Clustering Techniques to Determine Origins of Wine

Mikey Joyce
Department of Electrical Engineering and Computer Science
University of Missouri
Columbia, MO 65201
mpjyky@umsystem.edu

1. Introduction

Due to new technologies that connect the globe, in the modern era, competition is greater in magnitude and severity in basically every single field. Whether that means the individual is attempting to break into the technology and information systems industry, or the journalism field, it is apparent they will need to hone their craft more than ever before to be able to compete with their peers. This is true within the sommelier field, especially since it is a skill that is hard to gain, if a peer has an edge that could be catastrophic for a sommelier. To help a sommelier gain an edge against their peers, it is worthwhile to employ an unsupervised approach to the data. Unsupervised learning, or clustering, is a machine learning (ML) technique that allows a user to extract meaningful patterns, groups, and/or visualizations with data that doesn't have explicit class labels. Because of the nature of unsupervised learning, it provides a good launch pad for knowledge discovery which is why it would be so useful for a sommelier to employ to gain new understanding about the wine which they did not have prior. In this paper, an unsupervised approach to this problem will be discussed at length.

2. The Dataset

The dataset in question was taken from Kaggle and is called, Wine Dataset for Clustering, created by Harry Wang [3]. It is important to note, that at the end of all of the clustering, it was discovered that this dataset actually has a supervised counterpart, which is the same data, except it has the labels [5]. These labels were not known during the clustering process. At the end after the module was developed, these labels were used as a benchmark against the groups that were created with unsupervised learning to produce an accuracy score to observe how this method actually performs on the wine data. The wine data consists of 13 different features. These features represent the chemical makeup of the wines, the flavor profile of the wines, the color of the wines, etc. Out of all of the features there were 12 numerical features and one categorical feature. The categorical feature is labeled as "Proline" and at a first glance it appears to be a numerical feature. It is important to observe that this feature is a representation of an amino acid profile in the wine. And as such, it is a category because there are only so many combinations of the amino acid profile that could be present within the wine. Because of this, the proline feature was dropped from the dataset before clustering. It probably could have been utilized within the clustering and may even have given better results, but it just seemed slightly out of place, so it was decided to remove it from the feature set.

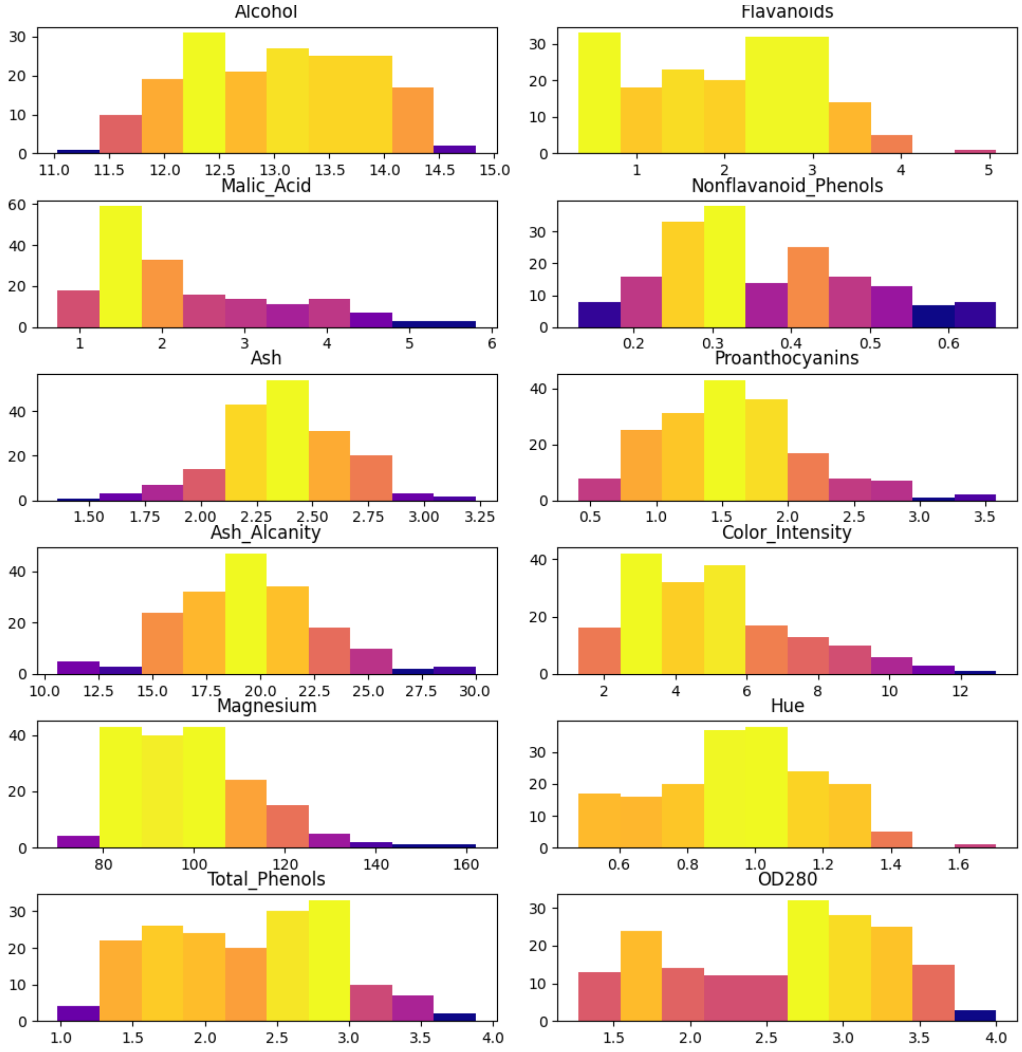

Figure 2.1: Shows the resulting 12 dimensional feature vector and the distribution of data for each feature

3. Unsupervised Learning Methods

In this section, many different unsupervised learning methods will be discussed. The first section will discuss a cluster tendency measure will be explored to determine the structure of the wine data. The second section will discuss the final clustering pipeline that involved four different steps and three different unsupervised learning techniques to gain knowledge from the wine data.

3.1 Cluster Tendency

One major question within the dataset is to determine the number of clusters that exist within the data. An interesting way to do this is to utilize the visual assessment of tendency (VAT) and the informational visual assessment of tendency (iVAT) algorithms [2]. These algorithms analyze the clusters and attempt to provide a visualization to us humans to allow for a greater understanding of the structure of the data. The VAT algorithm is noisier, but it shows more details due to this which allows it to show more locality. Where the iVAT algorithm is less noisy, so it shows more of a global structure of the data. Since they both show meaning, but in different ways, it is important to use both visualizations.

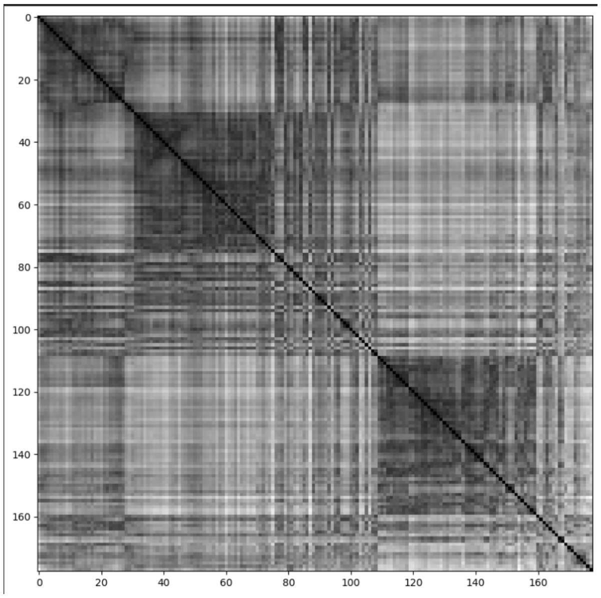

Figure 3.1: Shows the VAT algorithm output

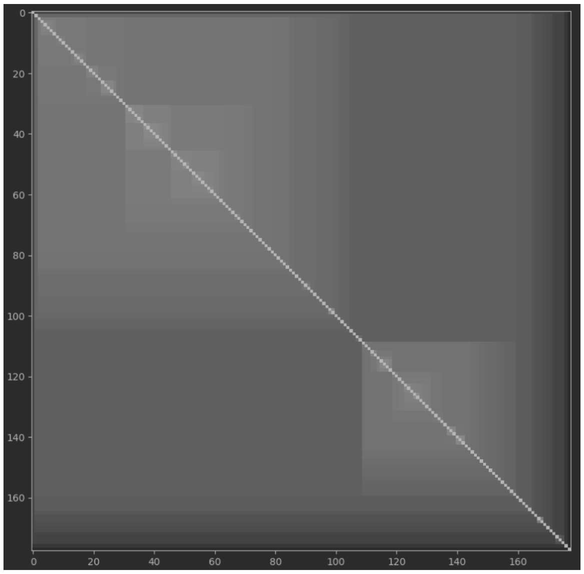

Figure 3.2: Shows the iVAT algorithm output

It is important to interpret the output of these algorithms before developing a technique to model the wine data. In the VAT visualization it appears that there is one large main group and a second group which is smaller. The large main group appears to have two subclusters within it. Looking at the iVAT visualization, it confirms the conjecture from the VAT visualization, that there are two main clusters. So, the conclusion can be made that there are two main clusters, and the larger cluster has two subclusters contained within. This is important because clustering directly as three clusters might not give strong results, instead it is better to cluster as two clusters and then take the larger cluster and cluster over that to obtain three clusters in the end.

3.2 Clustering Pipeline

The clustering techniques that are proposed involves a four step procedure to group the wines. (1) After normalizing the data using a min-max scaler, the first technique that is proposed will be to create a t-SNE embedding [4] that will allow for the 12 dimensional feature set to be mapped in 2 dimensional space, which provides for better visualizations and possibly even clearer clustering. (2) Next, with the 2 dimensional feature space, spectral clustering [1] will be utilized to separate the two main clusters from each other. Spectral clustering is proposed to deal with the imbalance of the size of each cluster and for the fact that t-SNE may produce clusters with wacky shapes that are not consistent. Spectral clustering will be robust to each of those challenges. (3) The third step will involve creating a new t-SNE embedding for only the larger cluster. This will allow for the mapping of the 12 dimensional feature space to a 2 dimensional feature space for the larger cluster. (4) due to the nature of the second t-SNE embedding, these clusters most likely won't be super clear. Because of this, it is important to account for uncertainty within the data. That is why the Fuzzy c-Means clustering algorithm [2] is proposed to break up the larger cluster into two sub clusters. At the end of this proposed pipeline there will be three clusters.

4. Experiments and Results

This section will review the clustering pipeline proposed in 3.2 to see what it looks like in practice and determine the performance of this pipeline.

4.1 t-SNE Global Embedding

Utilizing t-SNE, an embedding will be created to map all of the data into a 2 dimensional feature space. The hyperparameters for the t-SNE embedding are: multi scale perplexities of [5, 8, 15, 20], early exaggeration of 4, distance metric being Euclidean distance, and a PCA initialization before the t-SNE takes hold. With all of this in mind the following map was produced:

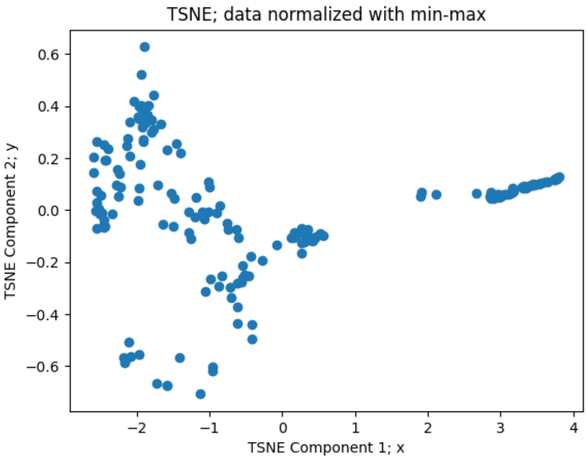

Figure 4.1: Shows the global t-SNE embedding

As one can see, there are two distinct clusters within the t-SNE map that was obtained. Another item to note is this t-SNE embedding that was produced allows for the mapping of new data. So, if this pipeline was put into production, if new data was seen, it would be possible to map that data to this exact pipeline.

4.2 Spectral Clustering

This part of the pipeline utilized spectral clustering to separate the t-SNE global embedding into two different groups.

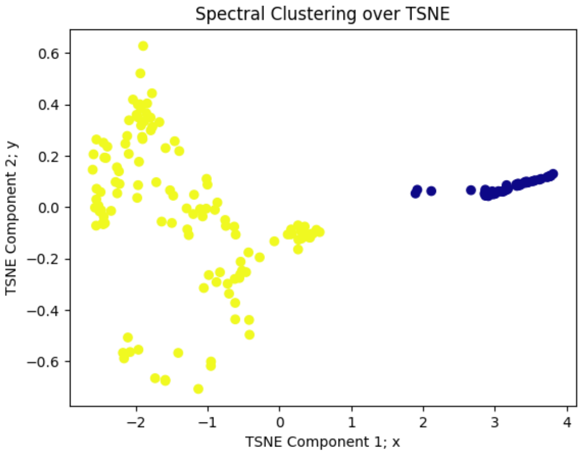

Figure 4.2: Shows spectral clustering over the global t-SNE map

As you can see, spectral clustering produced two clear and direct clusters that met the specification of being two different sizes. The yellow cluster is the target cluster for the rest of the pipeline, while we will save the blue cluster as a hard cluster for the final result.

4.3 t-SNE Local Embedding

Now, using the larger cluster, t-SNE will be utilized again to map only those data points into their own 2 dimensional plane. This is called local embedding because only the neighborhood of the larger cluster is being reviewed during this process. The hyperparameters of this t-SNE embedding are a bit different from the first. This time there is no multi scale perplexity, but instead has a single perplexity value of 20. The same distance metric, Euclidean distance, is utilized. Lastly, there is not an early exaggeration in this embedding, and it has the same PCA initialization as the first embedding.

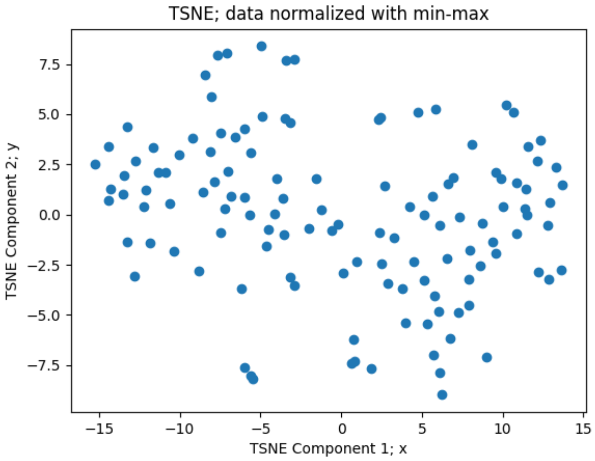

Figure 4.3: Shows the local t-SNE embedding that was obtained

4.4 Fuzzy c-Means Clustering (FCM)

Since the local t-SNE embedding that was obtained is not as clear and crisp as the first t-SNE, uncertainty must be accounted for. That is the main reason why the Fuzzy c-Means algorithm is employed because this algorithm measures uncertainty by allowing each data point to have a certain membership to each cluster. The higher the membership value, the more certain the algorithm is that the data point belongs to that cluster. Since it was observed in the VAT algorithm that the larger algorithm has two subclusters, the number of clusters for this algorithm will be two, just like in spectral clustering.

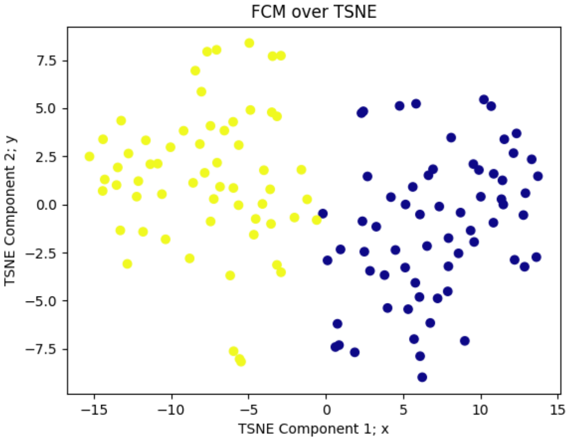

Figure 4.4: Shows the FCM algorithm run over the local t-SNE embedding

4.5 Final Result

Now that the final clustering has occurred, the final clustering plot can be visualized. Since the supervised counterpart to this dataset was found on Kaggle, it is also worth it to benchmark my results against the ground truth to see if this method was actually viable.

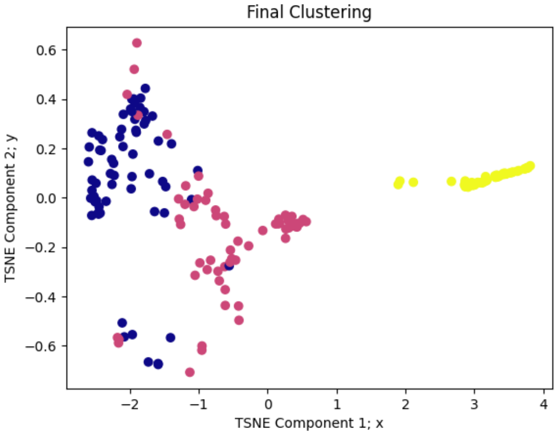

Figure 4.5: Shows the final clustering result visualized on the global t-SNE embedding

The clusters that we found are not exactly intuitive based on this plot, the pink and blue clusters seem to have a strange relationship with each other that could not have been discovered without doing the local t-SNE embedding. Using the supervised labels for this dataset, the accuracy score was calculated to be: **0.9269662921348315, or 92-93%.**

5. Conclusion and Future Work

Based on the accuracy score that was obtained and the cluster tendency measure that was utilized at the beginning, it is clear that this method is able to identify the origins of the wine strongly. In the future, it may be worth it to account for the proline feature instead of throwing it away. This might be able to provide a higher accuracy because there is more data for the unsupervised methods to find patterns in. It may also be worth it to test this method with unexpected data that may have come from cultivars that are not present in this dataset, just to see how it reacts. All-in-all, this project was successful in clustering the origins of wines that come from three separate cultivars.

References

[1] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," in Advances in Neural Information Processing Systems, vol. 14, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2001. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2001/file/801272ee79cfde7fa5960571fee36b9b-Paper.pdf.

[2] D. B. Fogel, D. Liu, and J. M. Keller, "Fuzzy Clustering and Classification," in Fundamentals of Computational Intelligence, John Wiley \& Sons, Ltd, 2016, pp. 147-182, ISBN: 9781119214403, DOI: 10.1002/9781119214403.ch8, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119214403.ch8.

[3] H. Wang, "Wine dataset for clustering," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering, 2023.

[4] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

[5] T. Elmetwally, "Wine dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/tawfikelmetwally/wine-dataset, 2023.