Clustering and Distance Methods

Justin Hood

University of Wisconsin-Stout

hoodj5402@uwstout.edu

May 6, 2019

Outline

Background

- Over the course of studying statistics at any level, it becomes natural to consider data that is categorical in nature.

Background

- Over the course of studying statistics at any level, it becomes natural to consider data that is categorical in nature.
- Oftentimes multivariate data exhibits a natrual grouping or clustering behavior that is challenging to analytically describe.

Background

- Over the course of studying statistics at any level, it becomes natural to consider data that is categorical in nature.
- Oftentimes multivariate data exhibits a natrual grouping or clustering behavior that is challenging to analytically describe.
- The goal of clustering analysis is to find the natural groupings that exist within the data without requiring assumptions by the statistician.

Cluster vs. Classification

- Classification is another tool that statisticians can use to group data structures together.

Cluster vs. Classification

- Classification is another tool that statisticians can use to group data structures together.
- It requires that the statistician choose a number of groupings before analysis. This can be problematic on large data sets as well as on data whose shape is not well known before.

Cluster vs. Classification

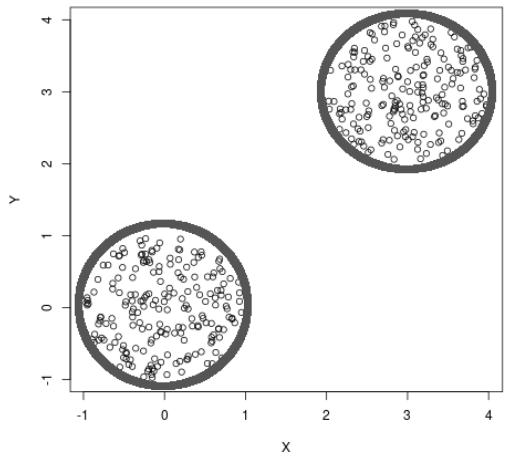
- Classification is another tool that statisticians can use to group data structures together.
- It requires that the statistician choose a number of groupings before analysis. This can be problematic on large data sets as well as on data whose shape is not well known before.
- Clustering is a more base and general process that allows the data to naturally sift into groupings based on the user defined "distance". We will focus on this method for this project.

What is Distance?

- Topologically speaking, we can think about clustering as wrapping n-dimensional hyperspheres around the data in n-dimensional space, where our goal is to find the most natural centers and radii that cover the data.

What is Distance?

- Topologically speaking, we can think about clustering as wrapping n-dimensional hyperspheres around the data in n-dimensional space, where our goal is to find the most natural centers and radii that cover the data.
  For a 2-D example, consider the image below,



- We see that these two clusters of points have natural circular boundaries.

Distance Continued

- For high dimensions and categorical data, distance is harder to visualize, but the rough idea is the same.

Distance Continued

- For high dimensions and categorical data, distance is harder to visualize, but the rough idea is the same.
- The problem is defining a meaningful metric to compare points.

Common Measures of Distance

- Euclidean distance: $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
- Minkowski Metric (Taxicab distance), $d(x, y) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}$
- Canberra metric $d(x, y) = \sum_i \frac{|x_i - y_i|}{x_i + y_i}$

Common Measures of Distance

- Euclidean distance: $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
- Minkowski Metric (Taxicab distance), $d(x, y) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}$
- Canberra metric $d(x, y) = \sum_i \frac{|x_i - y_i|}{x_i + y_i}$
- Other examples include hamming distance and the "max metric"

Similarity Coefficient

- For characteristic data, it is often helpful to introduce a binary variable for the purposes of comparison.

Similarity Coefficient

- For characteristic data, it is often helpful to introduce a binary variable for the purposes of comparison.
  For example, consider the data below,

	1	2	3	4	5
i	1	0	0	1	1
k	1	1	0	1	0

Similarity Coefficient

- For characteristic data, it is often helpful to introduce a binary variable for the purposes of comparison.
  For example, consider the data below,

	1	2	3	4	5
i	1	0	0	1	1
k	1	1	0	1	0

  We may then write,

$$d(i,k) = \begin{cases} 0 & i = k \\ 1 & i \neq k \end{cases}$$

Similarity Coefficient

- For characteristic data, it is often helpful to introduce a binary variable for the purposes of comparison.
  For example, consider the data below,

	1	2	3	4	5
i	1	0	0	1	1
k	1	1	0	1	0

  We may then write,

$$d(i,k) = \begin{cases} 0 & i = k \\ 1 & i \neq k \end{cases}$$

  So, d is a measurement of "difference" in the two variables

Similarity Coefficient Continued

In this way, we can write the comparisons in the table,

	$k = 1$	$k = 0$	Totals
$i = 1$	a	b	$a + b$
$i = 0$	c	d	$c + d$
Totals	$a + c$	$b + d$	$a + b + c + d$

Similarity Coefficient Continued

In this way, we can write the comparisons in the table,

	$k=1$	$k=0$	Totals
$i=1$	a	b	$a+b$
$i=0$	c	d	$c+d$
Totals	$a+c$	$b+d$	$a+b+c+d$

From this table, we can define many similarity coefficients depending on our desired measurement. These coefficients are then collected into matrices, upon which we can perform our more familiar statistical operations.

Methods

- Now that we have considered how we might compare our abstract data, let us now consider how we will go about sorting it.

Methods

- Now that we have considered how we might compare our abstract data, let us now consider how we will go about sorting it.
- We will first consider hierarchical clustering methods, which have two types,

Methods

- Now that we have considered how we might compare our abstract data, let us now consider how we will go about sorting it.
- We will first consider hierarchical clustering methods, which have two types,
  - Agglomerative (Start with n clusters and combine)
  - Divisive (Start with 1 cluster and partition)

Agglomerative Method Steps

For an agglomerative method, the basic steps are,

1. Start with n clusters, and an $n \times n$ matrix of distance values.

Agglomerative Method Steps

For an agglomerative method, the basic steps are,

1. Start with n clusters, and an $n \times n$ matrix of distance values.
2. Find the entry or entries with the lowest distance value.

Agglomerative Method Steps

For an agglomerative method, the basic steps are,

1. Start with n clusters, and an $n \times n$ matrix of distance values.
2. Find the entry or entries with the lowest distance value.
3. Merge these pairs of clusters, and update the matrix by removing the old entries and recalculating distances relative to the new merged groups.

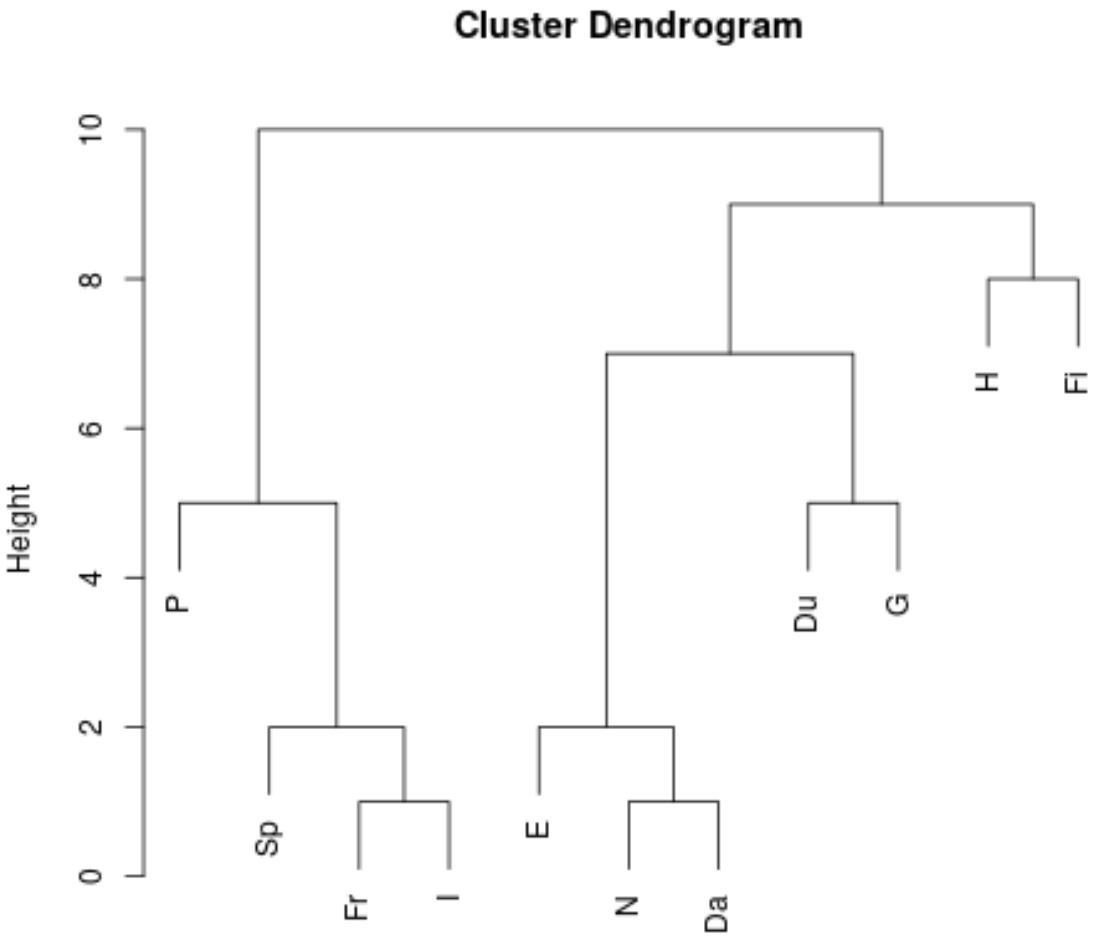
Agglomerative Method Steps

For an agglomerative method, the basic steps are,

1. Start with n clusters, and an $n \times n$ matrix of distance values.
2. Find the entry or entries with the lowest distance value.
3. Merge these pairs of clusters, and update the matrix by removing the old entries and recalculating distances relative to the new merged groups.
4. Rinse and repeat $n - 1$ times. Keep a record of each merge as it happens for plotting purposes.

Dendrogram Example

Using the agglomerative method on Concordance data from several different languages, we arrive at,



This result shows that French and Itailian are closely related, and that Danish and Norwegian are as well. To the French and Italian, we see that Spanish is also closely related, and so on.

Hierarchical Methods

When we compute the steps of these agglomerative methods, we must consider the different ways we can join the subgroups.

- Single Linkage - Find the smallest values in the distance matrix and join those groups together
- Complete Linkage - Much the same as the former, however our updated matrix uses the maximum distance between the subgroups when updating
- Average Linkage - Updating the matrix uses the average of all possible distances from the original grouping.
- Wards Method - Seeks to minimize ESS. Treats each cluster as an n-sphere and computes the distance to the centroid. Each succesive merger results in the smallest possible ESS based on these centers and groupings.

Hierarchical Continued

- As we see, there are many different ways to compute the iterative steps of these hierarchical methods.

Hierarchical Continued

- As we see, there are many different ways to compute the iterative steps of these hierarchical methods.
- Due to the nature of the methods, we have no statistically sound way of identifying and cleaning outliers from the data, which can result in bad clustering.

Hierarchical Continued

- As we see, there are many different ways to compute the iterative steps of these hierarchical methods.
- Due to the nature of the methods, we have no statistically sound way of identifying and cleaning outliers from the data, which can result in bad clustering.
- There is also no way to change a points assignment once it has been merged into a new cluster. As such, points can be grouped incorrectly.

Hierarchical Continued

- As we see, there are many different ways to compute the iterative steps of these hierarchical methods.
- Due to the nature of the methods, we have no statistically sound way of identifying and cleaning outliers from the data, which can result in bad clustering.
- There is also no way to change a points assignment once it has been merged into a new cluster. As such, points can be grouped incorrectly.
- Using multiple different methods and comparing the results is often the best way to decide if your results are stable and correct.

Nonhierarchical Methods

- We saw before that Hierarchical methods are designed to group variables into groups.
- We now consider methods that group items into an arbitrary number of groups by recursively updating.
- This is commonly done by choosing "seed" points at random within the space of the data.

K-means Method

We now consider a popular nonhierarchical method, K-means. The steps of this method are,

K-means Method

We now consider a popular nonhierarchical method, K-means. The steps of this method are,

1. Partition the points into K initial clusters

K-means Method

We now consider a popular nonhierarchical method, K-means. The steps of this method are,

1. Partition the points into K initial clusters
2. Iteratively compute the distances to the centroids of each cluster for each data point, and move the point to a better cluster if necessary.

K-means Method

We now consider a popular nonhierarchical method, K-means. The steps of this method are,

1. Partition the points into K initial clusters
2. Iteratively compute the distances to the centroids of each cluster for each data point, and move the point to a better cluster if necessary.
3. Update the affected centroids

K-means Method

We now consider a popular nonhierarchical method, K-means. The steps of this method are,

1. Partition the points into K initial clusters
2. Iteratively compute the distances to the centroids of each cluster for each data point, and move the point to a better cluster if necessary.
3. Update the affected centroids
4. Rinse and Repeat until each point is in the best possible centroid for itself

K-means Example

Consider the data,

Item	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

To perform a K-means clustering, let us at random choose an initial partition of (AB) and (CD). We find the centroid of each of these to be,

Cluster	x_1	x_2
(AB)	$\frac{5-1}{2} = 2$	$\frac{3+1}{2} = 2$
(CD)	$\frac{1-3}{2} = -1$	$\frac{-2-2}{2} = -2$

So, we have our centroids, $(2, 2)$ and $(-1, -2)$. Next, we compute the distances between each point and the centroids,

$$d(A, (AB)) = \sqrt{10}$$
$$d(A, (CD)) = \sqrt{61}$$
$$d(B, (AB)) = \sqrt{10}$$
$$d(B, (CD)) = \sqrt{9}$$

Here, we see that B is actually closer to the centroid of (CD), and as such, we shall move it, and update the centroids. Continuing,

$$d(C, (A)) = \sqrt{41}$$
$$d(C, (BCD)) = \sqrt{5}$$
$$d(D, (A)) = \sqrt{89}$$
$$d(D, (BCD)) = \sqrt{5}$$

K-means Continued

- We see now that none of the points need to be moved, and as such, we are done with our calculations.
- The resulting groups (A) and (BCD) make sense upon inspection, as BCD all lie at least partly within the negative ranges around the origin, while A sits alone in quadrant I with a ways between it and the other points.
- As before, it is recommended to re-run the algorithm with different initial conditions, to test stability.

NonHierarchical Method Notes

- It is important to make sure that our initial conditions do not inadvertently overlap.

NonHierarchical Method Notes

- It is important to make sure that our initial conditions do not inadvertently overlap.
- Outliers will tend to isolate themselves into their own groups.

NonHierarchical Method Notes

- It is important to make sure that our initial conditions do not inadvertently overlap.
- Outliers will tend to isolate themselves into their own groups.
- Even if you know the number of groups a population has, if the sampling method does not have points in each cluster, a forced number of clusters might have bad results.

Clustering Based on Statistical Models

- The methods that we have considered before have all been fairly intuitive in terms of the derivation of the steps, but there exist more statistically model based methods for analyzing the data.
- For example, consider the model where each cluster k has an expected proportion p_k of the points, and the associated pdf $f_k(x)$. Then, for k clusters,

$$f_{Mix}(x) = \sum p_k f_k(x)$$

- As one might expect, a common mixture model is a multivariate normal distribution.

Further Analysis of the Normal Model

- As one might expect, this normal distribution model will yield results close to the K-means and Ward's method processes.
- Choosing the number of clusters for an arbitrary data set can be challenging, but the formula

$$\#Clusters \propto -2\ln(L_{max}) - Penalty$$

  where L_{max} is defined as the product of our mixture functions.
- Under the AIC the penalty can be written as,

$$-2\ln(L_{max}) - 2N(\frac{k}{2}(p+1)(p+2) - 1)$$

Multidimensional Scaling

- As one might expect, a natural consequence of looking at these high dimensional abstract data sets will be to attempt to project them into a lower dimensional form for graphical interpretation.
- This, however will not always be a simple procedure, as information is lost through the projection into lower dimensions, as well as the fact that trying to encode categorical data always results in loss of precision as well.
- As such, we might consider a process similar to PCA to attempt to look at the data while retaining as much information as possible.
- Multidimensional scaling looks at a slightly different problem than PCA, however, as the relative distances between the points should be retained when possible, as this can provide interesting analytical results on high dimensional data.

Multidimensional Scaling Continued

- Looking at N observations, there are $M = N(N-1)/2$ similarities between points. Assuming the data is fine enough that equal values are not possible, it is trivially possible to order these distances,

$$s_1 < s_2 < \ldots < s_M$$

  As such we would expect that the associated distances to follow the form,

$$d_1^q > d_2^q > \ldots > d_M^q$$

  Where q is the desired dimensionality of the projection.
- For cases when a the transformation from s to d is not as simple, a computation of the Stress is used to align the data accordingly.
- An interesting consequence of this analysis is that on data about physical distances, the resultant plot will artificially start to resemble a map without the data ever being placed in that shape.

Correspondence Analysis

- Another way of graphically representing complex higher dimensional data is Corresponednce analysis.
- This method is used for frequency data.
- By representing relative frequency in the form of stacked bar charts, both the categories and the levels can be compared simultaneously.

Correspondence Analysis Numerics

- To