NML502 - Final Project

Elliot Smith, Eugen Hruska, Varun Suriyanarayana

4/22/2018

Introduction

In this report, we will be comparing the clustering results of a dataset of different beer types using two distinct methods, K Means and Self-Organizing Maps. Our previous experience tells us that when data reaches a particular level of complexity that K Means is no longer an acceptable method for clustering; we hope to prove or debunk this assertion. In addition, given our hypothesis is correct, that K Means is not a satisfactory method, we hope to prove that using an SOM network is a significantly stronger technique for identifying clusters in multi-dimensional data. We hope that you will learn something through this process as we have.

Motivation

There were many motivations for us to perform this analysis:

- We love beer! While trying to decide on a project topic and looking for similar interests, we found that each of us enjoyed beer and that trying to cluster multi-dimensional beer data would be both interesting and fun; it certainly made working on the project very engaging as we were curious to test the results of our analysis against our previous experience.
- Our group had a strong preference to explore Unsupervies Learning (as opposed to a Supervised approach). As a team, we were certainly more interested in the concept of trying to find the solution without providing the answer; the potential power of accurate prediction without supervision was very interesting to us. However, having the answers available to us was also important, so that we could test how well our analyses performed.
- Self-Organizing Maps is a topic that deserved further exploration. We decided that an SOM was a robust tool that was perhaps best equipped to solve our problem with an optimal solution and we also had an intrinsic desire to explore the network's abilities more in a multi-dimensional setting; this project represented the perfect opportunity.
- Each of us had the desire to pursue the topic of classification; it worked to our benefit that the data we chose was very well-suited to our task. The concept that we can try to determine a observations class (in this case, a style of beer) by developing a learning algorithm was a major driver for us. Classification power is an extremely robust topic; the ability to classify observations based on their features is a topic worth exploring!
- Finally, we wanted to compare how Self-Organizing Maps compared against a more conventional clustering method (in this case, K Means). Since SOMs are a relatively new topic to us, we wanted to understand how it would compare to methods that we had more familiarity with in an effort to expand our statistical toolsets and perhaps replace our primitive tools (K Means) with a more robust one (SOMs).

Objectives

Our objectives with this analysis are the following:

- Compare visually how effective K Means and SOMs are at clustering the data. We want to get a visual representation of how the two methods cluster the data in an effort to determine their efficiency at completing this task.
- Compare the successful classification rates for K Means and SOMs to determine which clustering performed better when we take into account the class labels. This simple number is our key statistic, it tells us that given our method used, we expect to correctly classify a particular percent of inputs.
- Determine if any subclasses exist within our defined classes based on the results of the clustering analysis. This result will work both ways in that it will help us understand how the clustering is realized, in addition to helping us understand differences internal to each class.
- Finally, we want to learn something new about beer that we did not know beforehand!

Our Data

Beer Data

Our data came from the kaggle.com website, a repository for publicly available data for personal or research analyses. The original data came with 73,861 observations and 23 distinct features. We decided to investigate the following beer styles:

- American Pale Ale
- Imperial IPA
- American IPA
- Saison
- American Brown Ale
- Witbier
- American Amber Ale
- Irish Red Ale
- American Stout
- American Light Lager

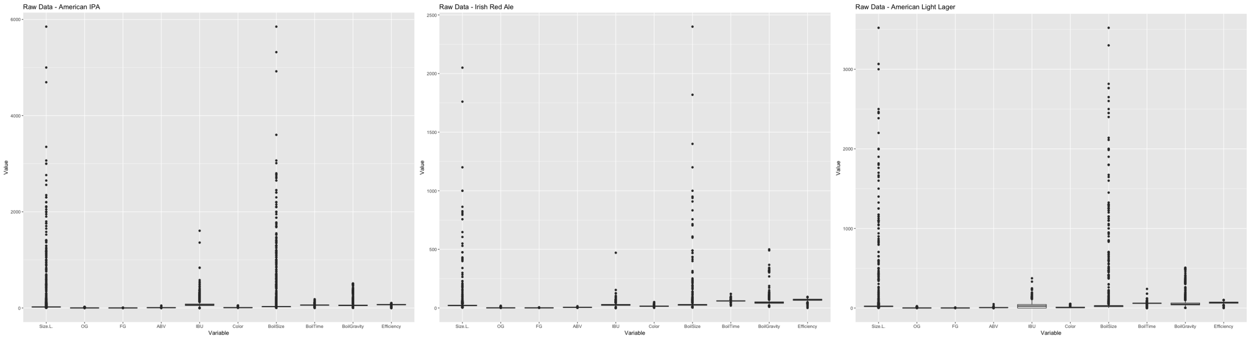
In addition, we decided to investigate the impacts of the following features:

- Size (L) - the amount brewed for each observation
- OG - the original density of the wort (key ingredient in the brewing process) compared to the water before fermenation
- FG - the final density of the wort compared to the water after fermentation
- ABV - alcohol by volumne
- IBU - international bittering units (an international standard for the beer's level of bitterness)
- Color - a numerical scale representing light to dark beer coloring
- BoilSize - the amount of fluid at the beginning of the boil process
- BoilTime - the amount of time the wort is boiled
- BoilGravity - density of the wort compared to the water before the boil
- Efficiency - how efficient the mash was at extracting sugar from the grain

Our reduction down to these selected beer styles (and after removing any rows that had any of the selected variables marked as NA) left us with 31,419 observations.

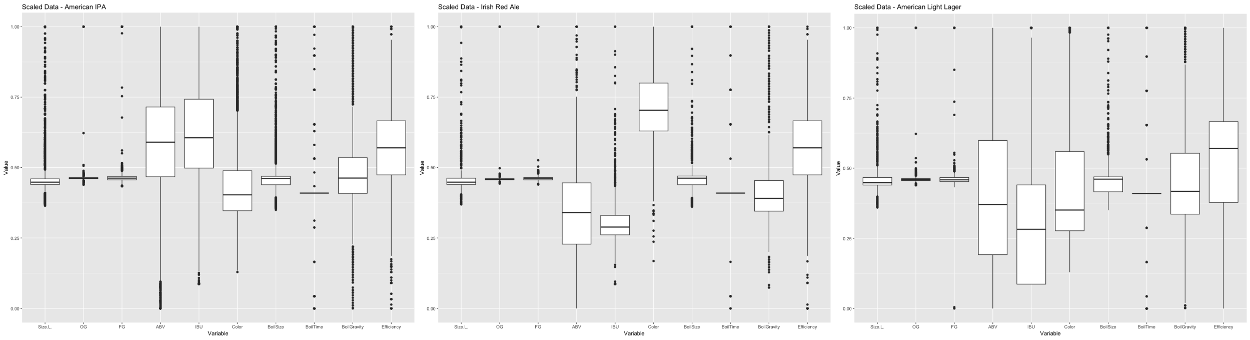
Raw Data Issues

A preliminary look at our raw data told us that each of our features were heavily affected by outlying values. Directly below is a look at three of our selected beer styles (from left to right: American IPA, Irish Red Ale and American Light Lager) where we have created a box plot for each to visualize how spread the data is for each feature. As we can see, our data is heavily impacted by outlying values. This is something that we will want to remedy so that points that are particularly outside of the inter-quartile range do not adversely affect our analysis by giving those values too much weight.



Data Transformation

Our solution to the outlying data issues was to scaled 2 standard deviations of our data between 0 and 1. This transformation allowed us to remove all outlying values that would overly affected our analysis and normalize between 0 and 1 so that no single feature with a large range would dominate the analysis. Please see below for the transformation on the previously displayed beer styles (from left to right: American IPA, Irish Red Ale and American Light Lager).

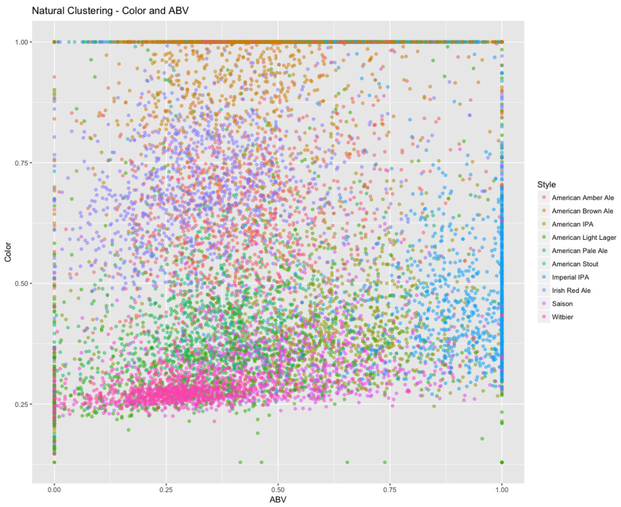


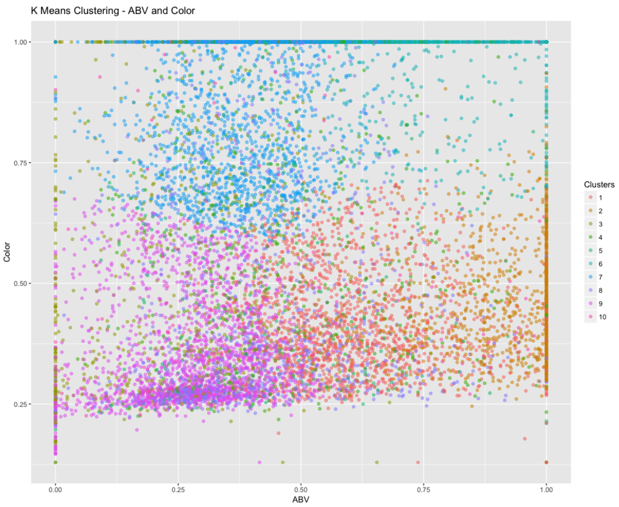
Our Analysis

K Means

To begin our K Means analysis, we will begin by selecting notable features (based on our discretion); we have selected Color, ABV and IBU. We will then plot the natural clusters based on our cleaned data, to show the relation between each of these variables, based on the accompanying class label. Then, we will use R's kmeans() function with our data, specify that we are looking for ten clusters, and then visualize the same plots as above, however, in the K Means cases, we will color them observations based on the cluster they have been assigned to. In this way, we can compare the natural clusters against how K Means clusters the data.

Color and ABV

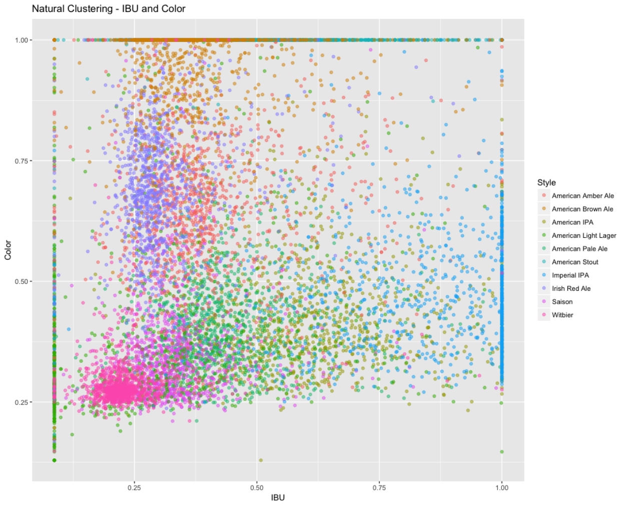


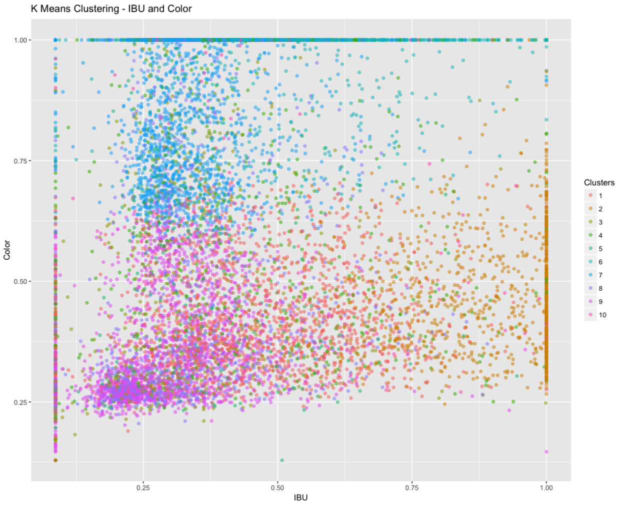


We draw the following notable conclusions:

- With the natural clustering, we see approximately five dominant clusters:
  - American Brown Ale - Light Brown
  - Irish Red Ale - Purple
  - American Light Lager - Green
  - Witbier - Magenta
  - Imperial IPA - Blue
- With K Means, we see approximately four dominant clusters:
  - Cluster 1 - Red
  - Cluster 2 - Light Brown
  - Cluster 7 - Blue
  - Cluster 9 - Light Purple
- Overall, we see much stronger, specific color density with the K Means results; we can infer this is because closer observations are lumped together in the same cluster
- It appears that Cluster 7 is a combination of American Brown Ale and the Irish Red Ale with a higher Color value
- Cluster 9 appears to be a combination of Witbier and the Irish Red Ale with a lower Color value

IBU and Color

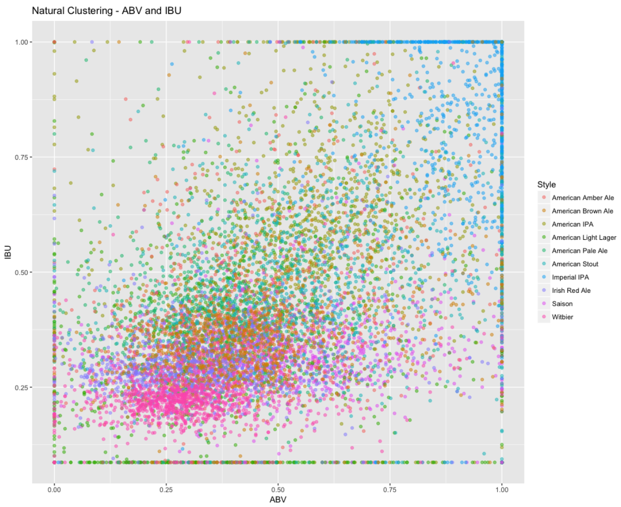


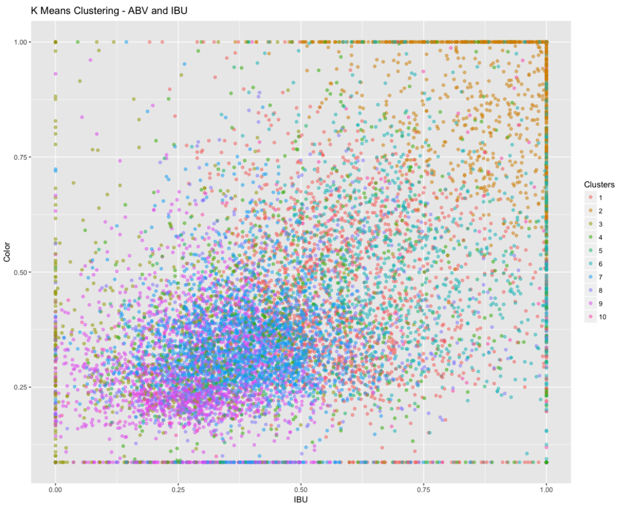


We draw the following notable conclusions:

- With the natural clustering, we see approximately seven dominant clusters:
  - American Amber Ale - Red
  - American Brown Ale - Light Brown
  - Irish Red Ale - Purple
  - American Light Lager - Green
  - Witbier - Magenta
  - Saison - Light Purple
  - Imperial IPA - Blue
  - American IPA - Green-Brown
- With K Means, we see approximately three dominant clusters:
  - Cluster 2 - Light Brown
  - Cluster 7 - Blue
  - Cluster 9 - Light Purple
- We see very defined natural clusters, which is impressive because there are seven distinct ones; however, our clusters with K Means are much less strong
- It looks as though K Means transformed the natural cluster amalgam of American Light Lager and American Brown Ale and split them up evenly between two clusters (Clusters 2 and 9)
- In this case, we believe that K Means negatively impacted the natural clustering process; where there were seven very strongly defined clusters naturally, K Means created only three dominant clusters which are quite ill-defined

ABV and IBU





We draw the following notable conclusions:

- With the natural clustering, we see approximately four dominant clusters:
  - Irish Red Ale - Purple
  - American Light Lager - Green
  - Witbier - Magenta
  - Imperial IPA - Blue
- With K Means, we see approximately three dominant clusters:
  - Cluster 2 - Light Brown
  - Cluster 7 - Blue
  - Cluster 9 - Light Purple
- Aside from Imperial IPA and Witbier, we dont't see impressive natural clustering; the large center cluster seems to be a mixture of Irish Red Ale, American Brown Ale and American Light Lager
- The K Means again did a poor clustering in this case; there appears to be the three clusters we defined (with Clusters 7 and 9 being very similar) and not much else
- There is clearly a one-to-one relationship between the Natural Cluster of Imperial IPA and K Means Cluster 2

Classification Rate

To determine the K Means classification rate, we first determined how many of each Beer Style contributed to each of our contrived K Means clusters (columns 1-10 below); in this way, we can see which Beer Styles contribute the most to each Cluster. The table below shows us this result:

	1	2	3	4	5	6	7	8	9	10
American Amber Ale	65	124	22	336	107	106	17	98	20	133
American Brown Ale	175	7	20	590	77	88	20	38	5	8
American IPA	32	69	26	12	84	82	22	50	157	494
American Light Lager	56	330	56	83	49	87	32	129	39	167
American Pale Ale	4	430	29	21	99	58	22	80	21	264
American Stout	444	4	23	333	44	106	27	35	5	7
Imperial IPA	75	6	5	1	19	41	24	6	780	71
Irish Red Ale	53	137	43	478	111	73	22	76	2	33
Saison	21	267	28	27	227	55	18	65	8	312
Witbier	2	532	26	8	244	42	25	99	5	45

Using the above result, we can see the fraction of Beer Style contributions to each K Means Cluster. Taking it one step further, we can determine which Beer Style is the "winner" (drawing a parallel with the winning methodology used by SOM) of the corresponding Cluster based on the highest percent contributor to that cluster; and can then refer to the observations that map to these clusters of the winning Beer Style as correctly classified. The table below shows this percent result:

	1	2	3	4	5	6	7	8	9	10
American Amber Ale	0.0701	0.0651	0.0791	0.1779	0.1008	0.1436	0.0742	0.1450	0.0192	0.0867
American Brown Ale	0.1888	0.0037	0.0719	0.3123	0.0726	0.1192	0.0873	0.0562	0.0048	0.0052
American IPA	0.0345	0.0362	0.0935	0.0064	0.0792	0.1111	0.0961	0.0740	0.1507	0.3220
American Light Lager	0.0604	0.1731	0.2014	0.0439	0.0462	0.1179	0.1397	0.1908	0.0374	0.1089
American Pale Ale	0.0043	0.2256	0.1043	0.0111	0.0933	0.0786	0.0961	0.1183	0.0202	0.1721
American Stout	0.4790	0.0021	0.0827	0.1763	0.0415	0.1436	0.1179	0.0518	0.0048	0.0046
Imperial IPA	0.0809	0.0031	0.0180	0.0005	0.0179	0.0556	0.1048	0.0089	0.7486	0.0463
Irish Red Ale	0.0572	0.0719	0.1547	0.2530	0.1046	0.0989	0.0961	0.1124	0.0019	0.0215
Saison	0.0227	0.1401	0.1007	0.0143	0.2139	0.0745	0.0786	0.0962	0.0077	0.2034
Witbier	0.0022	0.2791	0.0935	0.0042	0.2300	0.0569	0.1092	0.1464	0.0048	0.0293

Based on the above methodology, we calculated our Classfication Rate (correct classifications) as: 32.86%.

Self-Organizing Map

The Comparison

Conclusion