James Taylor

Assignment 5

University of Maryland University College

DATA630-9041 Spring 2020

Dr. Gates

james.taylor.ps3@gmail.com

**Introduction**

Escherichia coli (E. coli) is type of bacteria that exists in most people’s gut biome. While many types of it are harmless, some types cause intestinal infections. This is commonly the source of the term ‘food sickness’. While it is possible to ingest it on undercooked meat, the most common sources are ones that aren’t cooked like lettuce and spinach. This means here is no opportunity for a restaurant to kill the bacteria effectively.

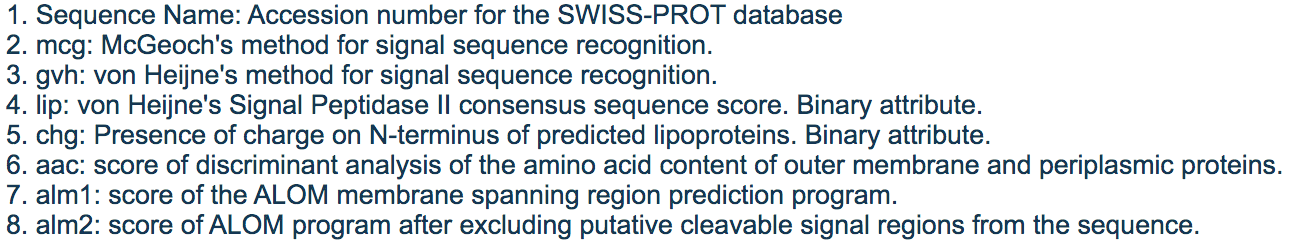
Food related illnesses can have disastrous consequences for restaurants. The restaurant Jack in the Box had 500 illnesses and three children die in 1992 from a meat-borne E. coli outbreak. This lead to then President Clinton to enforce random testing in ground beef. In 2006, at least 199 cases and three deaths across 26 states were attributed to an E. coli outbreak in spinach, with 31 people developing a dangerous complication that can cause kidney failure (CNN, 2019).

It is therefore vital for food providers to detect these particularly dangerous bacteria. This paper will look at the ability for an unsupervised, k-means clustering algorithm to distinguish between types of E. coli. This will firstly see if the different E. coli types are discernable on the provided variables in the dataset. This paper will also look into which variables are going to be useful separating the bacteria types. The number of clusters is another parameter that needs to be determined; there are eight bacteria types but this may not be the optimal amount.

**Analysis and Model Development**

About Dataset

The data’s source was Kenta Nakai at the Institute of Molecular and Cellular Biology at Osaka University. Each observation is a type of E. coli along with attributes of it. There are 336 observations on the eight variables. Below is the list of features in the dataset.



Preprocessing

There were no missing values in the dataset. To create the model, the class (E. coli type) wasn’t to be included. The variables were also normalized to aide in the calculations to optimize the clusters.

Exploratory Data Analysis

 One of the features that may help in separating the classes is McGeoch’s method for signal sequence. The first thing of note is there are only four imL and imS observations in the dataset. The ‘cp’ and ‘im’ groups look reasonably distinguishable just from this mcg feature. Other than 3 ‘pp’ observations there are nearly zero observations below .5 mcg value that aren’t either ‘cp’ or ‘im’.

The Von Heijne signal sequence method shows different distributions, but it is attempting to measure the same thing as the above figure. With this method ‘cp’ and ‘im’ don’t nearly have the range of values as McGeoch’s method. The other four types of E. coli have broader, more distinguishable distributions with this new measure.



The ‘lip’ feature will likely not be very useful for the unsupervised clustering. It is a binary value that only had 10 positive values across all 336 observations. There is a chance that it will help the model distinguish between E. coli types but with only 10 positive values it will likely not have a large effect. This is similar to the ‘chg’ feature, but in that case there is only one positive value.

The metrics related to ALOM show promise for helping to decipher between the types of E. coli. Ideally every different type would have a distribution with a small range that does not overlap with any other E. coli type. For example, types ‘cp’ and ‘im’ have little overlap in their distributions on the ‘alm1’ value. This helps because the data points will be closer together, allowing for a centroid to capture all of them easily without incidentally including other E. coli types. The model will not perform well when types have very similar distributions, like ‘im’ and ‘imU’ were they both have a mean of about 0.75 alm1.

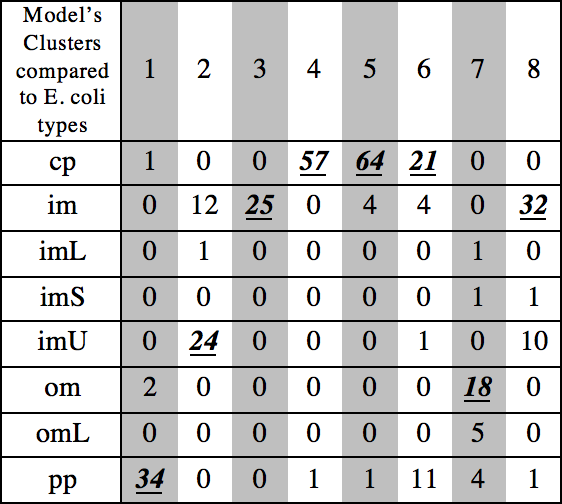


Model

The model used was a K-means cluster algorithm. This was from the ‘cluster’ package (version 2.1.0) in R. It will group observations into a specified number of non-overlapping groups based on the observation’s position the dimensional space, or simply values of its features. The algorithm used the Hartigan-Wong approach, which aims to minimize the within-cluster sum of squares of errors. The algorithm creates centroids to which the closest observations are then assigned to. The algorithm finds the mean (centroid) of those assigned data points and uses that as the centroid for the next iteration. This occurs until no observation switching from one cluster to another reduces the within-cluster SSE, at which point the iterations end. The model was trained on variables that were normalized. For each model, the E. coli type for the observation was removed. Because these models are unsupervised, they are not needed to train but can be used to check the model’s clustering ability.

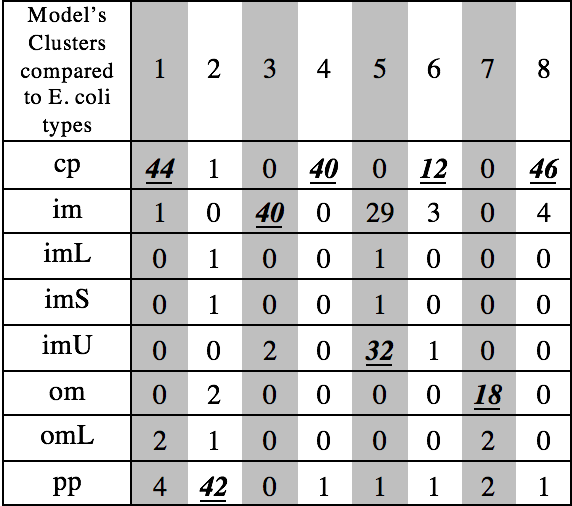
**Results and Model Evaluation**

The first unsupervised model used all available variables (excluding the E. coli type). The model was specified to find 8 clusters because this is the amount of E. coli types provided in the dataset. Below is a table showing the groups the model made against the E. coli type of the observations. The bolded and underlined number represents for that cluster what the dominant E. coli type is, or what this cluster is predicting.

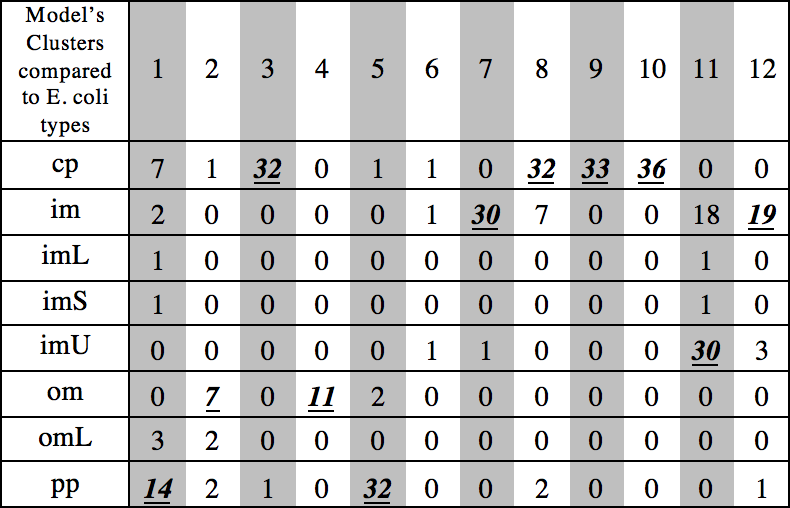
The purity of this model was 81% (this is a measure of accuracy- sum the largest values in each cluster then divide by number of observations). The ideal outcome would have every outcome only contain observations from one type of E. coli. This would mean no incorrect categorization would be possible. For example, cluster four below has 57 type ‘cp’ and one type ‘pp’. It obviously makes sense to deem observations in cluster four as ‘cp’, but this would incorrectly mark the one ‘pp’ E. coli type as ‘cp’. The issue persists in the first cluster, since it’s dominant E. coli type was ‘pp’ but there are also two type ‘om’ in the cluster.

The E. coli types imL, imS and omL have no cluster were they are the dominant type. That means this model would never give a prediction for those E. coli types. That being said, those types only make nine observations, a mere 2.6% of the dataset. Attempting to get those last 2.6% in accuracy may mean overfitting the model.

New Model

 One of the downsides to this k-means clustering approach is with weights of the features. The model treats each feature as equally important in clustering the data. Features that don’t help distinguish between the E. coli types are then making the decision boundaries of the model not optimal. The two obvious variables that may be detrimental to the model’s accuracy was ‘lip’ and ‘chg’, which were both binary values with 10 and one positive values across the data set respectively. The table below is the results of a new model with those two variables removed which had a purity of 81%. This reduced model had just about the same purity of the previous model but with E. coli type ‘cp’ is the dominant type in four clusters instead of three. This model was least accurate at isolating the ‘im’ E. coli. There was only one cluster where it was the dominant type, leaving 48% of that type in other clusters where they aren’t the dominant type. Types ‘imL’, ‘imS’ and ‘omL’ remained unable to be dominant in any clusters.

Looking at the E. coli type ‘cp’ above, it was the dominant type in half of the clusters. There may be lots of variation in that type, or subgroups within that single type that are significantly different from each other. The next model will have 12 clusters instead of 8 while excluding the ‘lip’ and ‘chg’ features. The results below show this model’s ability to group bacteria types.



The purity of the above expanded model was 82%, a slight improvement over the other models. The bacteria type ‘cp’ was the dominant type in four clusters, meaning there are too many clusters or lots of variation in that type. Cluster six has no dominant type which may be evidence of too many clusters.

There are more clusters than types of E. coli, but this isn’t by itself a negative. The within-cluster SSE is lower, the between-cluster SSE is higher and the purity is higher. It simply requires that a type is categorized by more than one cluster. Although it may be unintuitive that there are more clusters than real categories it isn’t a difficult thing to grasp if this model was used by another person not too familiar with it.

The tradeoff is overfitting the data. This model was trained on a data sample, so new/unseen data may be different. Take for example cluster two where the dominant type only had seven observations out of 12. This cluster may be stochastic noise, or a cluster that exist in this training data that may not exist in unseen data. The lower the purity or number of observations of cluster, the more aware the researcher should be of overfitting. Below is the plot of the observations and clusters. They overlap significantly, but without it the purity of the clusters would be significantly lower.



**Conclusion**

There is more than one way to measure the accuracy, but all the models had the same accuracy in predicting the correct bacteria type (around 81%). The different models, with their different parameters, returned different accuracy levels on each bacteria type. This is important because not every type I created equal. Some are harmless and others can kill.

A model that has the best ability discovering dangerous types when they exist should be used. A model that takes low risk prediction to this respect (predict more observations to be a dangerous type than are actually present), this would lead to food producers to throw out more food than necessary. There is a clear dichotomy between having a risk-averse model and only disposing of food that is dangerous. Where a particular company chooses to fall on this trade-off can have consequences in the millions of dollars, both in lawsuits and efficiency.

Efforts should be put towards a more intuitive model. The models in this paper used a dataset that had complex features like von Heijne's method for signal sequence recognition and score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins. These seem complicated for food workers in chain restaurants to be dealing with because, ideally, this detection model will be cheap and easy enough for every restaurant to test their food. Attain the requisite features for the model should not need a degree in microbiology.

References

CNN. E Coli. Outbreaks Fast Facts. July 2019. Retrieved from <https://www.cnn.com/2013/06/28/health/e-coli-outbreaks-fast-facts/index.html>