Association Rule Analysis for Syngenta Soybean Data

Association Rule Mining for Data Understanding

Ian Johnson

Southern Methodist University

November 6, 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.

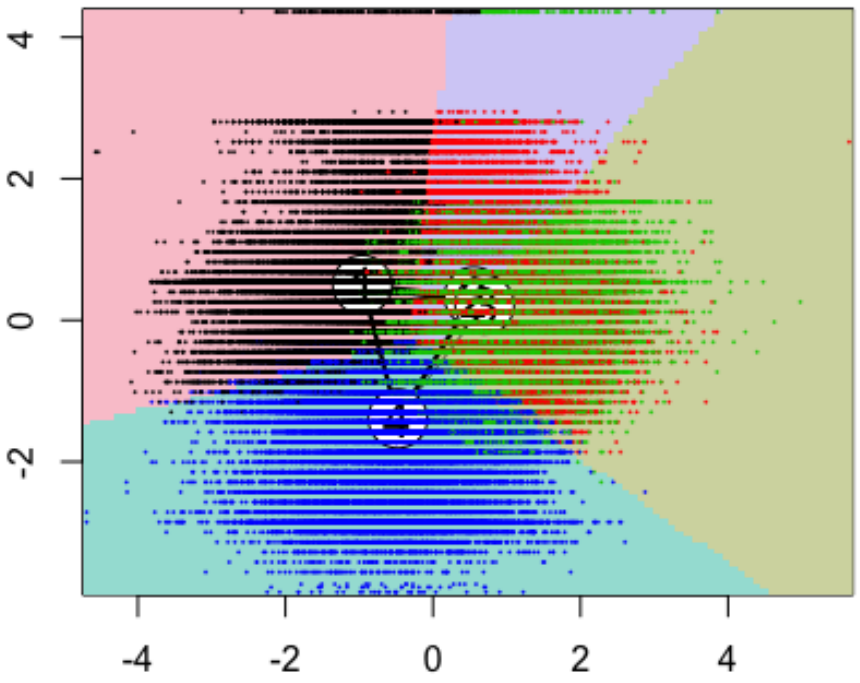

Figure 1.1 - A visualization of the K-means clustering algorithm output, with K=4

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.

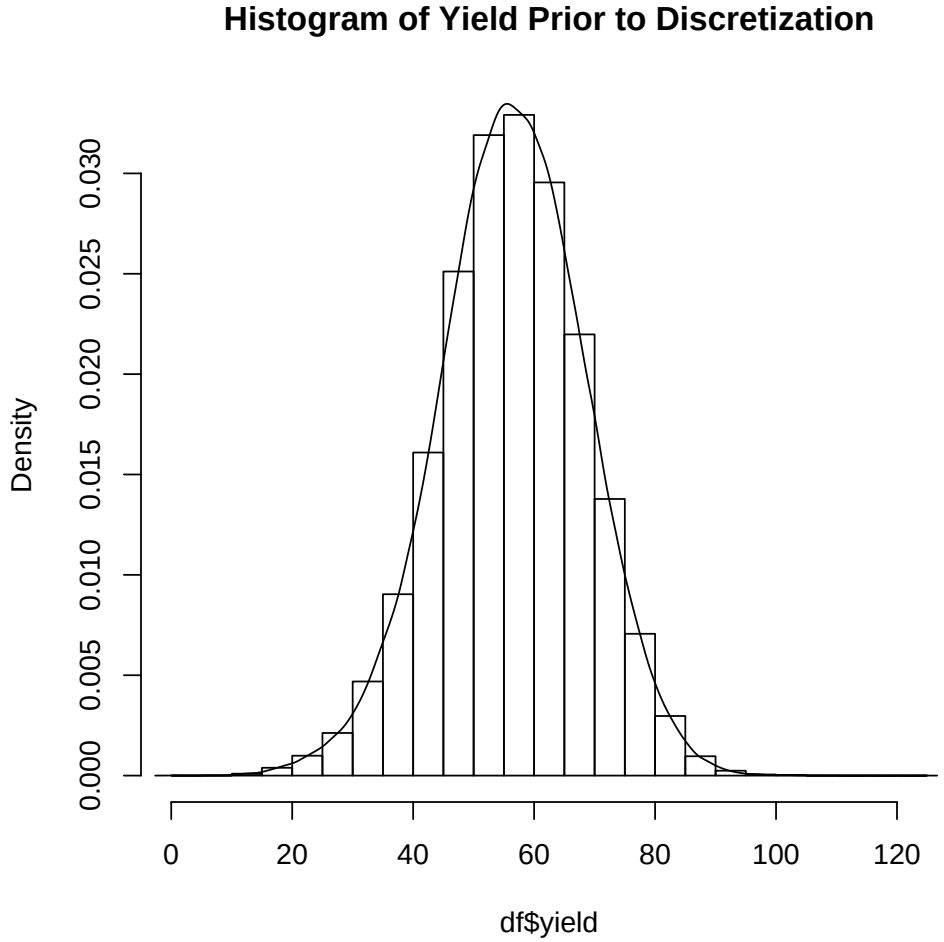

Figure 1.2 - Histogram of yield prior to discretization

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.

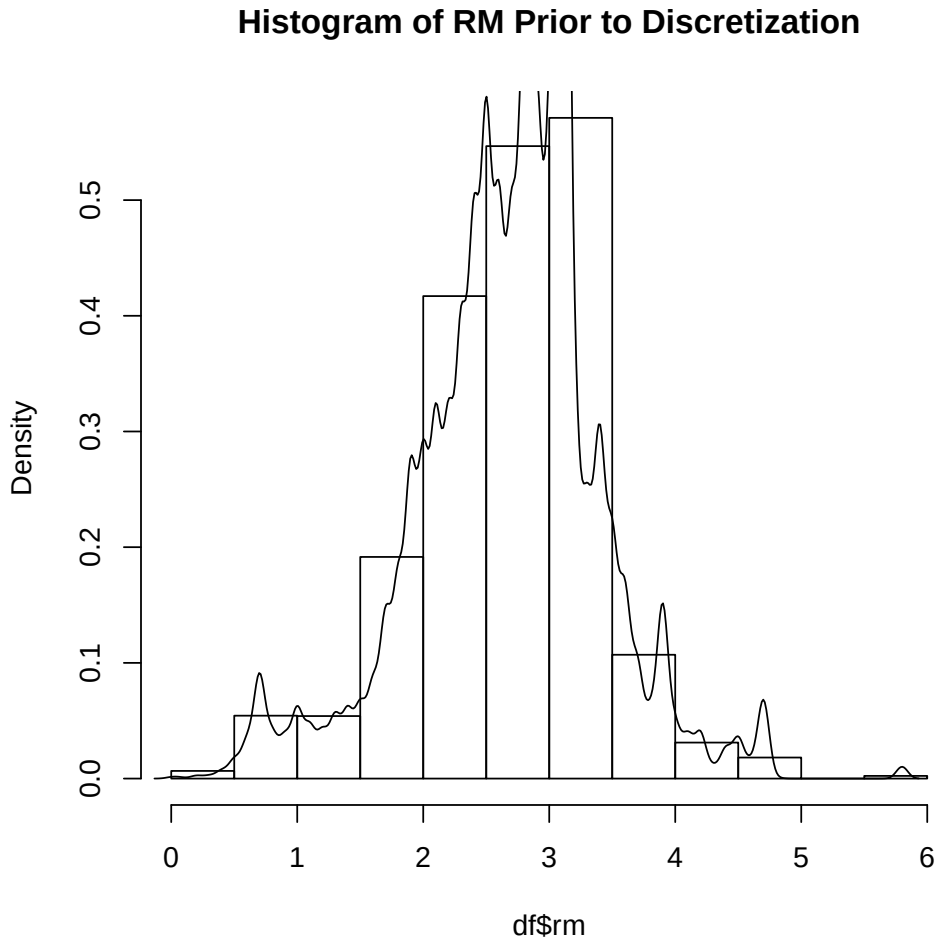

Figure 1.4 - Histogram of relative maturity prior to discretization

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	0.05	0.08	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1
30	20	14	73	72	194	451	870	2510	1029	962	1650	1196
1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2	2.1	2.2	2.3	2.4
1076	1453	1565	1688	2153	3765	4487	7097	7234	8168	7990	10148	12510
2.5	2.6	2.7	2.8	2.9	3	3.1	3.2	3.3	3.4	3.5	3.6	3.7
15034	12758	12284	17130	15203	13214	46167	7996	6095	7979	5507	4341	2627
3.8	3.9	4	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	5.8	
1449	4194	1207	997	1092	247	706	975	365	1963	20	298	

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.

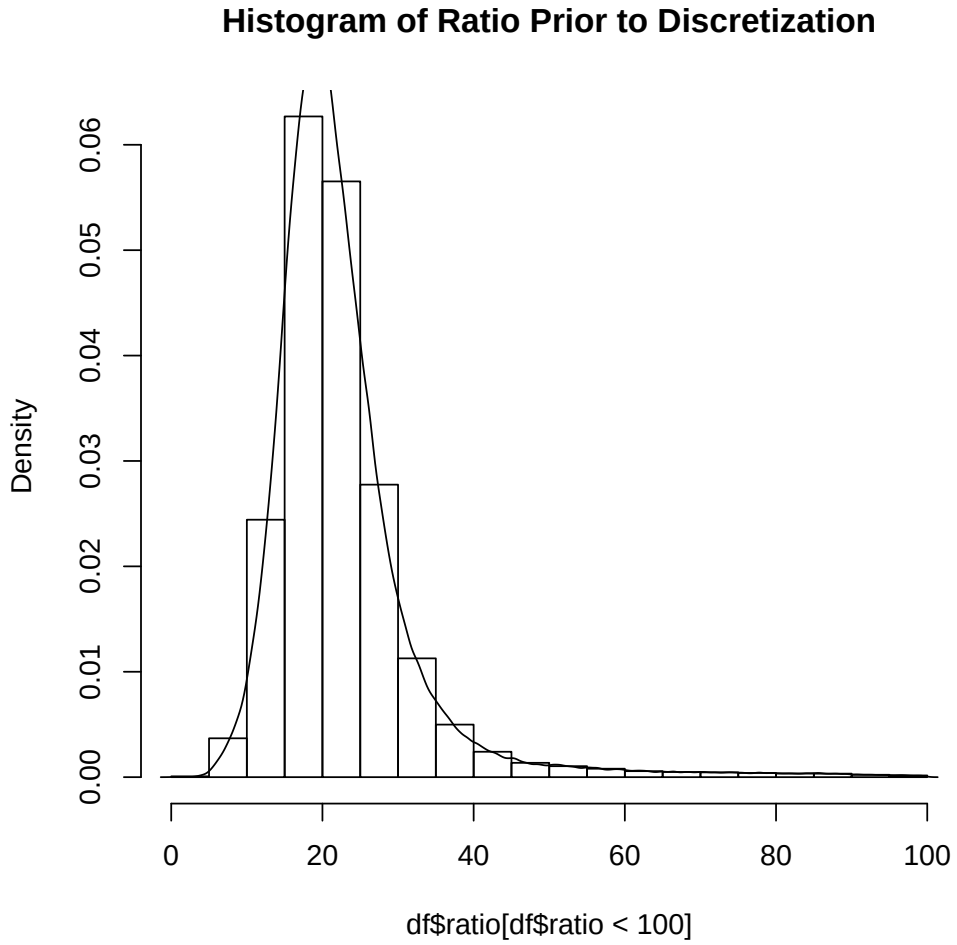

Figure 1.6 - Histogram of yield ratio prior to discretization

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.

2 Modeling

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