Defining Dialect Groups with Hierarchical Agglomerative Clustering

Sarah Darrow - sd1209
CSE 4990: Algorithms for AI
May 3, 2017

I. Introduction

For my final project, I chose to explore hierarchical clustering as we did not get a chance to look at unsupervised learning in depth in class, and I wanted to expand my breadth of knowledge of learning algorithms. I implemented this algorithm on data describing the dialects in the United States because the data clearly illustrates the benefits of hierarchical clustering due to the dialect clusters having a significant and interesting internal structure. The relationship between the measured dialects of two or more states is valuable information, as is the emergence of larger clusters. I implemented two different approaches to this clustering algorithm and compared their results on the dialect dataset.

II. Overview of Hierarchical Clustering

Hierarchical clustering differs from other clustering algorithms as it provides a structured view of the clusters. Whereas traditional K-Means Clustering divides a dataset into k clusters where each data point belongs to only one cluster, Hierarchical Clustering organizes the data into a hierarchy of clusters, with each data point belonging to the main cluster containing all the data as well as possibly belonging to one or more sub-clusters. This hierarchy is traditionally shown in a dendrogram, or a tree structure where branch height is proportional to the distance between clusters. [1]

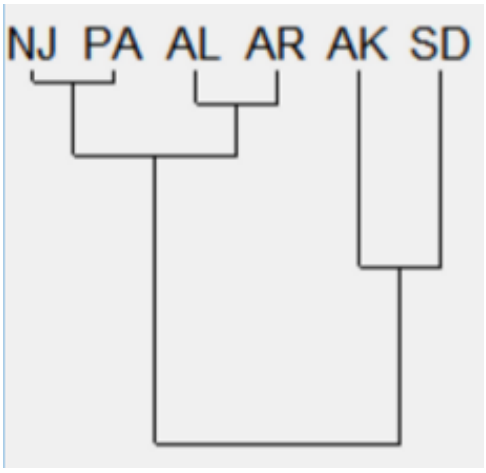

Fig. 1. An example of a dendrogram

A. Approaches to Hierarchical Clustering

The two main approaches to Hierarchical Clustering are agglomerative and divisive clustering.

Divisive, or top down, clustering begins with all of the instances of data belonging to one cluster, then separating instances or groups of clusters to create sub-clusters. This process is repeated until all of the data instances are in a sub-cluster containing only themselves. This is commonly accomplished by creating a minimal spanning tree of the instances with the edges having weights of the distance between the elements. Sub-clusters are created by removing edges with the largest weights until there are no edges left. [1]

Agglomerative, or bottom up, clustering is the more popular version of Hierarchical Clustering. begins with all of the instances of data in a clusters containing only one instance. These clusters are combined until all of the sub-clusters are merged into one. This approach requires the distances between the clusters to be recalculated when clusters are combined. There are multiple ways to do this, including centroid, group average, single linkage, and complete linkage. [2] [3]

- *Centroid* clustering calculates a centroid, or midpoint, for each cluster and defines the distance of the clusters by the distance between their centroids [4]
- *Group Average* clustering defines the distance between two clusters as the average of the distances between each instance in one cluster with each instance in the other cluster [5]
- *Single Linkage* clustering defines the distance between two clusters as the minimum distance between any instance in one cluster and any instance in the other cluster
- *Complete Linkage* clustering defines the distance between two clusters as the maximum distance between any instance in one cluster and any instance in the other cluster

III. Implementation on Dialect Dataset

The goal of this exercise was to use Hierarchical Clustering to determine the relationship of dialects in each of the states. As linguists have not been able to determine strict definitions of dialects in the United States, a hierarchical representation is appropriate. Dividing the states into a set number of clusters is possible, but not sufficient enough to describe the complex definitions of the dialects. A hierarchical approach better illustrates the groups and subgroups of dialects in the United States.

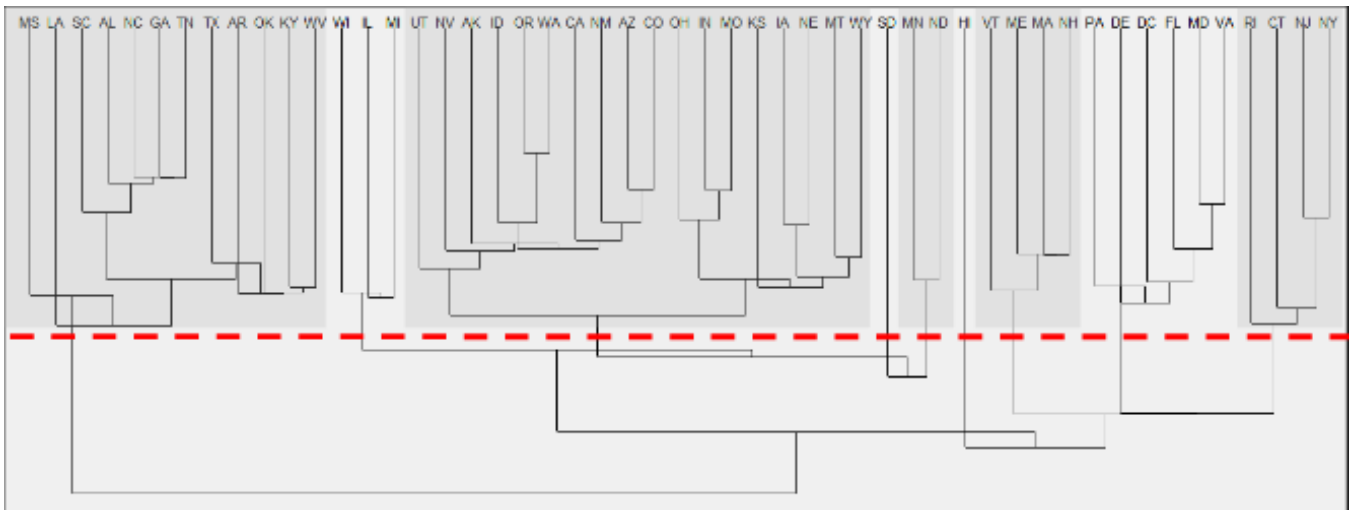

Fig. 2. The dendrogram found with Agglomerative Centroid Hierarchical Clustering

A. Harvard 2003 Dialect Survey

The dataset I chose to cluster is a survey performed by Bert Vaux and Scott Golder of the Harvard Computer Society [6]. They collected data from 30,934 participants from all 50 states and the District of Columbia. This data consists of the participants answers to questions about their pronunciation and word choice on language that is used different regionally. The survey included 122 questions about multiple aspects of language, including:

- *Pronunciation:* How do you pronounce caramel?
- *Emphasis:* THANKSgiving or thanksGIVing?
- *Word Choice:* What do you call the miniature lobster found in lakes and streams? (crawfish/crayfish/crawdad/etc)
- *Word Meaning:* What is the difference between dinner and supper?

This data was previously used to construct maps that represent the answers to each individual question, as well as to predict where a survey participant is from.

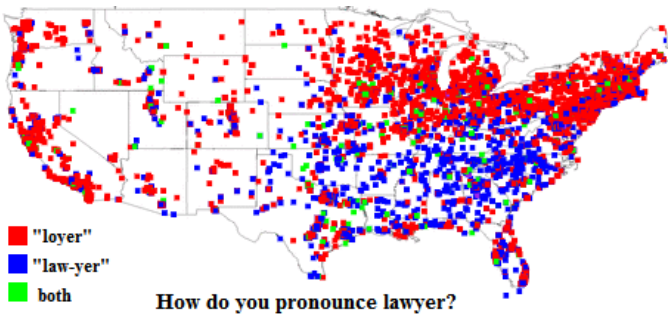

Fig. 3. A map generated by the Harvard Computer Society illustrating the regional differences in pronouncing the word "lawyer"

B. Agglomerative Clustering on Dialect Data

As the agglomerative approach is more widespread, I implemented this method of hierarchical clustering.

The dataset consisted of the percentage of each state that chose each answer of each question. To determine the distance between the data from each state, I calculated the root squared error between each answer on each question, then summed the error on each question as the distance between the states on all questions. I chose to sum differences between individual questions because the questions did not have an equal number of answers, and I wanted to somewhat equalize the influence of each question.

To determine the distance between clusters of more than one state, I implemented both the centroid and group average methods.

IV. Results

I generated a dendrogram of the states' dialects, and used the dendrogram to divide states into 9 clusters. The dendrogram provides more detailed information on the relationship between the states' dialects, and the defined clusters provides a method to create a simpler representation of a possible definition of dialects.

A. Agglomerative Centroid Hierarchical Clustering Results

The centroid cluster distance method provided a dendrogram that could be abstracted to 9 clusters, generally clusters for the West, North Central, South Dakota, Great Lakes, South, Mid-North East, North East, and New England.

Centroid clustering is not monotonic, so inversions can occur causing lines on the dendrogram to cross [4]. This is because sometimes including an instance of data in a cluster can make the cluster 'closer' to another cluster than it was before the merge. Some think of this property as a problem of the centroid distance approach, but it does provide information about the relationships in the data. For example, an inversion occurs when adding Mississippi to the South cluster after

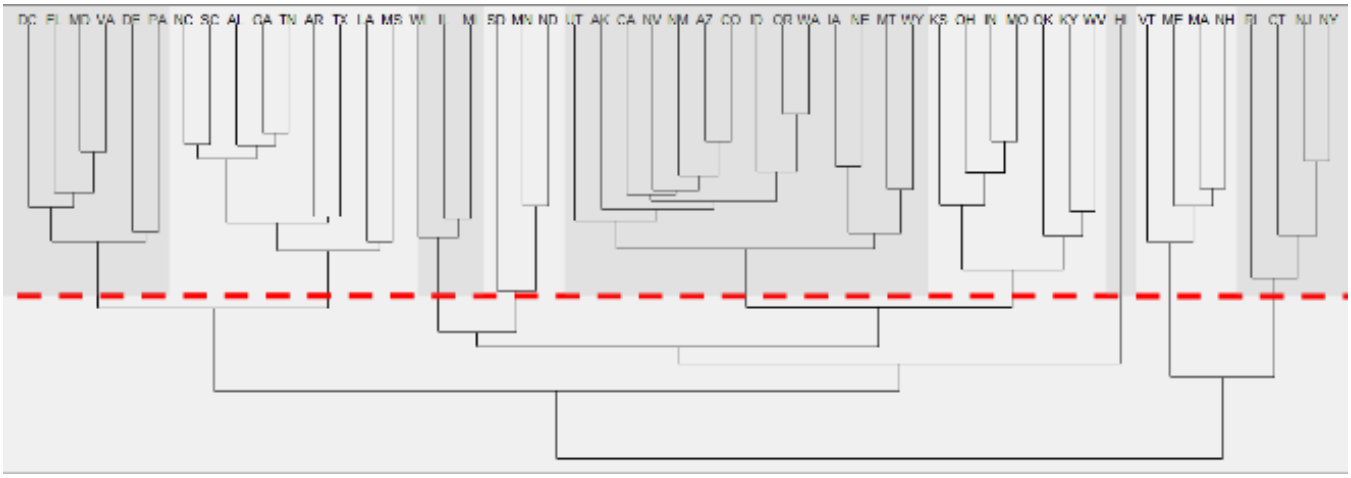

Fig. 4. The dendrogram found with Group Average Centroid Hierarchical Clustering

adding Louisiana. This indicates that the dialect of Mississippi is between the dialects of Louisiana and the rest of the South.

B. Agglomerative Group Average Hierarchical Clustering Results

The Group Average method of determining the distance between clusters resulted in a similar 9 clusters being discovered, with the only difference being the emergence of a Mid-America cluster and South Dakota being merged into the North-Central cluster. South Dakota's merge, however, still has a distance very close to the maximum distance used to create the 9 clusters.

Group Average Clustering is monotonic, so the inversions present in the Centroid method cannot occur [2]. Most of the large very similar clusters (like the North Carolina/South Carolina/Georgia/Tennessee/Alabama and New Mexico/Arizona/Colorodo clusters) stayed exactly the same. A main difference between the dendrograms resulting from the Centroid and Group Average clusterings is that the Group Average clustering had more two-state sub-clusters (16 in the Group Average dendrogram vs 12 in the Centroid dendrogram).

The Group Average Clustering was also faster than the Centroid clustering, probably due to not having to calculate the centroids. Both methods have a time complexity of $O(n^2 logn)$ in relation to the number of states, or generally instances, being clustered, but the Centroid Clustering has an extra factor of the number of questions, or generally the dimension of the data.

V. Conclusion

Agglomerative Hierarchical Clustering provided a detailed representation of the relationship between states' dialects. The Centroid and Group Average methods of finding the difference between clusters provided similar results, but had different nuances in the structures of their dendrograms.

References

[1] https://www.cs.utah.edu/ piyush/teaching/4-10-print.pdf.
[2] https://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html.
[3] https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html.
[4] https://nlp.stanford.edu/IR-book/html/htmledition/centroid-clustering-1.html.
[5] http://www.saedsayad.com/clustering_hierarchical.htm.
[6] https://www4.uwm.edu/FLL/linguistics/dialect/staticmaps/states.html.

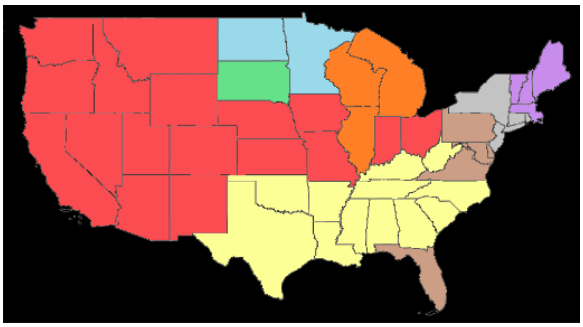

Fig. 5. The 9 clusters found with Agglomerative Centroid Hierarchical Clustering

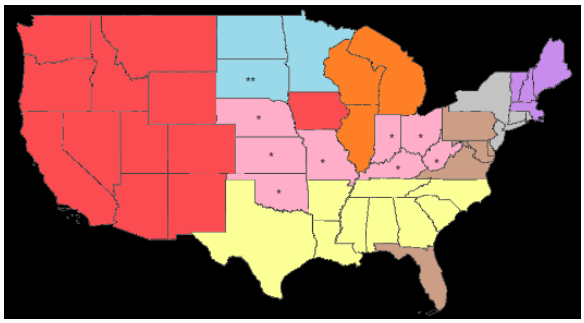

Fig. 6. The 9 clusters found with Group Average Centroid Hierarchical Clustering. Asterisks indicate differences from Centroid Clustering