendix B

Relevant R Output Images

```
> str(burn)
'data.frame':   1000 obs. of  9 variables:
 $ ï..ID   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ FACILITY: int  11 1 12 1 1 6 22 1 1 1 ...
 $ DEATH   : Factor w/ 2 levels "alive","dead": 1 1 1 1 1 1 1 1 1 1 ...
 $ AGE     : num  26.6 2 22 37.3 52.1 50.2 2.5 53.8 31.9 41.1 ...
 $ GENDER  : Factor w/ 2 levels "female","male": 2 1 1 2 2 2 1 1 2 2 ...
 $ RACEC   : Factor w/ 2 levels "non-white","white": 2 1 1 2 2 2 1 2 2 2 ...
 $ TBSA    : num  25.3 5 2 2 6 7 7 0.9 2 22 ...
 $ INH_INJ : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ FLAME   : Factor w/ 2 levels "no","yes": 2 1 1 1 2 1 1 2 1 2 ...
>
```

Figure 1. Initial Data Structure of Burn Dataset.

```
> summary(burn)
     ï..ID          FACILITY       DEATH          AGE            GENDER          RACEC          TBSA          INH_INJ
 Min.   :   1.0   Min.   : 1.00   alive:850   Min.   : 0.10   female:295   non-white:411   Min.   : 0.10   no :878
 1st Qu.: 250.8   1st Qu.: 2.00   dead :150   1st Qu.:10.85   male  :705   white    :589   1st Qu.: 2.50   yes:122
 Median : 500.5   Median : 8.00               Median :31.95                                Median : 6.00
 Mean   : 500.5   Mean   :11.56               Mean   :33.29                                Mean   :13.54
 3rd Qu.: 750.2   3rd Qu.:18.25               3rd Qu.:51.23                                3rd Qu.:16.00
 Max.   :1000.0   Max.   :40.00               Max.   :89.70                                Max.   :98.00
 FLAME
 no :471
 yes:529
```

Figure 2. Descriptive Statistics of Initial Variables in Burn Dataset.

```
> apply(burn, 2, function(credit) sum(is.na(credit)))
FACILITY    DEATH      AGE   GENDER    RACEC     TBSA  INH_INJ    FLAME
       0        0        0        0        0        0        0        0
```

Figure 3. Number of Missing Values for all Variables.

```
> summary(burn$FACILITY)
      1 [ 2, 5) [ 5, 9) [ 9,15) [15,25) [25,40]
    214     149     149     159     174     155
> summary(burn$AGE)
[ 0.1, 3.0) [ 3.0,17.9) [17.9,32.0) [32.0,46.1) [46.1,59.7) [59.7,89.7]
        167         167         166         169         165         166
> summary(burn$TBSA)
[ 0.1, 2.1) [ 2.1, 3.8) [ 3.8, 6.5) [ 6.5,11.2) [11.2,22.7) [22.7,98.0]
        239         107         175         148         165         166
> |
```

Figure 4. Discretized Variables Based on Equal Frequency.

```
> rules <- apriori(burn, parameter= list(supp=0.15, conf=0.85, minlen=2))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target   ext
       0.85    0.1    1 none FALSE            TRUE       5    0.15      2     10  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 150

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[28 item(s), 1000 transaction(s)] done [0.00s].
sorting and recoding items ... [23 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 done [0.00s].
writing ... [86 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
>
```

Figure 5. Apriori Algorithm Parameters and Generation.

```
> rules.sorted <- sort(rules, by="lift")
> inspect(rules.sorted[1:10])
     lhs                                         rhs            support confidence lift
[1]  {DEATH=alive,AGE=[ 0.1, 3.0),INH_INJ=no} => {FLAME=no}     0.152   0.9440994  2.004457
[2]  {AGE=[ 0.1, 3.0),INH_INJ=no}             => {FLAME=no}     0.152   0.9382716  1.992084
[3]  {DEATH=alive,AGE=[ 0.1, 3.0)}            => {FLAME=no}     0.153   0.9329268  1.980736
[4]  {AGE=[ 0.1, 3.0)}                        => {FLAME=no}     0.153   0.9161677  1.945154
[5]  {AGE=[ 0.1, 3.0),FLAME=no}               => {DEATH=alive} 0.153   1.0000000  1.176471
[6]  {AGE=[ 3.0,17.9),INH_INJ=no}             => {DEATH=alive} 0.153   1.0000000  1.176471
[7]  {AGE=[ 0.1, 3.0),INH_INJ=no,FLAME=no}    => {DEATH=alive} 0.152   1.0000000  1.176471
[8]  {AGE=[ 0.1, 3.0),INH_INJ=no}             => {DEATH=alive} 0.161   0.9938272  1.169208
[9]  {TBSA=[ 0.1, 2.1),INH_INJ=no}            => {DEATH=alive} 0.229   0.9828326  1.156274
[10] {AGE=[ 0.1, 3.0)}                        => {DEATH=alive} 0.164   0.9820359  1.155336
>
```

Figure 6. Initial List of Association Rules, Sorted by Lift.

```
> rules.pruned <- rules.sorted[!redundant]
> summary(rules.pruned)
set of 44 rules

rule length distribution (lhs + rhs):sizes
 2  3  4
22 19  3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.500   2.568   3.000   4.000

summary of quality measures:
    support         confidence          lift
 Min.   :0.1510   Min.   :0.8567   Min.   :0.9884
 1st Qu.:0.1600   1st Qu.:0.9162   1st Qu.:1.0486
 Median :0.2080   Median :0.9486   Median :1.1179
 Mean   :0.2646   Mean   :0.9409   Mean   :1.1721
 3rd Qu.:0.2797   3rd Qu.:0.9751   3rd Qu.:1.1412
 Max.   :0.8000   Max.   :1.0000   Max.   :2.0045

mining info:
 data ntransactions support confidence
 burn          1000    0.15       0.85
>
```

Figure 7. Descriptive Statistics of Pruned Rules List.

```
> inspect(rules.pruned[1:10])
     lhs                                                rhs           support confidence lift
[1]  {DEATH=alive,AGE=[ 0.1, 3.0),INH_INJ=no} => {FLAME=no}    0.152   0.9440994  2.004457
[2]  {AGE=[ 0.1, 3.0),INH_INJ=no}             => {FLAME=no}    0.152   0.9382716  1.992084
[3]  {DEATH=alive,AGE=[ 0.1, 3.0)}            => {FLAME=no}    0.153   0.9329268  1.980736
[4]  {AGE=[ 0.1, 3.0)}                        => {FLAME=no}    0.153   0.9161677  1.945154
[5]  {AGE=[ 3.0,17.9),INH_INJ=no}             => {DEATH=alive} 0.153   1.0000000  1.176471
[6]  {AGE=[ 0.1, 3.0),INH_INJ=no}             => {DEATH=alive} 0.161   0.9938272  1.169208
[7]  {TBSA=[ 0.1, 2.1),INH_INJ=no}            => {DEATH=alive} 0.229   0.9828326  1.156274
[8]  {AGE=[ 0.1, 3.0)}                        => {DEATH=alive} 0.164   0.9820359  1.155336
[9]  {TBSA=[ 0.1, 2.1),FLAME=no}              => {DEATH=alive} 0.164   0.9820359  1.155336
[10] {TBSA=[ 0.1, 2.1)}                       => {DEATH=alive} 0.234   0.9790795  1.151858
>
```

Figure 8. Pruned List of Association Rules, Sorted by Lift.

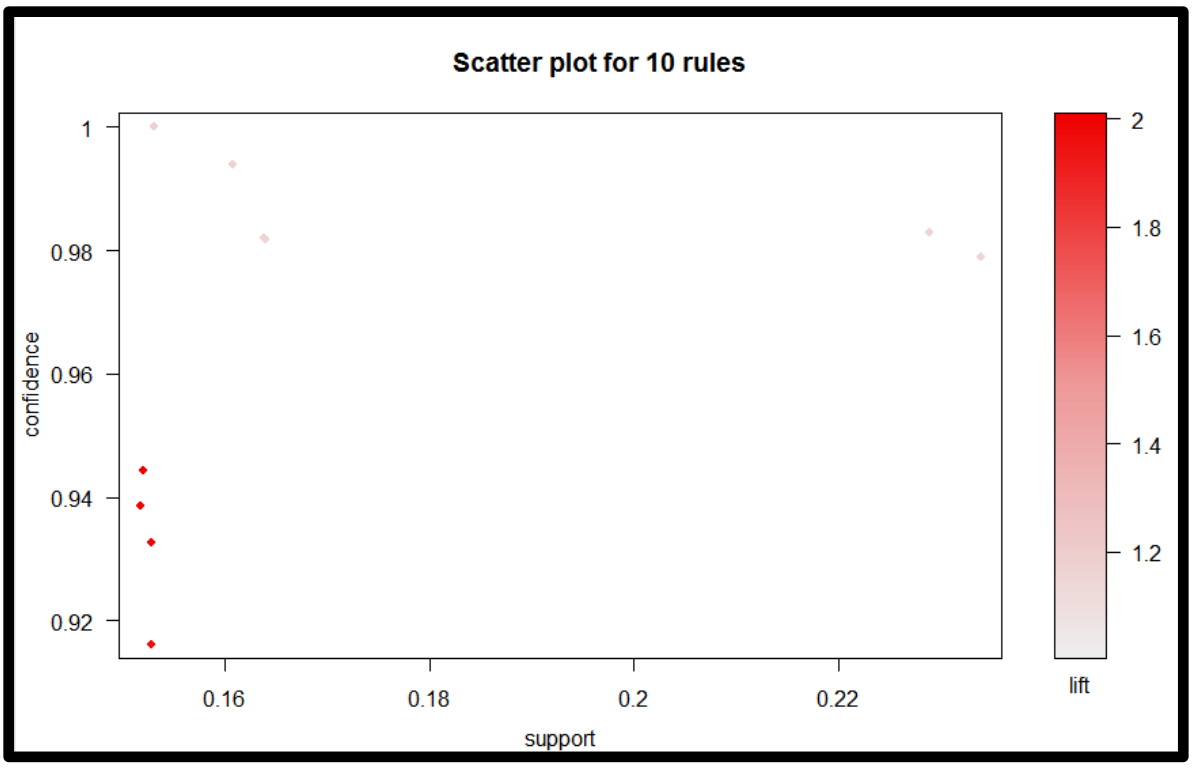

Figure 9. Scatterplot for Top 10 Association Rules.

```
> inspect(rules_dead.p)
    lhs                                                rhs          support confidence lift
[1] {TBSA=[22.7,98.0],INH_INJ=yes,FLAME=yes} => {DEATH=dead} 0.057   0.7402597  4.935065
[2] {TBSA=[22.7,98.0],INH_INJ=yes}           => {DEATH=dead} 0.058   0.7250000  4.833333
[3] {TBSA=[22.7,98.0],FLAME=yes}             => {DEATH=dead} 0.092   0.6524823  4.349882
[4] {INH_INJ=yes,FLAME=yes}                  => {DEATH=dead} 0.070   0.6034483  4.022989
[5] {RACEC=white,TBSA=[22.7,98.0]}           => {DEATH=dead} 0.061   0.5980392  3.986928
[6] {TBSA=[22.7,98.0]}                       => {DEATH=dead} 0.098   0.5903614  3.935743
[7] {INH_INJ=yes}                            => {DEATH=dead} 0.072   0.5901639  3.934426
[8] {AGE=[59.7,89.7],FLAME=yes}              => {DEATH=dead} 0.063   0.5250000  3.500000
>
```

Figure 10. Pruned List of Rules Resulting in Death, Sorted by Lift.

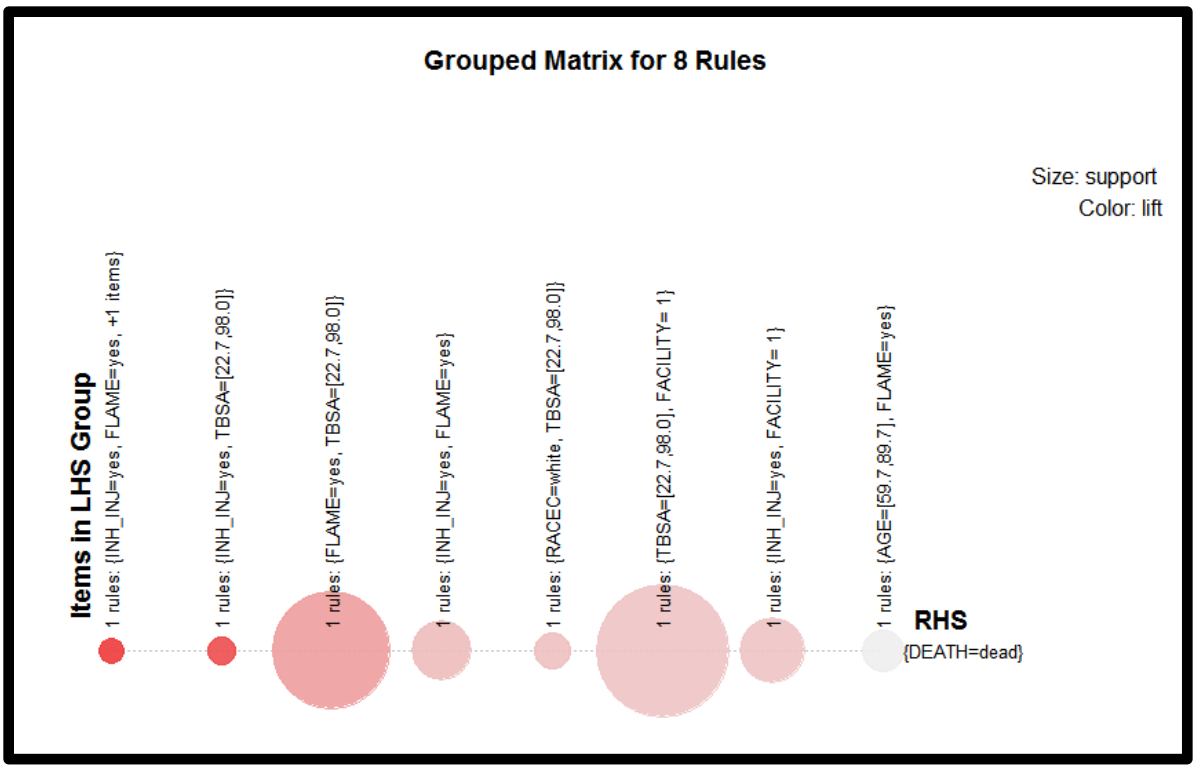

Figure 11. Grouped Plot for Rules Resulting in Death.

```
> inspect(rules_inh.p)
    lhs                 rhs                    support confidence lift
[1] {INH_INJ=yes} => {TBSA=[22.7,98.0]} 0.080   0.6557377  3.950227
[2] {INH_INJ=yes} => {DEATH=dead}       0.072   0.5901639  3.934426
[3] {INH_INJ=yes} => {FLAME=yes}        0.116   0.9508197  1.797391
[4] {INH_INJ=yes} => {RACEC=white}      0.075   0.6147541  1.043725
[5] {INH_INJ=yes} => {GENDER=male}      0.087   0.7131148  1.011510
>
```

Figure 12. Pruned List of Rules Involving Inhalation Injuries, Sorted by Lift.

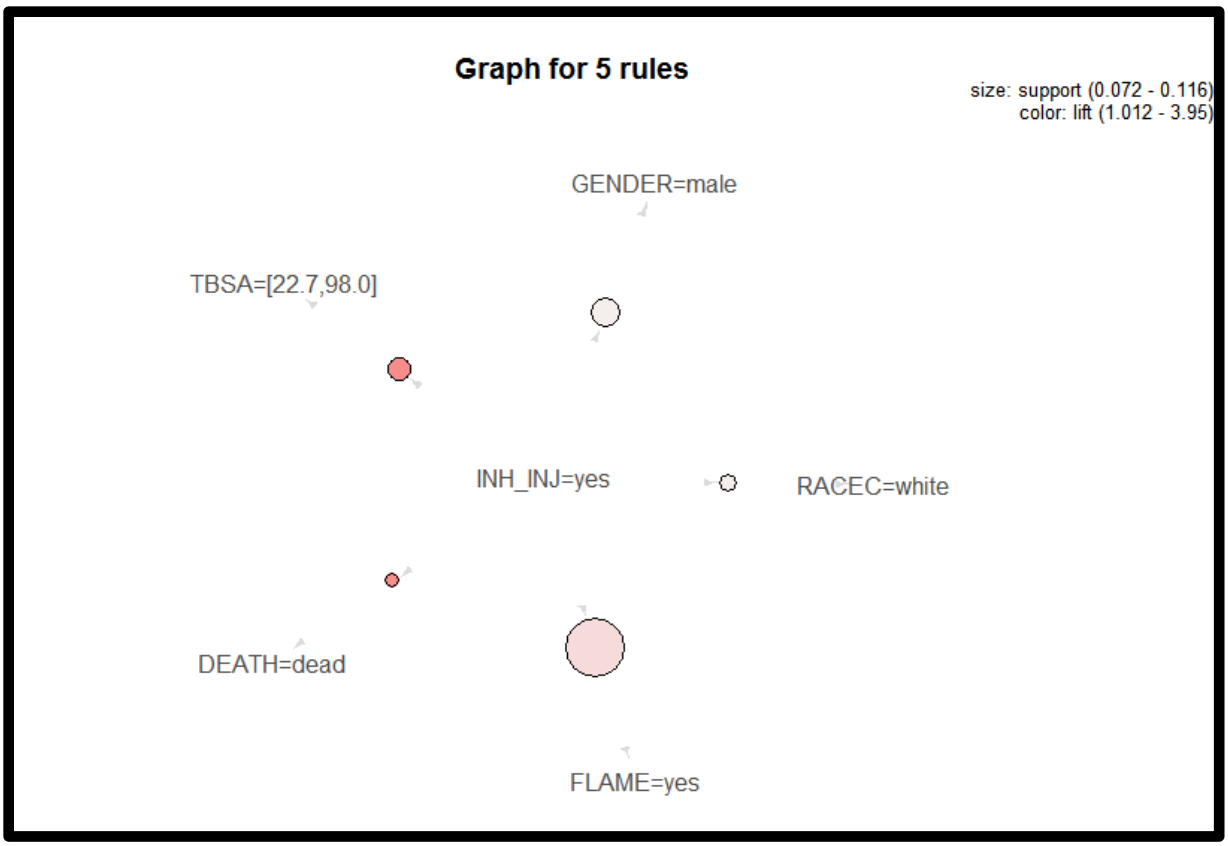

Figure 13. Circle Graph for Rules Involving Inhalation Injuries.