AsssignmentReport

Introduction

Since the significance of the knowledge about the edibility of mushroom, we believe it's necessary to extract meaningful infomation from the mushrooom dataset.This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family.Each species is identified as edible or poisonous.This report mainly focus on the analysis of Mushroom dataset, including some static analysis in data exploration part, and association rules mining in data analysis part.

Data Exploration

Load the dataset at first.

```
##load Mushroom dataset
url <- "./data/mushroom/agaricus-lepiota.data"
mushrooms <- read.csv(file = url, header = FALSE)
#assign column names
names(mushrooms) <- c("class", "cap-shape", "cap-surface",
                       "cap-color", "bruises", "odor", "gill-attachment", "gill-spacing",
                       "gill-size", "gill-color", "stalk-shape", "stalk-root",
                       "stalk-surface-above-ring", "stalk-surface-below-ring",
                       "stalk-color-above-ring", "stalk-color-below-ring",
                       "veil-type", "veil-color", "ring-number", "ring-type",
                       "spore-print-color", "population", "habitat")
```

Then check out the dimensionality of the dataset.There is 8124 samples and 23 features.

```
dim(mushrooms)
```

```
## [1] 8124   23
```

And its column names.

```
names(mushrooms)
```

```
##  [1] "class"                    "cap-shape"
##  [3] "cap-surface"              "cap-color"
##  [5] "bruises"                  "odor"
##  [7] "gill-attachment"          "gill-spacing"
##  [9] "gill-size"                "gill-color"
## [11] "stalk-shape"              "stalk-root"
## [13] "stalk-surface-above-ring" "stalk-surface-below-ring"
## [15] "stalk-color-above-ring"   "stalk-color-below-ring"
## [17] "veil-type"                "veil-color"
## [19] "ring-number"              "ring-type"
## [21] "spore-print-color"        "population"
## [23] "habitat"
```

We can further explore these colunms.The first column indicates the edibility of the species. The remaining column is about various features of the species which covers different part of the mushroom from the cap to the root, and also describe the characteristics of the mushroom from different sensory levels, including vision, smell, touch, etc.Note that all the columns are factors.

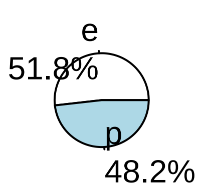
```
str(mushrooms)
```

```
## 'data.frame':    8124 obs. of  23 variables:
```

```
## $ class                    : Factor w/ 2 levels "e","p": 2 1 1 2 1 1 1 1 2 1 ...
## $ cap-shape                : Factor w/ 6 levels "b","c","f","k",..: 6 6 1 6 6 6 1 1 6 1 ...
## $ cap-surface              : Factor w/ 4 levels "f","g","s","y": 3 3 3 4 3 4 3 4 4 3 ...
## $ cap-color                : Factor w/ 10 levels "b","c","e","g",..: 5 10 9 9 4 10 9 9 9 10 ...
## $ bruises                  : Factor w/ 2 levels "f","t": 2 2 2 2 1 2 2 2 2 2 ...
## $ odor                     : Factor w/ 9 levels "a","c","f","l",..: 7 1 4 7 6 1 1 4 7 1 ...
## $ gill-attachment          : Factor w/ 2 levels "a","f": 2 2 2 2 2 2 2 2 2 2 ...
## $ gill-spacing             : Factor w/ 2 levels "c","w": 1 1 1 1 2 1 1 1 1 1 ...
## $ gill-size                : Factor w/ 2 levels "b","n": 2 1 1 2 1 1 1 1 2 1 ...
## $ gill-color               : Factor w/ 12 levels "b","e","g","h",..: 5 5 6 6 5 6 3 6 8 3 ...
## $ stalk-shape              : Factor w/ 2 levels "e","t": 1 1 1 1 2 1 1 1 1 1 ...
## $ stalk-root               : Factor w/ 5 levels "?","b","c","e",..: 4 3 3 4 4 3 3 3 4 3 ...
## $ stalk-surface-above-ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
## $ stalk-surface-below-ring: Factor w/ 4 levels "f","k","s","y": 3 3 3 3 3 3 3 3 3 3 ...
## $ stalk-color-above-ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...
## $ stalk-color-below-ring  : Factor w/ 9 levels "b","c","e","g",..: 8 8 8 8 8 8 8 8 8 8 ...
## $ veil-type                : Factor w/ 1 level "p": 1 1 1 1 1 1 1 1 1 1 ...
## $ veil-color               : Factor w/ 4 levels "n","o","w","y": 3 3 3 3 3 3 3 3 3 3 ...
## $ ring-number              : Factor w/ 3 levels "n","o","t": 2 2 2 2 2 2 2 2 2 2 ...
## $ ring-type                : Factor w/ 5 levels "e","f","l","n",..: 5 5 5 5 1 5 5 5 5 5 ...
## $ spore-print-color        : Factor w/ 9 levels "b","h","k","n",..: 3 4 4 3 4 3 3 4 3 3 ...
## $ population               : Factor w/ 6 levels "a","c","n","s",..: 4 3 3 4 1 3 3 4 5 4 ...
## $ habitat                  : Factor w/ 7 levels "d","g","l","m",..: 6 2 4 6 2 2 4 4 2 4 ...
```

In order to know the split of between edible mushrooms and poisonous ones in the dataset, we can plot a pie chart, from which we can see the dataset is quite balanced.

```r
library(dplyr)
tab <- mushrooms$class %>% table()
precentages <- tab %>% prop.table() %>% round(3) * 100
txt <- paste0(names(tab), '\n', precentages, '%')
pie(tab, labels=txt)
```



Data Analysis:Association Rules Mining

Now we use Apriori to mine association rules.Note that we limit the length to [2,5],and confidence=1 to filter out over complex rules and less reliable ones.And we can print the number of rules we get.

```r
library(arules)
library(arulesViz)
```

```r
rules <- apriori(mushrooms, control = list(verbose=F),
                 parameter = list(minlen=2, maxlen=5, confidence=1),
                 appearance = list(rhs=c("class=p", "class=e"),
                                   default="lhs"))
quality(rules) <- round(quality(rules), digits=3)
paste('The numer of rules we got is',nrow(quality(rules)))
```

```
## [1] "The numer of rules we got is 6484"
```

Before analysis, prune the redundant rules.

```r
#prune redundant rules
subset.matrix <- is.subset(rules, rules)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- F
redundant <- colSums(subset.matrix) >= 1
rules.pruned <- rules[!redundant]
#check out how many rules left after pruning
paste("The number of rules left after pruning is",nrow(quality(rules.pruned)))
```

```
## [1] "The number of rules left after pruning is 226"
```

Now we can inspect the first 8 rules after sorting the rules by lift and support.

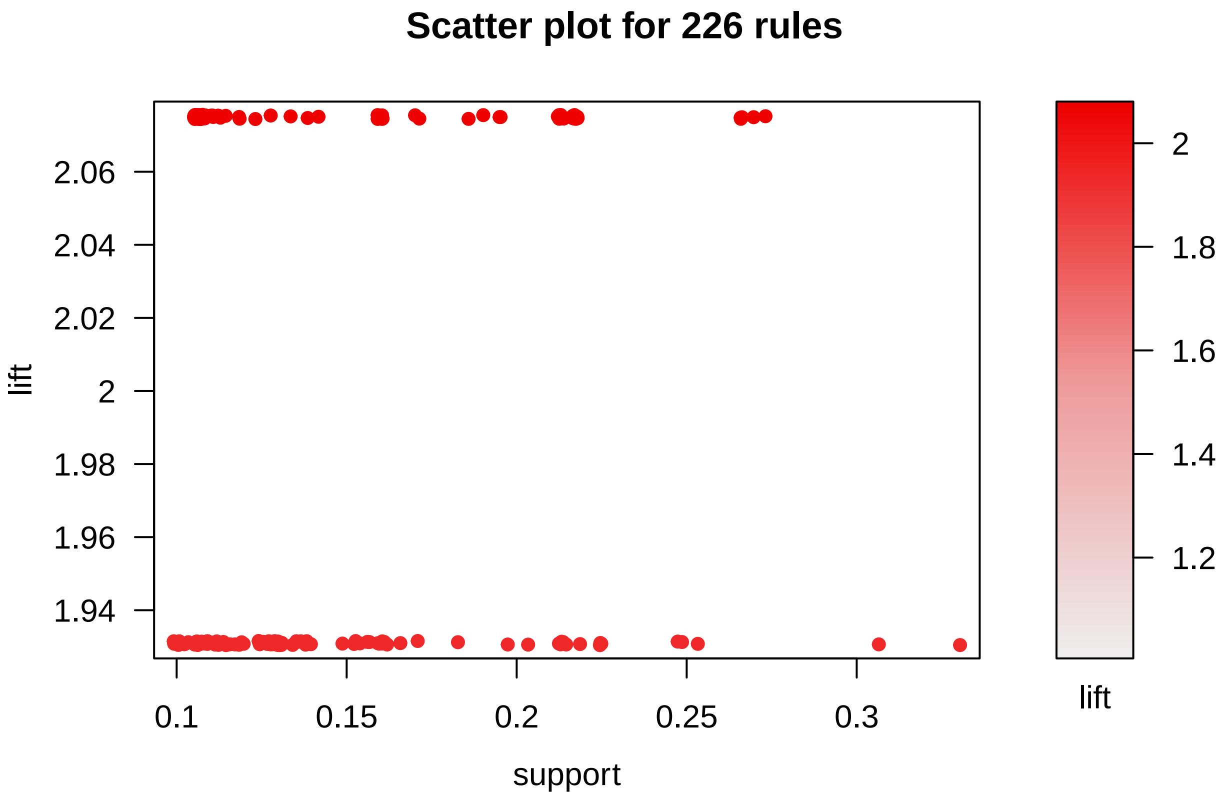
```r
#Sort rules by lift and support and inspect the first 8 ones.
rules.pruned.sorted <- sort(rules.pruned, by=c("lift","support"))
inspect(head(rules.pruned.sorted, 8))
```

```
##     lhs                              rhs       support confidence  lift count
## [1] {gill-spacing=c,
##      stalk-surface-above-ring=k} => {class=p}   0.274          1 2.075  2228
## [2] {stalk-surface-above-ring=k,
##      ring-number=o}              => {class=p}   0.270          1 2.075  2192
## [3] {odor=f}                     => {class=p}   0.266          1 2.075  2160
## [4] {gill-spacing=c,
##      stalk-surface-below-ring=k} => {class=p}   0.266          1 2.075  2160
## [5] {stalk-surface-below-ring=k,
##      ring-number=o}              => {class=p}   0.266          1 2.075  2160
## [6] {gill-size=n,
##      stalk-root=?,
##      spore-print-color=w}        => {class=p}   0.217          1 2.075  1760
## [7] {stalk-root=?,
##      spore-print-color=w,
##      population=v}               => {class=p}   0.217          1 2.075  1760
## [8] {stalk-root=?,
##      ring-number=o,
##      spore-print-color=w}        => {class=p}   0.217          1 2.075  1760
```

As we can see,all the rules displayed has lift=2.075, which means mushrooms having these characteristics in the lhs will have 2 times more probability to become poisonous. And most of the rules has the length of 3 or 4, but the 2nd important rule has the length of 2, so we may speculate that the feature "odor" is rather significant to determine the edibility.

And by using a scatter plot we can visualize the distribution of the rules.

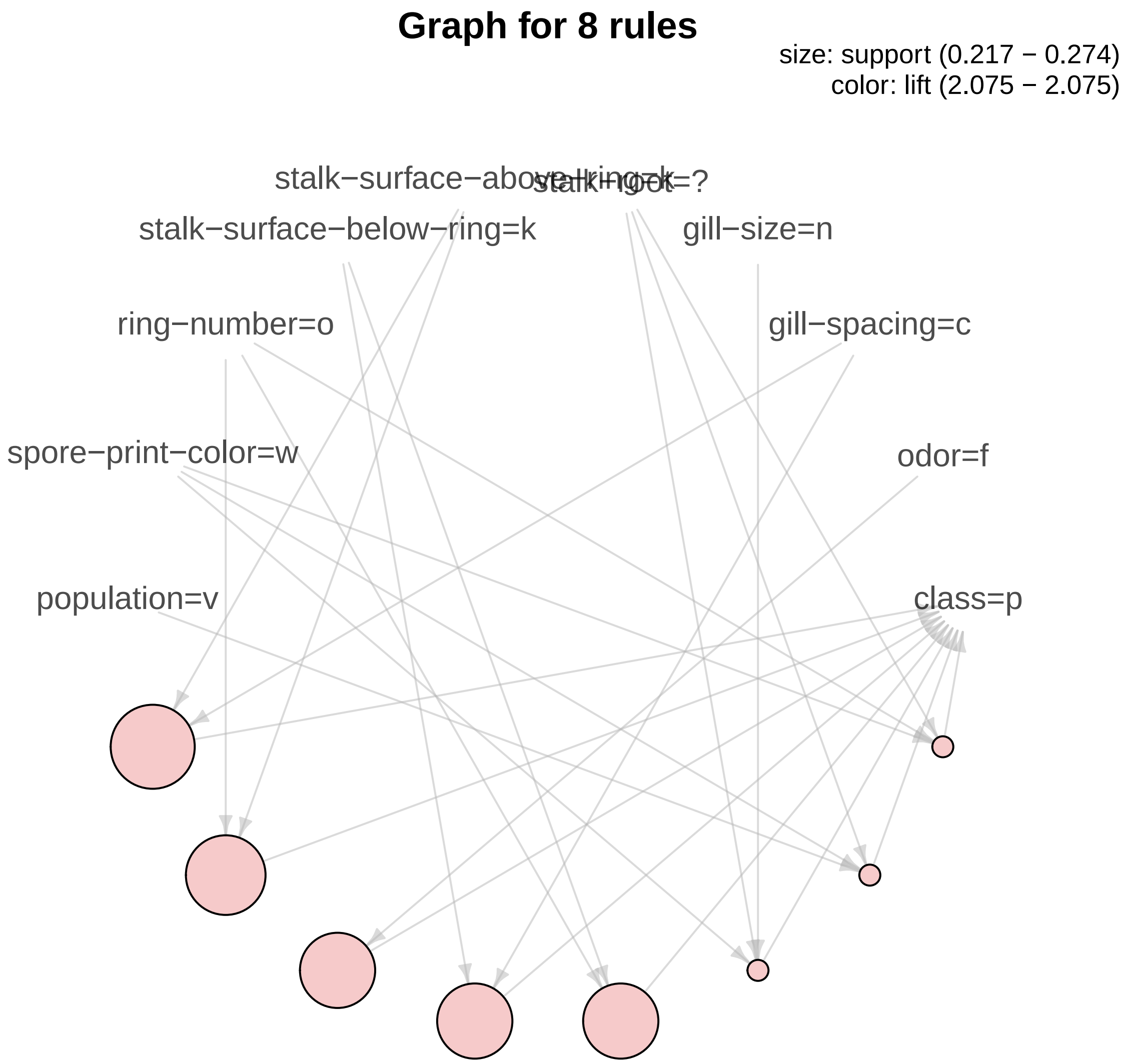
```r
plot(rules.pruned.sorted, measure=c("support","lift"))
```



Now we find an interesting phenomenon that basically all the rules has the lift of around 2.06 or 1.94, whereas the support of these rules varies.

In order to further analyse the association between these 8 rules, we can visualize them by ploting a graph.

```
#plot the graph of the rules to the association between them
head(rules.pruned.sorted,n=8) %>% plot(method="graph",  control=list(layout=igraph::in_circle()))
```



By obserbing the number of out edges of each feature node, we know that spore-print-color=w, ring-number=0 are most frequently shown in lhs, which indicates these feature may be highly related to poisonous mushrooms. And the size of the pink nodes represents the lift of a certain rule.So by taking that into consideration, stalk-surface-above-ring and gill-spacing may be more reliable indicators for edibility.

At last, we can study how well a single feature indicates the edibility by mining association rules of length=2.This time We don't limit the confident to be 1, otherwise there wouldn't be many rules left.

```
#limit the rules' length to 2 and mine again
rules2 <- apriori(mushrooms, control = list(verbose=F),
                  parameter = list(minlen=2, maxlen=2),
                  appearance = list(rhs=c("class=p", "class=e"),
                                    default="lhs"))
quality(rules2) <- round(quality(rules2), digits=3)
```

Again we sort the rules by confidence, lift and support and print the top 8 rules.

```
#sort by confidence, lift and support
rules2.sorted <- sort(rules2, by=c("confidence","lift","support"))
```

```r
inspect(head(rules2.sorted, 8))
```

```
##     lhs                              rhs         support confidence lift 
## [1] {odor=f}                      => {class=p} 0.266   1.000      2.075
## [2] {gill-color=b}                => {class=p} 0.213   1.000      2.075
## [3] {ring-type=l}                 => {class=p} 0.160   1.000      2.075
## [4] {spore-print-color=h}         => {class=p} 0.195   0.971      2.014
## [5] {odor=n}                      => {class=e} 0.419   0.966      1.865
## [6] {stalk-surface-above-ring=k} => {class=p} 0.274   0.939      1.949
## [7] {stalk-surface-below-ring=k} => {class=p} 0.266   0.938      1.945
## [8] {gill-spacing=w}              => {class=e} 0.148   0.915      1.766
##     count
## [1] 2160 
## [2] 1728 
## [3] 1296 
## [4] 1584 
## [5] 3408 
## [6] 2228 
## [7] 2160 
## [8] 1200
```

The result validates the importance of feature odor, and gill-color=b and ringt-type=l are also strong indicators. As for gill-spacing and stalk-surface-above-ring, they are pretty well in the first-round analysis, but the result above shows that they are less reliable as an individual feature, they may work better when combined with other features.

Conclusion

To determine whether a mushroom is edible is fairly complicated, so we try to avoid concluding a simple rule, since there is no such thing as a silver bullet.Instead we try to explore which feature may be more useful than others such as ordor, and understand the interaction of different features using visualization such as rules graph, under controled complexity.

Since we only use Apriori to mine association rules, the sight may be very limited.For further exploration, other methods can be exploited.

Also some classification methods can be applied to this dataset.Building a decision tree may bring more interesting and interpretable result, and it's also promising to use random forest to calculate the importance of each features.