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.

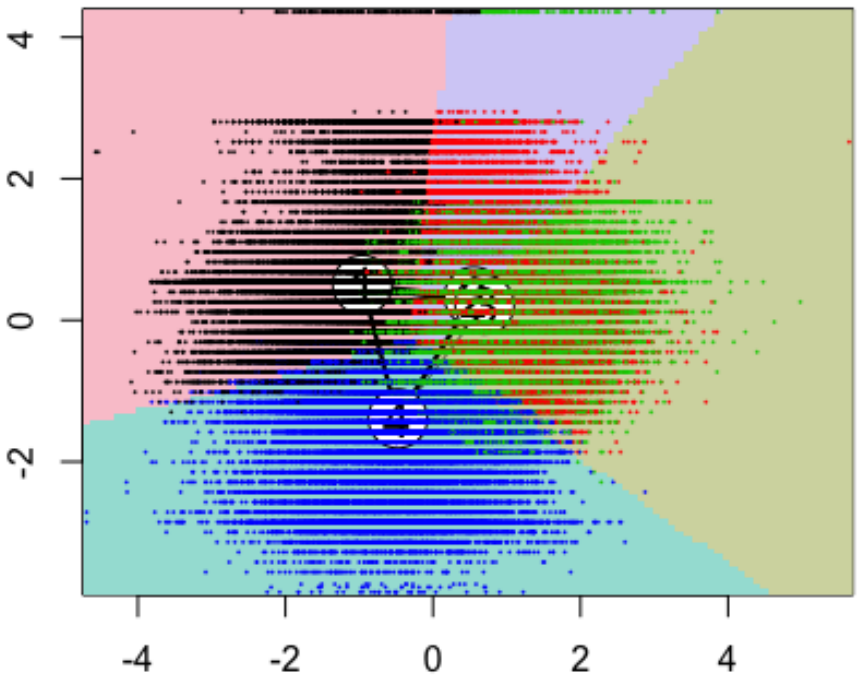

Figure 1.1 - A visualization of the K-means clustering algorithm output, with K=4

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

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.

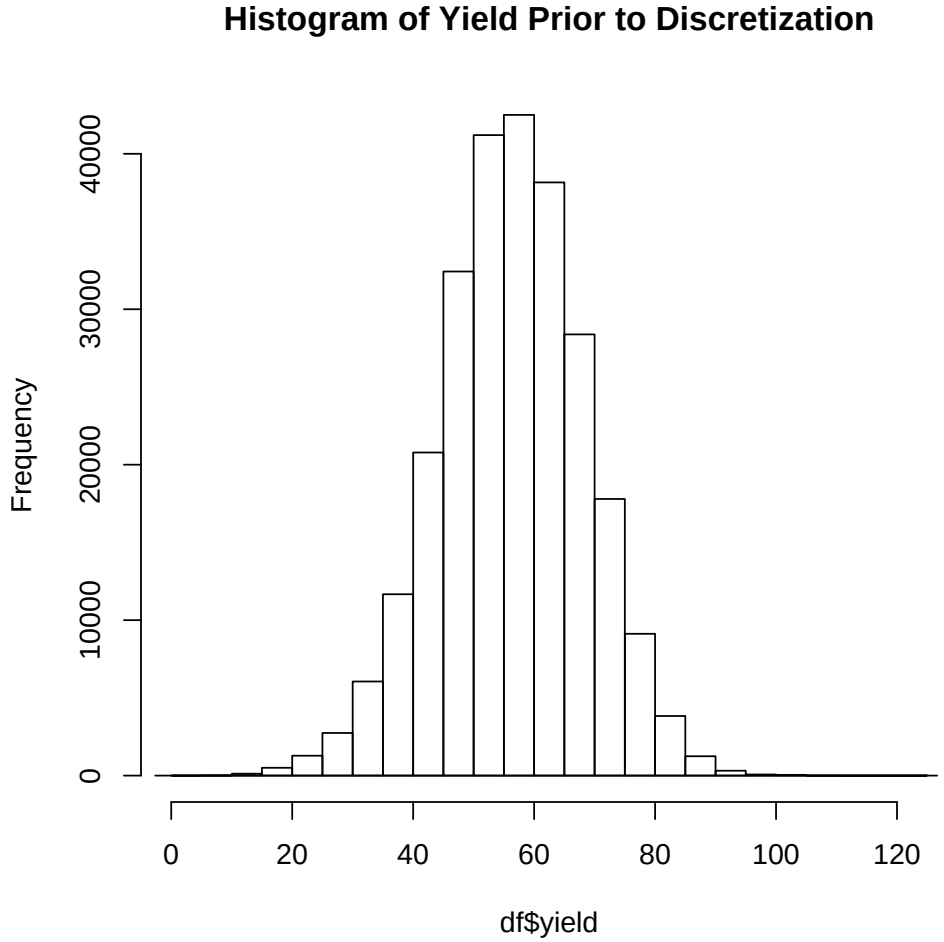

Figure 1.1 - Histogram of yield prior to discretization

1.4 Transaction Set Creation

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