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 Unsupervised 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

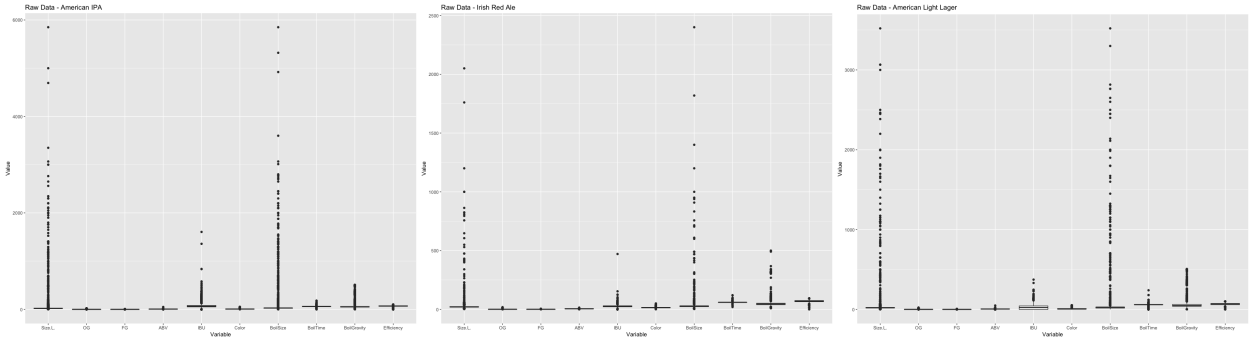
In addition, we decided to investigate the impacts of the following features:

- 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.

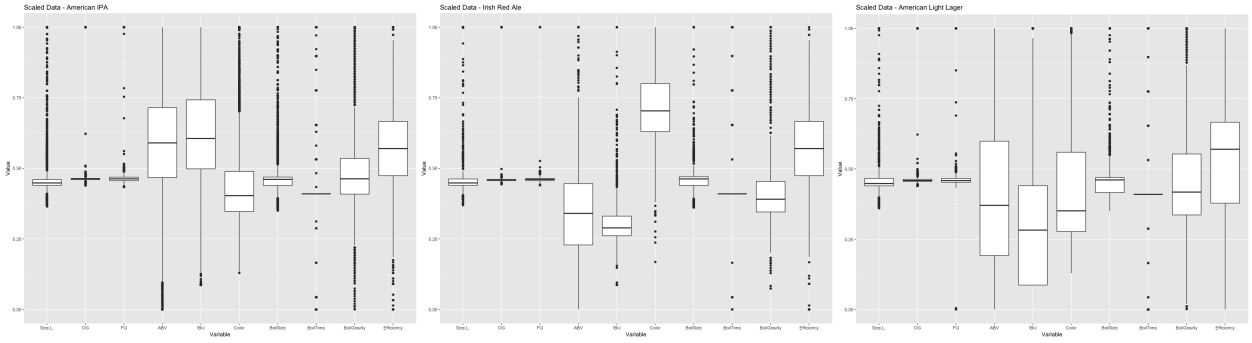
Raw Data Issues

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).

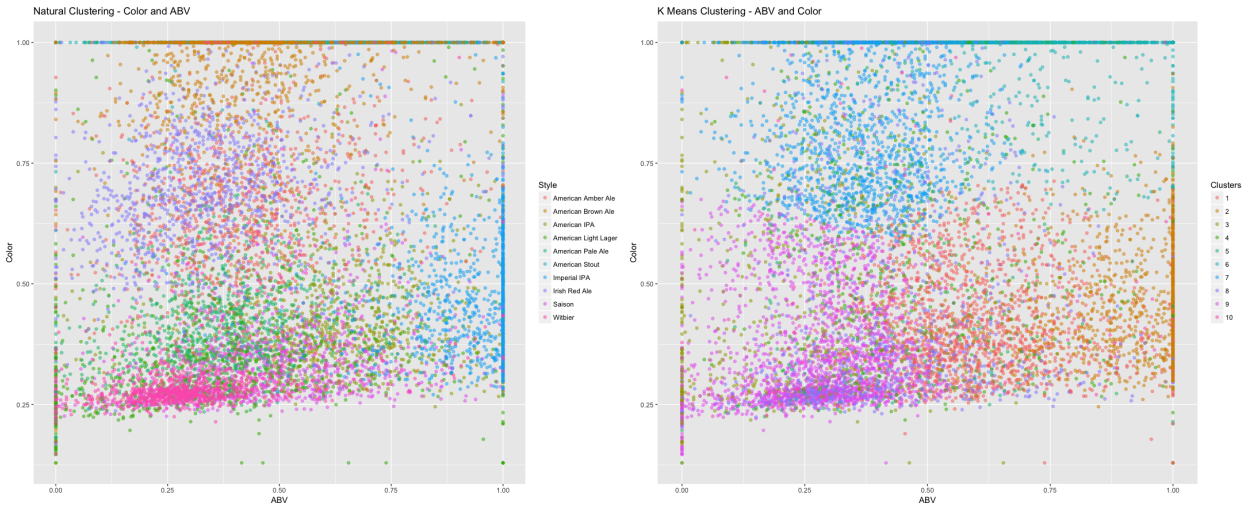


Our Analysis

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	48	44	404	81	93	20	75	165	15	83
American Brown Ale	5	11	575	6	72	20	125	11	11	192
American IPA	441	96	15	56	81	17	6	209	80	27
American Light Lager	110	49	79	151	73	26	32	437	10	61
American Pale Ale	91	36	27	132	101	25	2	603	3	8
American Stout	8	21	279	3	46	22	145	7	17	480
Imperial IPA	465	109	1	3	10	25	3	11	352	49
Irish Red Ale	4	13	532	58	119	21	78	141	3	59
Saison	98	31	26	76	241	21	6	491	11	27
Witbier	16	12	8	132	251	25	5	573	4	2

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	0.0373	0.1043	0.2076	0.1160	0.0856	0.0901	0.1572	0.0623	0.0296	0.0840
American Brown Ale	0.0039	0.0261	0.2955	0.0086	0.0662	0.0901	0.2621	0.0042	0.0217	0.1943
American IPA	0.3429	0.2275	0.0077	0.0802	0.0745	0.0766	0.0126	0.0789	0.1581	0.0273
American Light Lager	0.0855	0.1161	0.0406	0.2163	0.0672	0.1171	0.0671	0.1650	0.0198	0.0617
American Pale Ale	0.0708	0.0853	0.0139	0.1891	0.0929	0.1126	0.0042	0.2277	0.0059	0.0081
American Stout	0.0062	0.0498	0.1434	0.0043	0.0423	0.0991	0.3040	0.0026	0.0336	0.4858
Imperial IPA	0.3616	0.2583	0.0005	0.0043	0.0092	0.1126	0.0063	0.0042	0.6957	0.0496
Irish Red Ale	0.0031	0.0308	0.2734	0.0831	0.1095	0.0946	0.1635	0.0532	0.0059	0.0597
Saison	0.0762	0.0735	0.0134	0.1089	0.2217	0.0946	0.0126	0.1854	0.0217	0.0273
Witbier	0.0124	0.0284	0.0041	0.1891	0.2309	0.1126	0.0105	0.2164	0.0079	0.0020

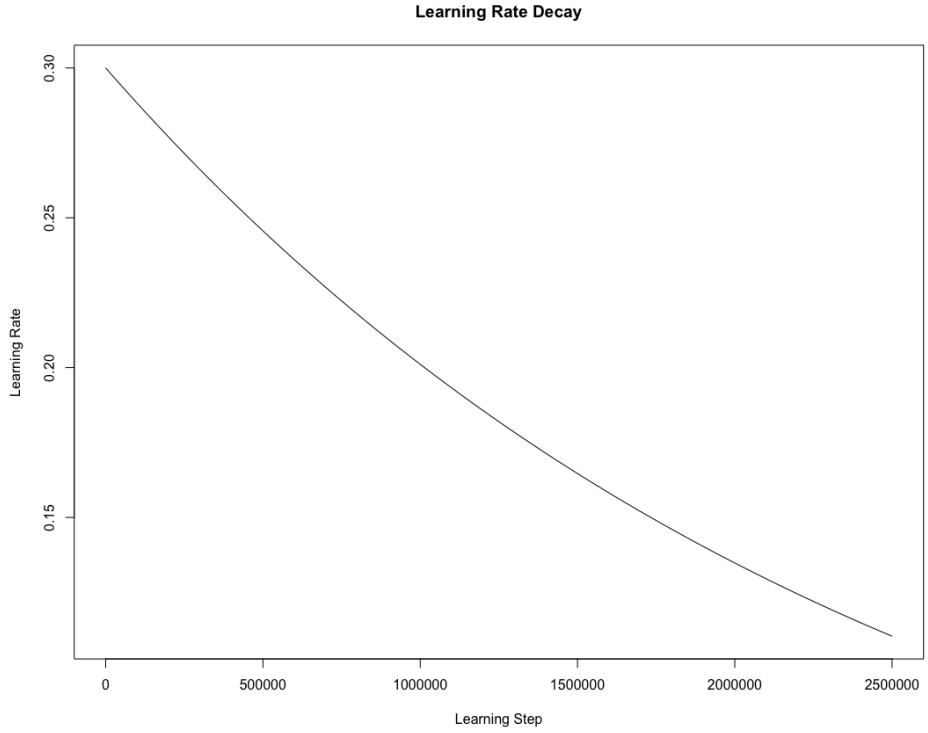
Based on the above methodology, we calculated our Classfication Rate (correct classifications) as: 32.86%.

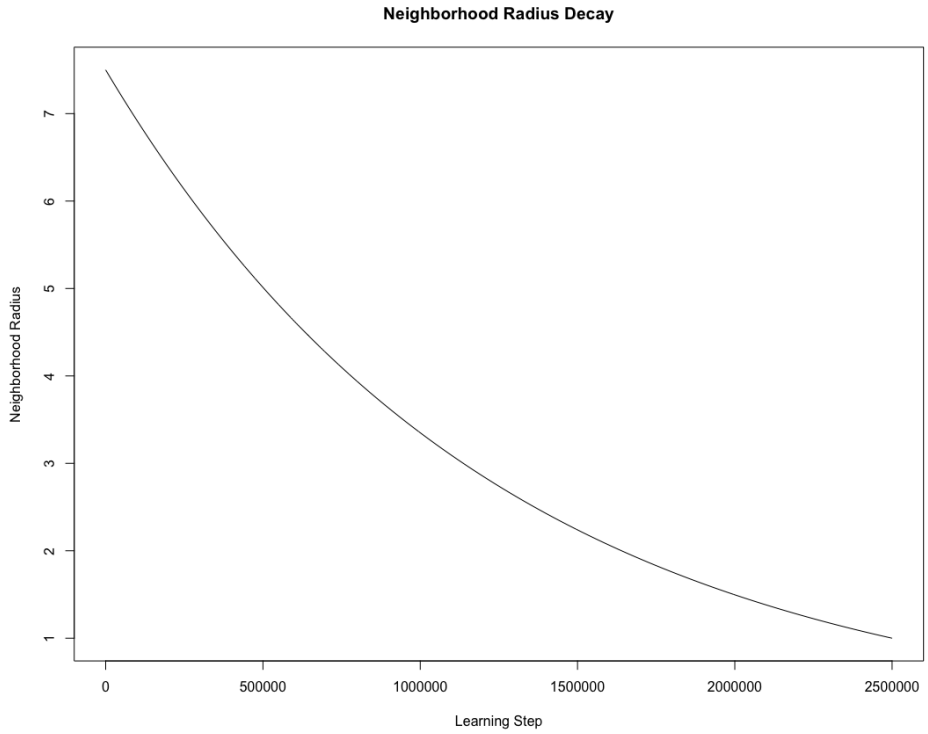
Self-Organizing Map

Network Parameters

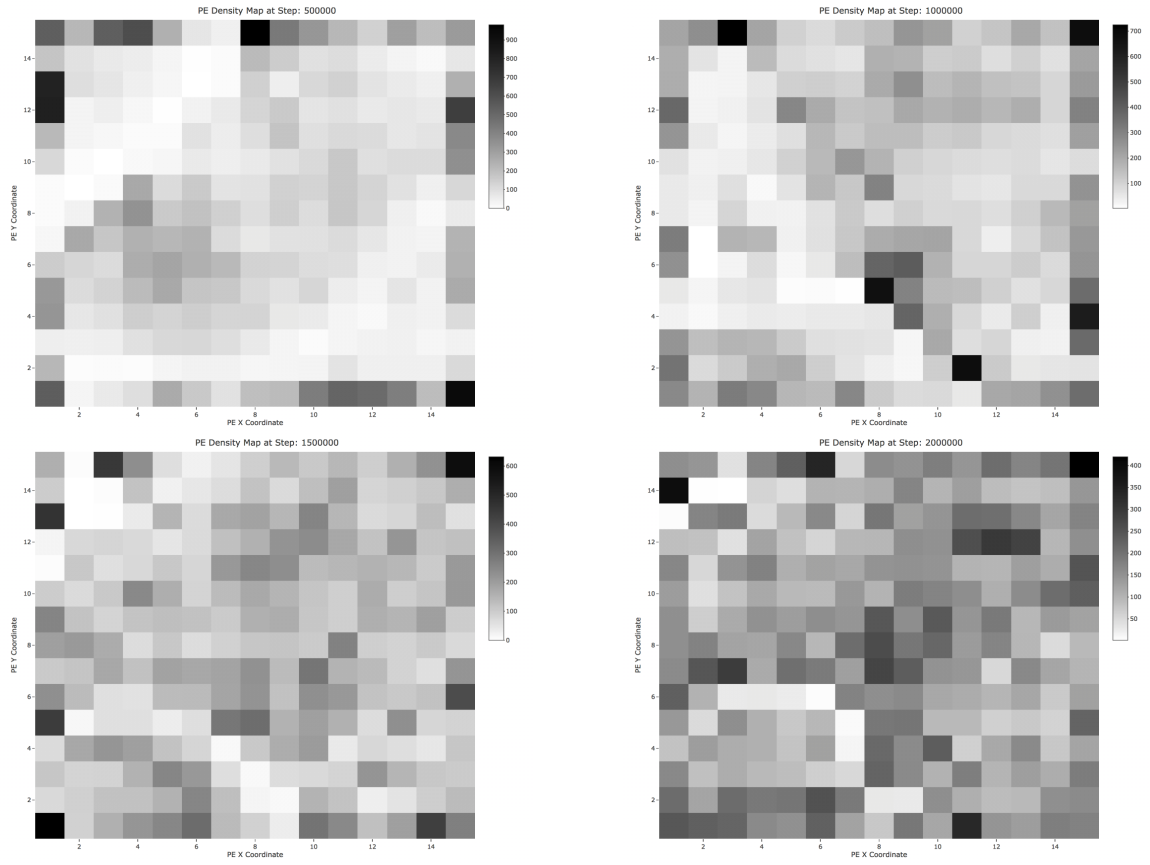
- Network Parameters
  - Topology: 225 PEs (15x15 2-dimensional square lattice)
  - Each PE is a 10-dimensional vector (see below for weight draws)
- Learning Parameters
  - Initial Weights: Drawn from U[-0.5, 0.5]
  - Initial Learning Rate (α_{init}): 0.3
  - Learning Decay Rate: $\alpha_{init} * e^{\frac{-i}{n}}$
    - i = current learning step
    - n = total number of learning steps
  - Initial Radius (θ_{init}): 7.5
  - Radius Decay Rate: $\theta_{init} * e^{\frac{-i}{\mu}}$
    - i = current learning step
    - n = total number of learning steps
    - $\mu = \frac{n}{log(\theta_{init})}$
  - Momentum: None
  - Stopping Criteria: 2,500,000 learning steps
- Input Data
  - As previously described
- Error and Performance Measure
  - Learning Steps Performed: 2,500,000
  - Monitoring Frequency: Every 500,000 Learning Steps

Decay Graphs





Learning Process - PE Density

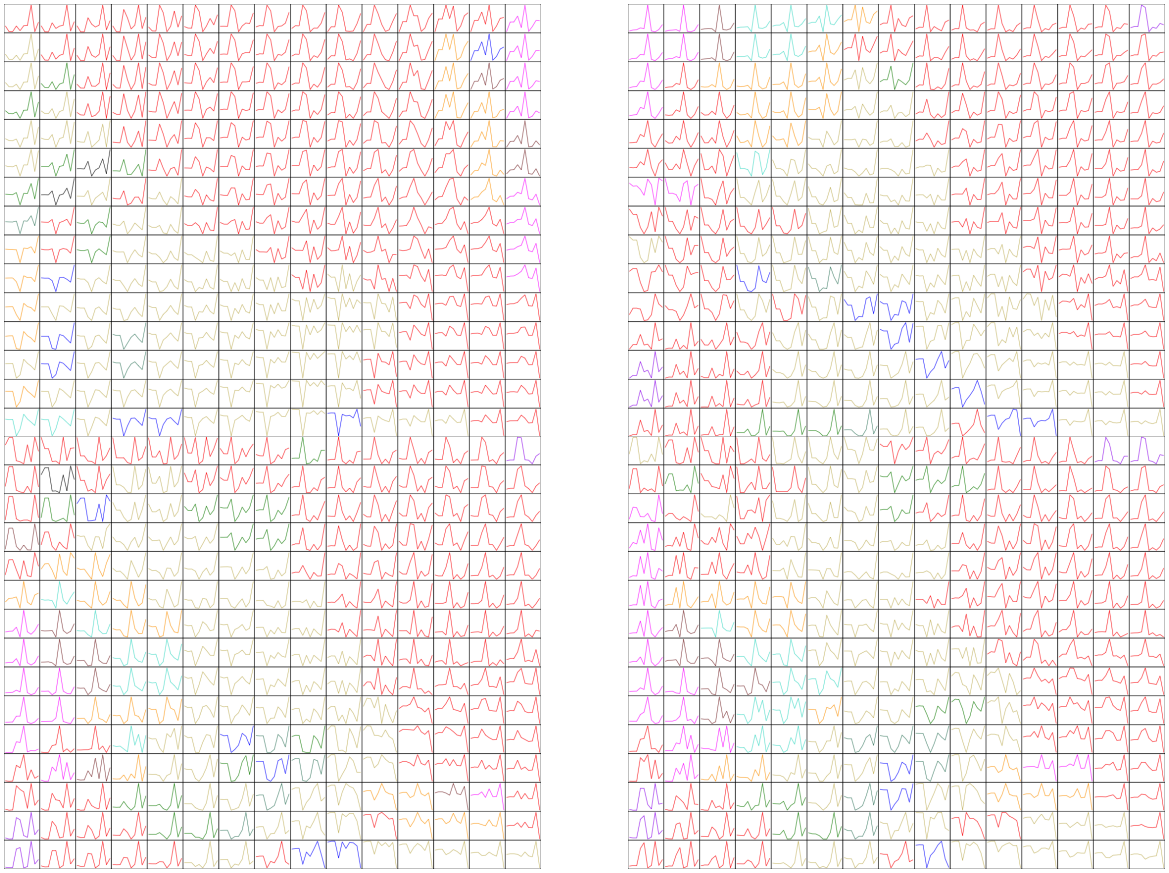


The above matrix of graphs represents the number of inputs that map to each of the PEs in our 15x15 network lattice at several learning steps. The images are arranged as follows:

- Top Left - 500,000 Learning Steps
- Top Right - 1,000,000 Learning Steps
- Bottom Left - 1,500,000 Learning Steps
- Bottom Right - 2,000,000 Learning Steps

As we can see, during the learning process, the number of mapped inputs to each PE increases over time, that is, as learning progresses, the inputs are mapped to a wider range of PEs, instead of just a few at the beginning of the learning process. We can see that at learning step 500,000, most of the inputs are mapped to a select few PEs mostly at the outside of the lattice, while there does appear to be a cluster forming in the bottom left area that is not part of the fringe of the lattice. At the 1,000,000 learning step, much of the inputs that were originally mapped to PEs on the outside of the lattice are now moving towards the center of the lattice; we can see a strong cluster forming near the middle and bottom of the lattice. At the 1,500,000 learning step, we are still seeing a strong favoritism of the inputs to map to PEs on the fringes of the lattice, however, the density in the middle of the lattice is starting to become much more uniform. Finally, at the 2,000,000 learning step, the lattice is almost completely uniform; this is a great result. We will show the final PE density latttice in our final SOM conclusion section; the lattice at the 2,500,000 learning step.

Learning Process - PE Dimensionality



The above matrix of graphs represents the vector of dimensions to each of the PEs in our 15x15 network lattice at several learning steps. That is, each PE is represented by a line that shows each of its dimensions in 10-dimensional space. In addition, each line is colored by the Beer Style for which most the inputs mapped to. So, for example, if most of the inputs that mapped to the lattice at (1, 1) were American IPA, we would color the drawn line red. The images are arranged as follows:

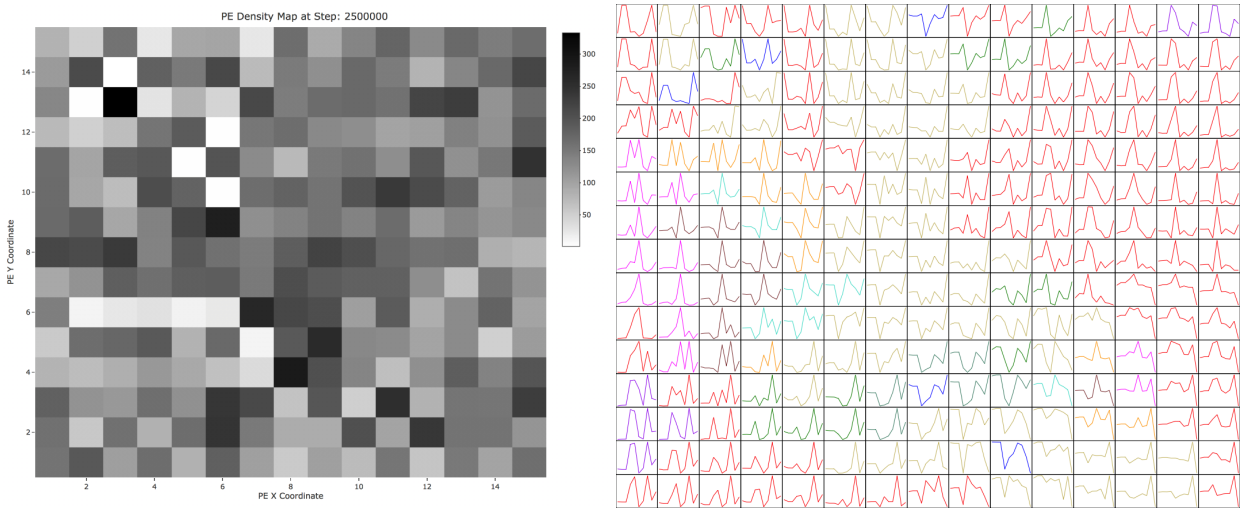
- Top Left - 500,000 Learning Steps
- Top Right - 1,000,000 Learning Steps
- Bottom Left - 1,500,000 Learning Steps
- Bottom Right - 2,000,000 Learning Steps

The colors for the lines are defined as follows according to the Beer Style:

- Red - American IPA
- Sandy - American Pale Ale
- Blue - American Light Lager
- Green - Saison
- Orange - American Amber Ale
- Purple - Imperial IPA
- Magenta - American Stout
- Turquoise - Irish Red Ale
- Dark Red - American Brown Ale
- Dark Turquoise - Witbier
- Black - No PE Mapped

At learning step 500,000 we can see that almost half of the PEs map to American IPA (Red), while another fraction (approximately one in five) map to American Pale Ale (Sandy). We can see some marginally substantial clusters starting to form, such as American Stout (Magenta) in the top left and American Amber Ale (Orange) in the bottom left. As we move from learning step 500,000 to learning step 1,000,000, we can see that American IPA (Red) is beginning to form at least two distinct subclasses in its right cluster, whereas there is no clear definition of its rightmost cluster. The Amber Ale (Orange) cluster has moved closer to the top where it is blending with a cluster formed from Irish Red Ale (Turquoise). A Saison (Green) cluster has formed near the bottom of the lattice trying to differentiate itself from the neighboring American IPA (Red) and American Pale Ale (Sandy). Moving from learning step 1,000,000 to 1,500,000, the structure of our large American IPA (Red) and American Pale Ale (Sandy) clusters remain the same, with each having many subclasses. Meanwhile, American Stout (Magenta) is solidifying its cluster in the west side of the lattice, competing with Irish Red Ale (Turquoise) and American Amber Ale (Orange). There are many distinct American Amber Ale (Orange) clusters forming across the lattice. Finally, moving onto the 2,000,000 learning step from learning step 1,500,000, we are seeing a second American Stout (Magenta) subclass forming that is quite distinct from the first. Otherwise, there weren't many drastic changes. We will show the final vector dimensionality latttice in our final SOM conclusion section; the lattice at the 2,500,000 learning step.

Final Results



As we can see on the PE Density Plot, not much has changed since learning step 2,000,000; however, we are continuing to see the density per PE decreasing as inputs are continuing to be mapped to a wider variety of PEs.

As for the final PE dimensionality plot, we are seeing the following characteristics of our data:

- American IPA (Red) and American Pale Ale (Sandy) are very ill-defined, not only are their winning PEs spread out, but withing the connected clusters we are seeing several, often indistinct subclasses
- American Stout (Magenta) is one of the most well-defined clusters (on the left side of the lattice) with two apparent and distinct subclasses
- American Brown Ale (Dark Red) is also quite well defined, like American Stout (Magenta) it appears to have two reasonably distinct subclasses
- Saison (Green), while not having a true distinct cluster in the lattice space, does appear to have two reasonalby distinct subclasses (ignoring the PEs it won near the top and top-left of the lattice)
- Imperial IPA (Purple) also has a very well-defined cluster at the bottom left of the lattice; interstingly, it appears to have attracted an American IPA (Red) cluster, which makes sense since they are both the same style of beer

Classification Rate

Our methodology for calculating the classification rate it quite simple and compares nicely with our methodology in the K Means case. Essentially, we find the winning Beer Type of each PE based on which Beer Type had the most inputs mapped to each PE. This will allow us to declare a class winner for each PE. Then, we sum all of the inputs that mapped to a PE that was won by their particular class and then divide by all of the inputs; we have the following classification rates at each monitoring frequency:

- Learning Step 500,000: 52.24%
- Learning Step 1,000,000: 53.53%
- Learning Step 1,500,000: 53.98%
- Learning Step 2,000,000: 56.09%
- Learning Step 2,500,000: 55.86%

Conclusion

The Comparison

- Based on the Classification Rate of K Means (32.86%) v. the Classification Rate of our Self-Organizing Map (55.86%) we can confidently say that our SOM has much better classification power than does our K Means approach; however, we will discuss in our Further Investigation section some changes we would recommend to our study design and some further analyses we would be interested in performing.
- In regards to subclass detection, we see that generating the vector dimensionality plots are a great tool for the task; we were able to identify several subclasses within existing tasks using the results of our SOM. On the contrary, K Means was very poor at identifying subclasses; quite the opposite actually, performing K Means made it harder to identify where subclasses may have existed in our data!

Further Investigation

Below are some ways we would consider altering our study and some additional analyses we would consider to augment our project:

- Change Classification Rate: While we are very confident in our result, we would like to consider changing how we calculate our classification rates to be more uniform. As it stands, our classification rates for the K Means approach determined which Beer Style won in each of the 10 clusters, while our SOM approach determined which Beer Style won each of 225 PEs; this approach may have adversely affected K Means (though it just have likely may not have). We would like to consider a more uniform approach to classfication rate so that we can be more confident about our result.
- Conscience SOM: We would also like to work with a Conscience SOM paradigm, particularly for us since our dataset did not have an equal number of inputs per class; we believe that this may have skewed the results and a Conscience SOM may have remedied this issues.
- Increasse Dimensionality: It would be a wise consideration to try increasing our dimensionality to see if we can uncover some other interesting features to help build our map. Of the 23 available features, we chose to use 10 that we were familiar with, perhaps an approach of randomly selecting features based on some probabilty could lead to some interesting findings that we otherwise would have missed.
- Data Pre-Processing: Our data pre-processing approach was done so to scale the data between 0 and 1 and remove outlying values; we may want to consider other approaches that would serve a similar purpose but may perhaps affect our final result.

- SOM Distance Functions: Another possible further investigation would be to experiment with other distance functions when learning our SOM; applying different types of learning functions and seeing how they affect our results would be a great way to refine our output and make our analysis more robust.