Association Rule Analysis for Syngenta Soybean Data

Association Rule Mining for Data Understanding

Ian Johnson

Southern Methodist University

November 9, 2016

Contents

Executive Summary

Lorem Ipsum... TODO WRITE EXECUTIVE SUMMARY.

1 Data Preparation

1.1 Data Introduction

This report will analyze transaction data generated from the Syngenta Informs OR Competition dataset [1]. This data contains information about soybean seed tests which are used to decide which seed varieties will go to market and be sold by Syngenta. Potential seeds are tested in a 3-tiered testing system. First, all potential seeds are tested in 10 locations across the US. The top 15% of those seeds continue to the second and third tiers, at which all seeds are tested in 30 locations across the US, and the top 15% are selected and moved onto the next round. Seeds which make it through all 3 tiers of testing and outperform existing seeds become new seeds sold by Syngenta. Such seeds have sales data in the OR competition dataset.

The raw Informs OR dataset is table data where each row represents an individual test for a given seed. The columns are:

- Experiment Number (*nominal*) - A unique identifier for experiment represented by a given row
- Seed Variety (*nominal*) - A unique identifier for the variety of a seed
- Seed Family (*nominal*) - A unique identifier for the family of a seed (there are many varieties in each family)
- Location (*nominal*) - A 4-digit code for the location where the test occurred
- Check (*nominal*) - A binary attribute which is set to 1 (*true*) if the seed being tested is already at market and being used for comparison to potential new seeds
- RM (Relative Maturity) (*interval*) - A floating point number between 2 and 5 which represents the rate at which the variety of seed matures
- Class Of (*interval*) - The year that the seed "graduated" to market, or "." if the seed did not graduate
- Grad (*nominal*) - A binary attribute which is set to 1 (*true*) if the seed being tested graduated after the given test
- Bags Sold (*ratio*) - The number of bags of seeds sold in the first year after this seed went to market or "." if the seed ddn't go to market
- Yield (*ratio*) - The number of bushels of soybeans produced per acre by the seed during this test
- Replication Number (*nominal*) - An integer code used to deliniate two experiments which use the same seed and same location and same year
- Year (*interval*) - The year when the experiment occured

Before transaction data is generated from this raw dataset, it will be manipulated and formatted such that it becomes easiest to work with.

1.2 Data Formatting and Encoding

The first simple modification made to the dataset is that rows with matching experiment numbers and separate replication numbers are averaged (so that every single row in the data represents a unique year-variety-location combination). This simplifies further analysis, and also makes each row more representative of the seed variety in question.

Subsequently, the dataset is ordered by variety, year, and location. This is a simple housekeeping decision to make forthcoming analyses easier.

1.2.1 Location Clustering

There are 152 unique locations used for testing in the dataset. In order to reduce the number of possible values of the location factor, clustering is used to reduce the location to a location cluster.

K-means clustering will be used, with the implementation from CRAN package "flexclust" [2].

An in-depth look at optimal clustering strategies may be warranted. However, because this report is focused on association rule mining, K-means will be used as a novel first-attempt at clustering.

Location clustering was performed by aggregating the dataset by location and computing the mean of yield, relative maturity, and bags sold of the seeds tested at that location. The resulting dataset was clustered (without the location variable included), and the resulting cluster for each location was assigned to that location.

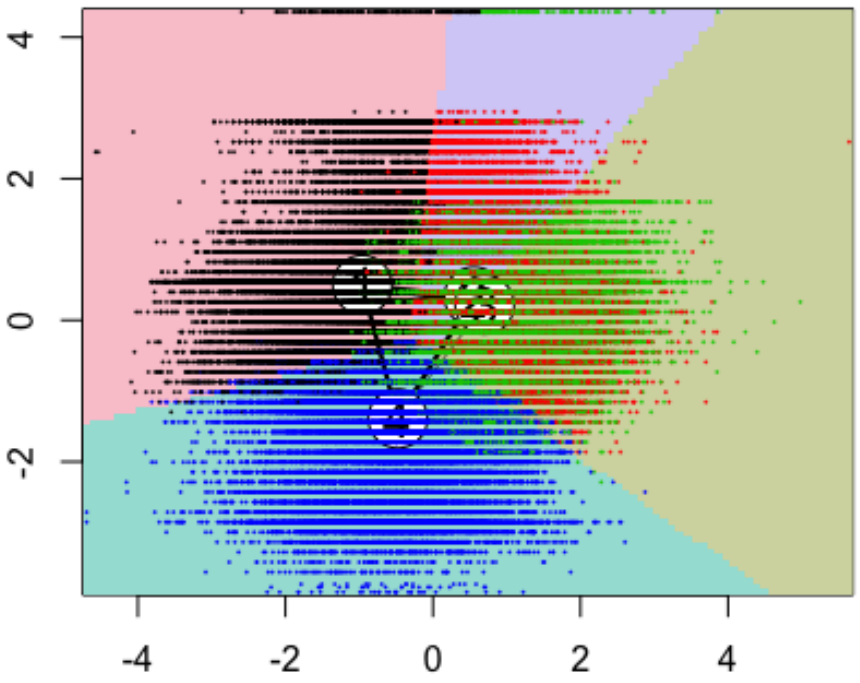

Figure 1.1 - A visualization of the K-means clustering algorithm output, with K=4

Figure 1.1 shows the result of clustering the data using K-means clustering with K=4. The result is visualized in 2-dimensions using the 2 top components from principle component analysis which is performed on the aggregated location data. PCA is performed using CRAN package "caret" [3].

Note that in this 2D space, the locations appear to be dispersed in a single large cluster, so the resulting clusters are not particularly meaningful. This will be considered when the association rules are analyzed.

After locations are clustered, a new attribute is added to the data table called locationGroup, which is a nominal integer in the range 1-4 which represents which cluster the location resides in.

1.2.2 A Divergence Between this Report and OR Competition Work

For the OR Competition work, I one-hot encode the year column, separate the dataframe into a set of dataframes by year, and then use columnar-binding to combine the data frames into one master data frame where each row includes data for each seed for each year, instead of including duplicate rows for any given seed variety. This format is used to facilitate classification or regression for bags sold information. However, for this report, the dataset is left as one large table with duplicate rows for each seed variety (no two rows are truly "duplicates" but there are multiple rows per seed). This is done in the hopes that it will allow for more insight from the association rule part of the analysis.

The dataset which will be used moving forward has the following columns:

- Seed Variety (*nominal*) - A unique identifier for the variety of a seed
- Seed Family (*nominal*) - A unique identifier for the family of a seed (there are many varieties in each family)
- Location (*nominal*) - A 4-digit code for the location where the test occurred
- Check (*nominal*) - A binary attribute which is set to 1 (*true*) if the seed being tested is already at market and being used for comparison to potential new seeds
- RM (Relative Maturity) (*interval*) - A floating point number between 2 and 5 which represents the rate at which the variety of seed matures
- Grad (*nominal*) - A binary attribute which is set to 1 (*true*) if the seed being tested graduated after the given test
- Bags Sold (*ratio*) - The number of bags of seeds sold in the first year after this seed went to market or "." if the seed ddn't go to market
- Yield (*ratio*) - The number of bushels of soybeans produced per acre by the seed during this test
- Year (*interval*) - The year when the experiment occured
- Location Group (*nominal*) - The cluster number for the location of this experiment

Note that a few columns have been removed, including experiment number and replication number. The experiment number has no real meaning, and the replication number has been aggregated away.

1.3 Data Discretization

1.3.1 Yield Discretization

Based on the Syngenta problem statement, that at each test year, only the top 15% of seeds of move on to the subsequent year, it is meaningful to discretize yield results into 7 sections by equal frequency, such that approximately 15% of the seeds fall into each section after discretization.

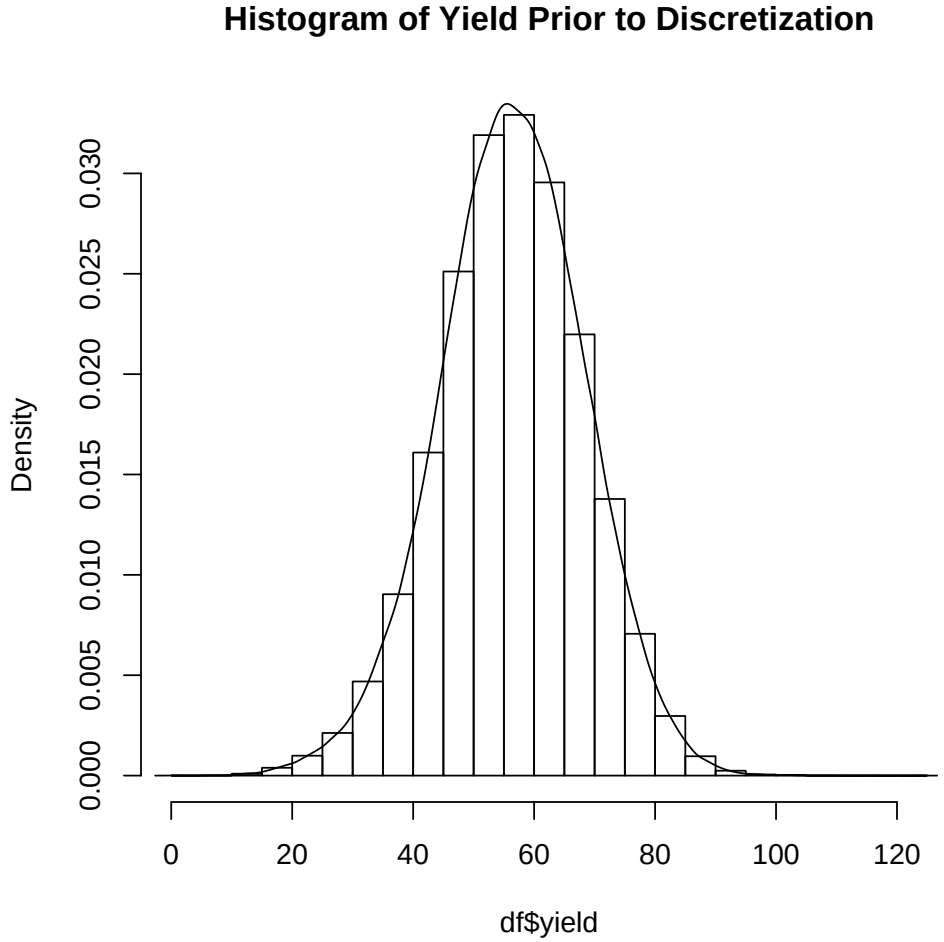

Figure 1.2 - Histogram of yield prior to discretization

Figure 1.2 shows that the yield of a soybean experiment follows a normal distribution, so equal-frequency discretization will yield much smaller ranges in the middle section of the data than on the outside sections. This is okay, however, as we're most interested in identifying which seeds are the highest and lowest performing without excluding too many seeds from the top tiers by using equal-width discretization.

```
[ 0.0, 45.0) [45.0, 51.4) [51.4, 56.4) [56.4, 61.6) [61.6, 67.9) [67.9,124.0]
       43043        43044        43040        43042        43042        43042
```

Figure 1.3 - Number of rows in each discretized yield category

Figure 1.3 shows that, after discretization, the yield attribute in the dataframe is split into 6 evenly-filled categories.

Discretization is performed using CRAN package "arules" [4]

1.3.2 Relative Maturity Discretization

Relative maturity, as defined by the syngenta problem statement, is a representation of the number of months that it takes for a soybean seed to mature, where lower values are preferred.

In order to inform the discretization strategy decision for relative maturity (RM), a histogram of the attribute is plotted in Figure 1.4.

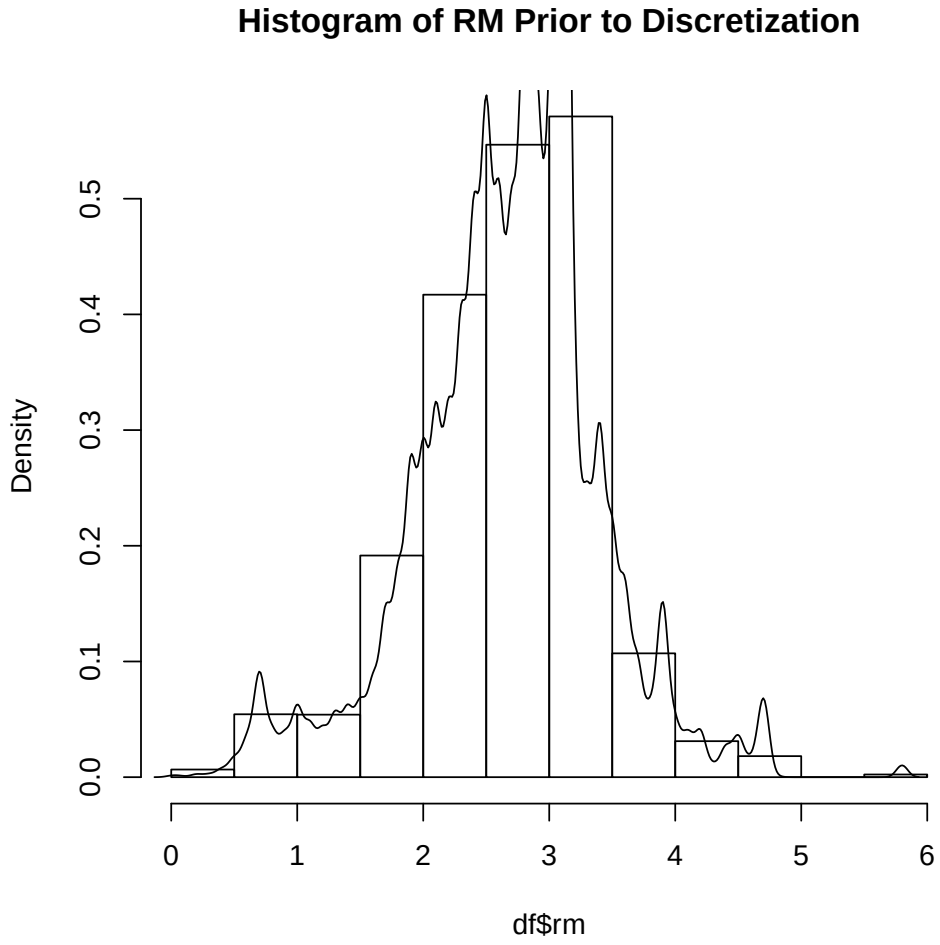

Figure 1.4 - Histogram of relative maturity prior to discretization

Figure 1.4 shows that RM, much like yield, is normally distributed. While RM is slightly skewed, it stands to reason that the same discretization strategy can apply.

[0.0,2.2)	[2.2,2.6)	[2.6,2.9)	[2.9,3.2)	3.2	[3.3,5.8]
47757	45682	42172	74584	7996	40062

Figure 1.5 - Number of rows in each discretized RM category

Figure 1.5 shows that, after discretization, the RM variable is split into 6 evenly-filled categories, much like yield.

1.3.3 An Additional Feature

An optimal seed will produce the maximal amount of soybeans in the minimum amount of time. Therefore, the yield-to-rm ratio may be a meaningful attribute when it comes to soybean seed evaluation.

The yield-to-rm ratio, which will be called 'yield ratio,' is computed using a row-wise quotient of yield over rm, and the distribution of the new attribute is shown in figure 1.6. Note that the row-wise quotient is computed before the yield and rm variables are discretized, so the result is a true division of the original attributes and is independent of previous discretization.

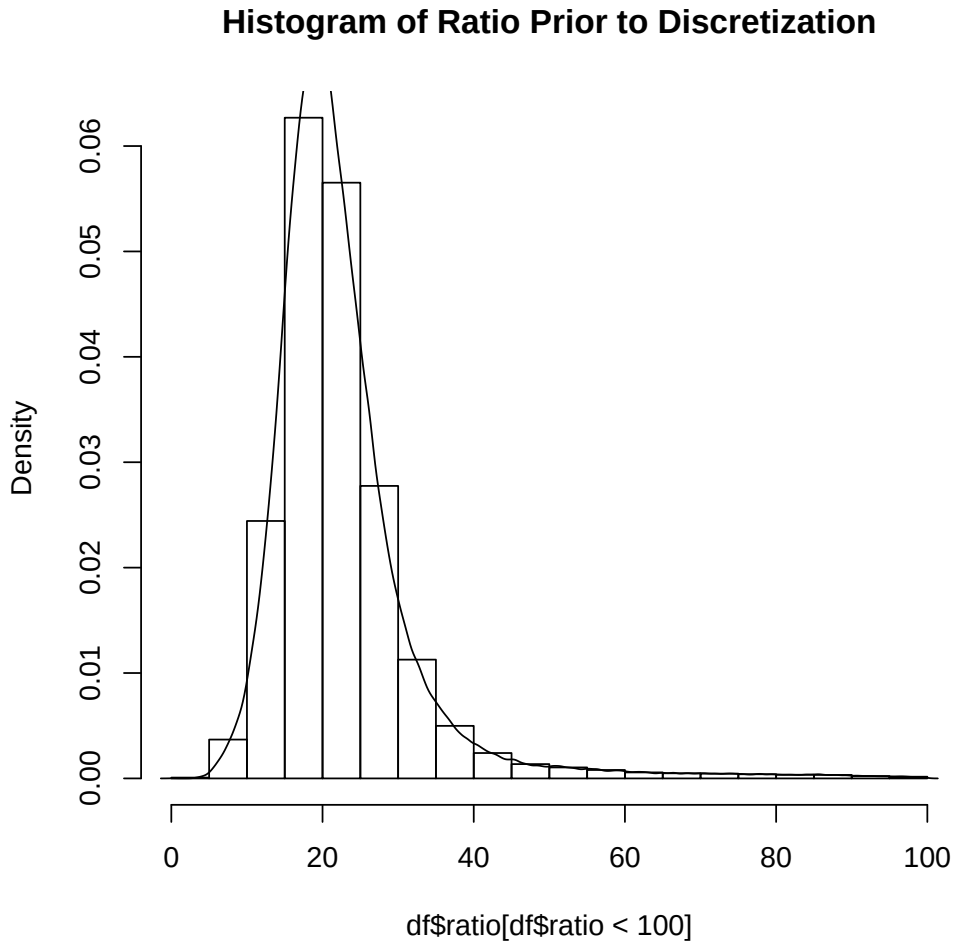

Figure 1.6 - Histogram of yield ratio prior to discretization

Yield ratio, like yield and RM, is normally distributed, so it can be discretized using the same strategy used for yield and RM (6-way frequency-based discretization).

```
[ 0.0,15.5) [15.5,18.3) [18.3,20.7) [20.7,23.5) [23.5,28.0) [28.0, Inf]
      43045       43043       43047       43034       43051       43033
```

Figure 1.7 - Number of rows in each discretized yield ratio category

Figure 1.7 shows that, after discretization, the yield ratio variable is split up into 6 evenly-filled categories.

1.4 Transaction Set Creation

The dataset will be partitioned into two transaction sets based on the "check" variable. The resulting two itemsets, therefore, will represent test seeds and check seeds, respectively. One set will include exclusively seeds that are already at market, and the other set will contain seeds that are currently being tested.

Check	42842
Test	215411

Figure 1.8 - Number of rows in each of the two transaction sets

Figure 1.8 shows the number of check and non-check rows in the original dataset, and therefore the number of rows in the two transaction sets being built. It is important to note that the non-check transaction set is much larger than the check transactions set.

1.4.1 Transaction Set Summary

```
transactions as itemMatrix in sparse format with
 42842 rows (elements/itemsets/transactions) and
 18326 columns (items) and a density of 0.0006548074

most frequent items:
         grad=.       bagsold=.      class_of=. locationGroup=3       year=2011
          33862           33862           33556           19757           14174
        (Other)
         378893

element (itemset/transaction) length distribution:
sizes
   12
42842

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     12      12      12      12      12      12

includes extended item information - examples:
     labels variables levels
1 year=2009      year   2009
2 year=2010      year   2010
3 year=2011      year   2011

includes extended transaction information - examples:
  transactionID
1             1
2             2
3             3
```

Figure 1.9 - Summary information for the check transaction set

Figure 1.9 shows a summary of the check transaction set, whose most frequent items are *grad=.* and *bagsold=.* (which represents all non-graduating seeds). This means that there will likely be a number of non-meaningful association rules built, as these items essentially represent missing data. However, this will be handled in rule post-processing by only selecting rules which are of interest.

The summary for the test transaction set is ommited in the interest of brevity.

1.4.2 Item Frequency Plot

Figure 1.10 shows an item frequency plot for individual items in the check transaction set.

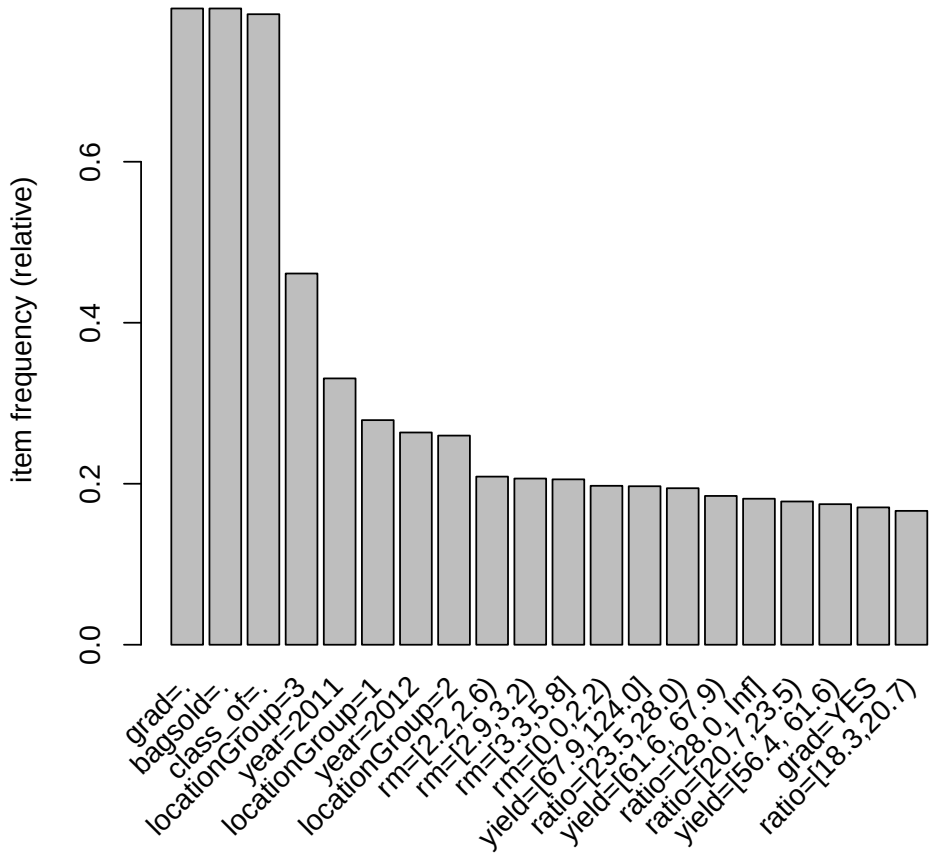

Figure 1.10 - Item frequency plot for check transaction set

Figure 1.10 shows that, beyond the 3 high frequency missing data items, the most frequent items are years and location groups. This is reasonable, since location group and year were discretized into 3 and 5 categories, respectively, while the yield, rm, and ratio statistics were discretized into 6 categories, and all of those categories are nearly equally filled.

2 Modeling

2.1 The Check Variety Itemset

First, the check variety itemset will be evaluated. A number of frequent itemsets and association rules will be generated with the check variety itemset, and they will be analyzed to unearth trends or meaningful information from the data.

2.1.1 Check Variety Frequent Itemsets

To begin our analysis, frequent itemsets will be mined from the check dataset with the default support constraint of 0.05. Figure 2.1 shows the first 10 most frequent itemsets (the first 10 items in the itemset after the set is sorted by support).

```
     items                          support
[1]  {bagsold=.}                    0.7903926
[2]  {grad=.}                       0.7903926
[3]  {grad=.,bagsold=.}             0.7903926
[4]  {class_of=.}                   0.7832501
[5]  {class_of=.,bagsold=.}         0.7832501
[6]  {class_of=.,grad=.}            0.7832501
[7]  {class_of=.,grad=.,bagsold=.} 0.7832501
[8]  {locationGroup=3}              0.4611596
[9]  {year=2011}                    0.3308436
[10] {year=2011,class_of=.}         0.3308436
```

Figure 2.1 - Top 10 most frequent itemsets

Figure 2.1 shows that the majority of the high-frequency itemsets include the "." characteristic, which represents missing data in the original dataset. Therefore, we will eliminate the rules that include that characteristic and show the new top-5.

```
    items              support
[1] {locationGroup=3} 0.4611596
[2] {year=2011}       0.3308436
[3] {locationGroup=1} 0.2790253
[4] {year=2012}       0.2636665
[5] {locationGroup=2} 0.2598151
```

Figure 2.2 - Top 5 most frequent itemsets, excluding "." items

Figure 2.2 shows that the most frequent itemsets are, in fact, single item itemsets which represent years and locations. This is expected, given the results shown in Figure 1.9.

2.1.2 Closed Itemsets

The list of mined itemsets from the previous section will be subsetted to isolate only closed itemsets. Closed itemsets are itemsets for which no immediate supersets have the same support as the itemset.

```
    items                support
[1] {locationGroup=3}    0.4611596
[2] {year=2012}          0.2636665
[3] {locationGroup=2}    0.2598151
[4] {rm=[2.2,2.6)}       0.2088605
[5] {yield=[67.9,124.0]} 0.1968862
```

Figure 2.3 - Top 5 most frequent closed itemsets, excluding "." items

Figure 2.3 shows the top 5 most frequent closed itemsets. Once again, the itemsets have been post-processed to exclude itemsets which include "." items. This is the first figure of frequent itemsets which includes some non-trivial itemsets. It shows that $rm = [2.2, 2.6)$ and $yield = [67.9, 124.0]$ are both frequent itemsets. However, because these are single-item itemsets, and the items in question were discretized using equal frequency, these particular itemsets are not particularly meaningful.

That being said, this may nonetheless meaningful, because it means that these particular relative maturity and yield ranges occur more frequently in the check dataset than in the test dataset, since the two datasets were discretized together. This relative maturity range is the second lowest range, so this may indicate that low relative maturity seeds are more likely to go to market, because the seeds at market being used at tests are most likely to have low relative maturity.

The same can be said for the yield range. The most frequent yield range in the check data is the highest possible yield range. This indicates that high-yield seeds are more likely to go to market, as the seeds at markey being used for testing have high yield. Therefore, at this point it can be, on a working basis, inferred that high-yield and low-rm seeds are optimal for going to market.

2.1.3 Maximal Itemsets

The original list of mined itemsets will be subsetted once more, but this time to isolate maximal itemsets. Maximal frequent itemsets are frequent itemsets for which none of the immediate supersets is frequent.

```
    items                    support
[1] {year=2012,
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=3}    0.13094627
[2] {rm=[2.6,2.9),
     class_of=.,
```

```
     grad=.,
     bagsold=.}          0.10571402
[3] {yield=[ 0.0, 45.0),
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=2}    0.09056533
[4] {year=2013,
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=3}    0.08578031
[5] {yield=[67.9,124.0],
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=3}    0.07357266
[6] {rm=[2.9,3.2),
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=3}    0.07347930
```

Figure 2.4 - Top 5 most frequent maximal itemsets

Figure 2.4 shows the top-frequency maximal itemsets. Here, the "." items have been included, as all of the frequent maximal itemsets include some non-"." attributes alongside the "." attributes. In this set of itemsets, we see less meaningful data than in the prior, as each itemset includes a locationGroup and year item, which are soemwhat novel attributes that, in isolation, don't provide much insight to our analysis.

2.2 Test Variety Frequent Itemsets

To compare with the check variety itemset frequent itemsets, a set of frequent itemsets will be generated from the test transaction set.

```
    items              support
[1] {year=2011}        0.4165618
[2] {locationGroup=3}  0.3718287
[3] {year=2012}        0.3338687
[4] {locationGroup=1}  0.3175372
[5] {locationGroup=2}  0.3106341
```

Figure 2.5 - Top 5 most frequent itemsets, exluding "." items, for the test data

Figure 2.5 shows the top 5 itemsets by frequency from the test dataset, exluding missing (".") items. Much like for the test dataset, we can see that these itemsets are not particularly meaningful, as they are all single-item itemsets representing years and location groups.

2.2.1 Closed Itemsets

Next, closed itemsets will be subsetted from the entire set of frequent itemsets from the test dataset.

```
    items              support
[1] {year=2011}        0.4165618
[2] {locationGroup=3}  0.3718287
[3] {year=2012}        0.3338687
[4] {locationGroup=1}  0.3175372
[5] {rm=[2.9,3.2)}     0.3051794
```

Figure 2.6 - Top 5 most frequent closed itemsets, excluding "." items

The most frequent closed itemsets, shown in Figure 2.6, are mostly trivial once again. However, there is one non-trivial itemset, which is the itemset $rm = [2.9, 3.2]$. This one-item itemset shows that the most frequent relative maturity range for the test data is the 3rd highest of the ranges. Since there are only 6 ranges for the relative maturity, this isn't particularly profound, but it does perhaps indicate that higher-rm seeds are more abundant in the test dataset than they are in the check dataset.

2.2.2 Maximal Itemsets

Finally, the maximal itemsets are subsetted from the entire frequent itemset set from the test dataset.

```
    items                      support
[1] {location=3210,
     class_of=.,
     grad=.,
     bagsold=.}          0.10123438
[2] {year=2011,
     rm=[3.3,5.8],
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=1}    0.07514008
[3] {year=2011,
     rm=[2.2,2.6),
     class_of=.,
     grad=.,
     bagsold=.}          0.07349207
[4] {yield=[45.0, 51.4),
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=2}    0.07345957
```

```
[5] {yield=[67.9,124.0],
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=3}    0.07182084
[6] {yield=[ 0.0, 45.0),
     class_of=.,
     grad=.,
     bagsold=.,
     locationGroup=2,
     ratio=[ 0.0,15.5)}  0.07169550
```

Figure 2.7 - Top 5 most frequent maximal itemsets

The high-support maximal itemsets shown in Figure 2.7 show no obvious patterns, but instead a slew of various locations, years, relative maturity ranges, andyield ranges. Interestingly, one ratio range appears, the lowest possible range, but this is a single item out of 6 items, and so it likely has no strong implications about the test data.

2.3 Itemset Comparison

As a first novel comparison of the two sets of itemsets, we can compare the distribution of itemset cardinalities between the two transaction sets. Figures 2.8 shows the distributions of size of itemset for the check and test sets.

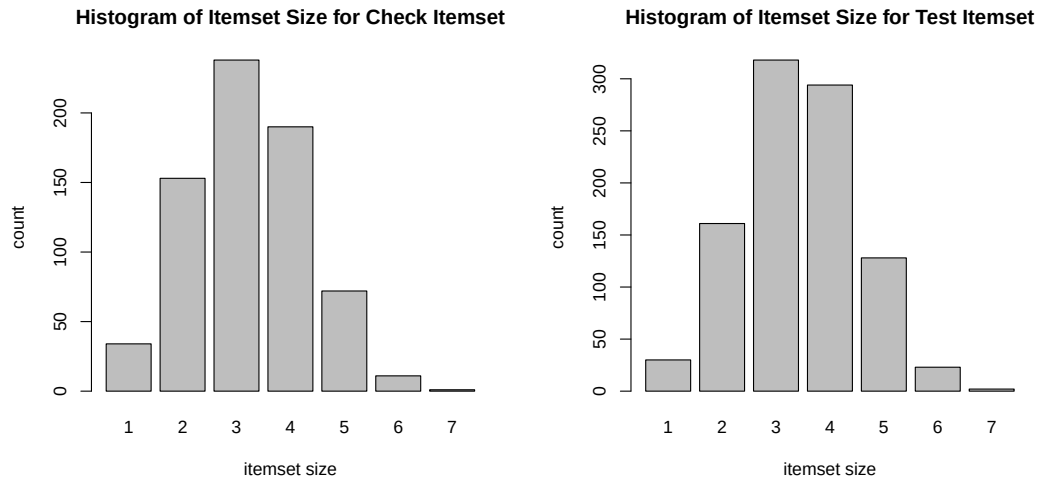

Figure 2.8 - Distribution of frequent itemset cardinality from check and test itemsets

Figures 2.8 show nearly identical distributions, which, on a structural level, indicates that the two datasets are quite similar. However, these plots provide little inference to any meaningful difference between the two datasets.

As a second comparison, the number of closed and maximal itemsets from the two datasets will be compared in Figure 2.9.

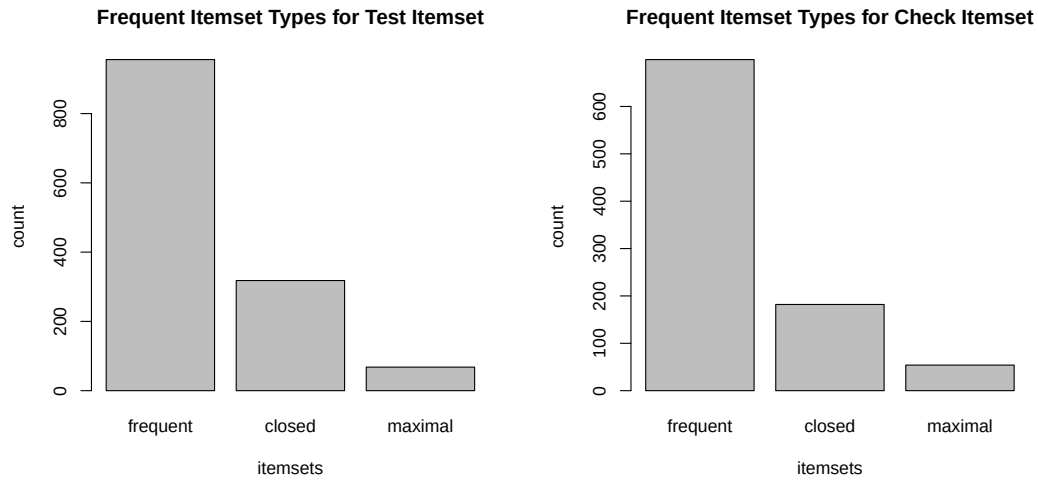

Figure 2.9 - Frequent itemset types for check and test itemsets

Figure 2.9 shows that, once again, the two itemsets appear to be very similar. Therefore, from an itemset-based perspective, there, in general, doesn't appear to be a large difference between the two. Association rules will now be mined to search for more insight into the difference between the two itemsets.

2.4 Test Variety Association Rules

To begin, association rules will be mined from the test transaction set with default apriori parameters, and the rules will be sorted by descending lift.

```
    lhs                rhs                  support confidence     lift
[1] {year=2011,
     rm=[0.0,2.2),
     class_of=.}    => {locationGroup=2} 0.1054821  0.8816887 2.838351
[2] {year=2011,
     rm=[0.0,2.2),
     bagsold=.}     => {locationGroup=2} 0.1054821  0.8816887 2.838351
[3] {year=2011,
     rm=[0.0,2.2),
     grad=.}        => {locationGroup=2} 0.1054821  0.8816887 2.838351
[4] {year=2011,
     rm=[0.0,2.2),
     class_of=.,
     bagsold=.}     => {locationGroup=2} 0.1054821  0.8816887 2.838351
[5] {year=2011,
     rm=[0.0,2.2),
     class_of=.,
     grad=.}        => {locationGroup=2} 0.1054821  0.8816887 2.838351
[6] {year=2011,
     rm=[0.0,2.2),
```

```
     grad=.,
     bagsold=.}    => {locationGroup=2} 0.1054821  0.8816887 2.838351
```

Figure 2.11 - Top-lift rules for the check transaction set

460 rules were mined from the test transaction set using default apriori parameters. Figure 2.10 shows the top rules from the mined ruleset, based on lift. None of these rules are particularly meaningful, since they are all predicting locationGroup. However, there are likely more meaty rules inside of the ruleset, which should become useful in visualizations comparing the check and test rulesets.

2.5 Check Variety Association Rules

The same apriori parameters are used to mine rules from the check transaction set.

```
    lhs                 rhs                   support confidence     lift
[1] {class_of=2011} => {grad=YES}        0.1052472  0.8755340 5.131276
[2] {rm=[0.0,2.2),
     class_of=.}    => {locationGroup=2} 0.1398161  0.8093501 3.115100
[3] {rm=[0.0,2.2),
     grad=.}        => {locationGroup=2} 0.1398161  0.8093501 3.115100
[4] {rm=[0.0,2.2),
     bagsold=.}     => {locationGroup=2} 0.1398161  0.8093501 3.115100
[5] {rm=[0.0,2.2),
     class_of=.,
     grad=.}        => {locationGroup=2} 0.1398161  0.8093501 3.115100
[6] {rm=[0.0,2.2),
     class_of=.,
     bagsold=.}     => {locationGroup=2} 0.1398161  0.8093501 3.115100
```

Figure 2.11 - Top-lift rules for the test transaction set

315 rules were mined from the check transaction set with default apriori parameters. Figure 2.11 shows the top rules from that ruleset, based on lift. The first rule is the only interesting rule among them, which shows that, with a confidence of 0.8755, seeds posed to be in the class of 2011 had a graduated value of "YES." This likely indicates that 2011 was a highly non-competitive year for the soybean seed graduates.

2.6 Comparison and Visualization

In the following section, a number of visualizations will be used to attempt to discern a distinction between the two transaction sets through the use of the association rules mined from those sets.

2.6.1 Support vs Confidence

The first visual comparison of the two sets of association rules will be a support vs confidence plot for all of the rules in each test set, color-coded by lift. Note that some jitter is added to each of the plots to separate duplicate points on the plots.

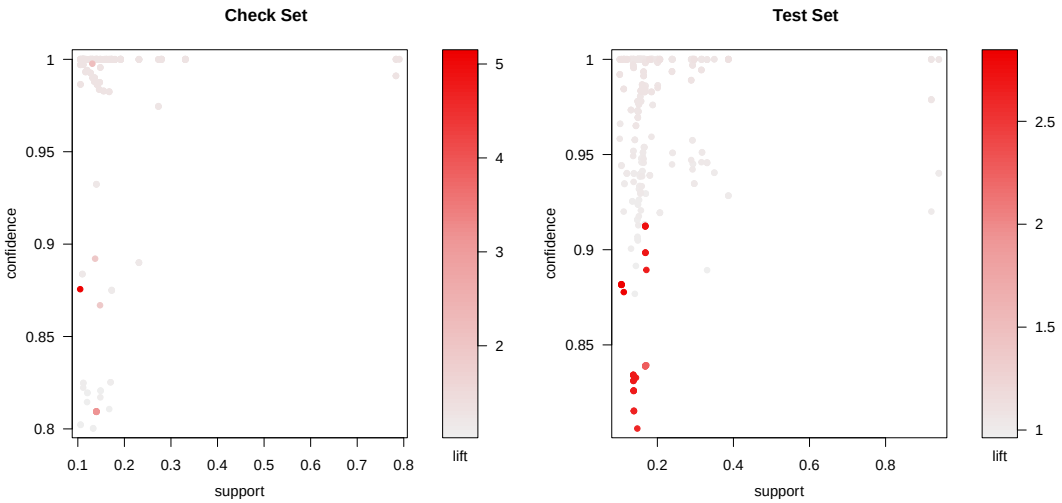

Figure 2.12 - Support vs Confidence plots for the two sets of association rules

2.6.2 Grouped Matrix Plots

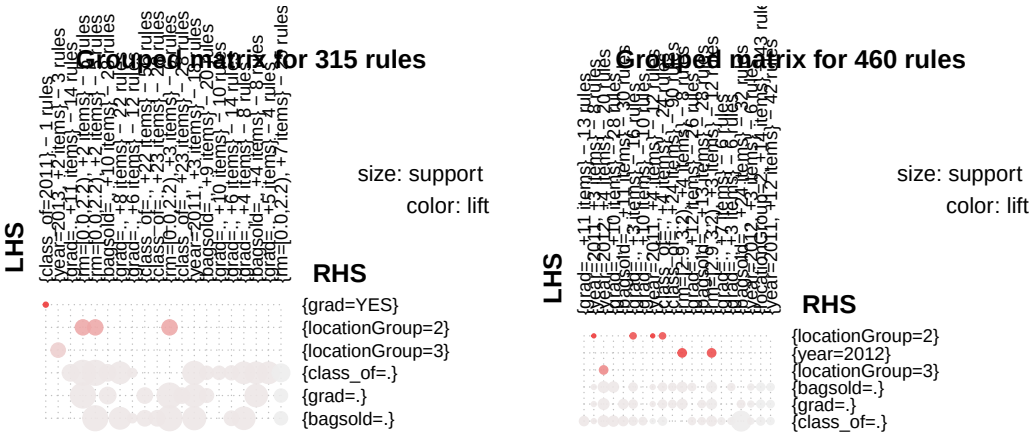

Figure 2.13 - Grouped matrix plots for the two rulesets

2.6.3 Ruleset Graph Plots

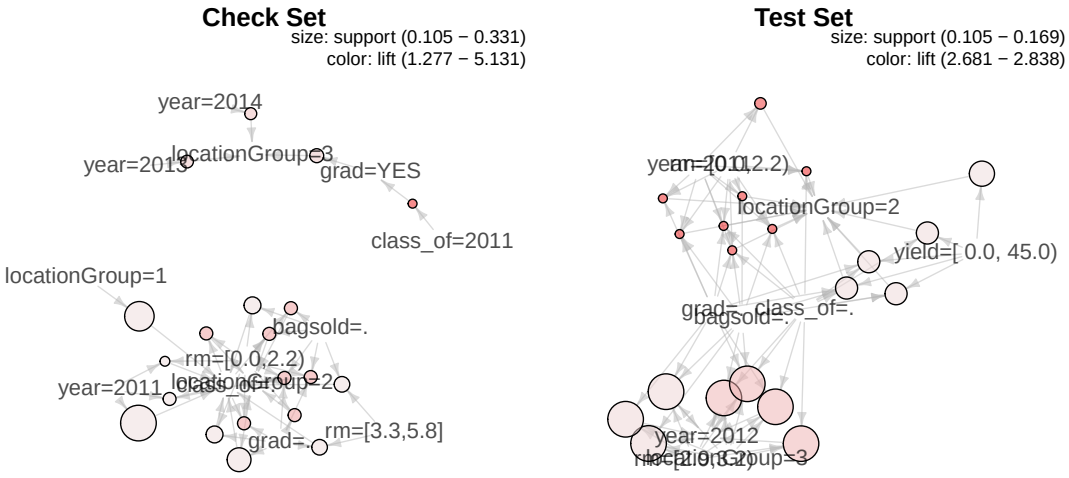

Figure 2.14 - Ruleset graphs for the two rulesets

3 Evaluation

4 Conclusion

References

[1] "2017 Problem." - INFORMS O.R. and Analytics Student Team Competition. N.p., n.d. Web. 03 Nov. 2016.

[2] Friedrich Leisch. A Toolbox for K-Centroids Cluster Analysis. Computational Statistics and Data Analysis, 51 (2), 526-544, 2006.

[3] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan. (2016). caret: Classification and Regression Training. R package version 6.0-68. https://CRAN.R-project.org/package=caret

[4] Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik (2016). arules: Mining Association Rules and Frequent Itemsets. R package version 1.5-0. https://CRAN.R-project.org/package=arules