Visualization of Census Data of Brazilians in America

Hui Li
Department of Computer Science
Boston University
huili93@bu.edu

Haoyu Zhang
Department of Computer Science
Boston University
haoyuz@bu.edu

Junhe Chen
Department of Computer Science
Boston University
junhec@bu.edu

Abstract

Digaai is a digital platform that aggregates and curates the cultural production of the Brazilian diaspora through text, audio, video and images that document the diversity of Brazilian immigrant communities around the world. The content of the platform is obtained both through the spontaneous contribution of participants and the via material collected by volunteers and collaborators. The platform provides statistical data towards Brazilians in United States. The statistic data demographically shows the number of Brazilian for each state. For this project, we are required to visualize the census data collected by the Digaai.
Furthermore, we not only visualize the demographic information for each state, we also perform a clustering and build a classification model to predict the state of an individual.

1 Problem Definition

As more and more Brazilians choose to live in United States, data of Brazilians in United States also also grows rapidly. With demographic state population of Brazilians, it is convenient for people to better understand the population distribution and its structure by visualizing the data on map.

In this project, our basic goal is to visualizing Brazilians in United States. But to make it useful in real life, we also approach to do the followings:

1. Understanding Brazilians population of states by clustering states based on location and population.
2. Predicting an individual's location according to demographic information with a Naive Bayes Model

Hence our visualizing part will include the result from the features mentioned above.

2 Related works

2.1 Data Visualization

Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning "information that has

been abstracted in some schematic form, including attributes or variables for the units of information".

A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Data visualization is both an art and a science. It is viewed as a branch of descriptive statistics by some, but also as a grounded theory development tool by others. Increased amounts of data created by Internet activity and an expanding number of sensors in the environment are referred to as "big data" or Internet of things. Processing, analyzing and communicating this data present ethical and analytical challenges for data visualization. The field of data science and practitioners called data scientists help address this challenge.

3 Data Related

In this part, we are talking about the data we used in this project and some of the processing on the data.

3.1 Data Description

Since this is an external project, here we have used the data given by the cooperator Digaai, it was a census data collected. In the census data, the statics are displayed according to the states. There are populations of Brazilians in each state, and the number of Brazilians with some specific attribute in each state. The data is relatively small but it take times to do some preprocessing.

3.2 Data Processing

Here we will list some process we did in our project, some of which may be mentioned later in other parts.

1. Eliminating some redundant rows and columns:
   Because the census data is more likely aiming to be read by person, there are some rows and columns are redundant for the project, and the headers are missing too. So First of all, we cleaned those rows and columns we don't need and use the proper header we will use later.

2. Change the number of Brazilians into possibilities:
   Since we will later use Naive Bayes to predict where an individual might show up, we tried to re-calculate the percentage of people in every states in order to have the possibility of those attributes. And meanwhile we found some attributes have a total possibility(percentage) over 1.0, that is, this process can ensure that later while doing Naive Bayes, every attribute can have a total possibility 1.0.

3. Separate attributes according to states and rename some attributes:
   Since while doing the visualization of the data as in the requirement, we are using Java Scrip, and while displaying features of each state, it is hard to read those attributes from the original form .csv file, we separated attributes of each state into a single csv file, hence we can continue visualizing. And by the way in this step, we have used regular expression to represent and alter some attributes' name because they are quite coding unfriendly. For example, we use '_' instead of the space and take over signs like '<' with plain English explanation, and eliminated some line changing characters in the data values. Besides, we

have once been stucked because we didn't realized the data was in windows form and we need to add a ''^{9} in the regular expression.

4 Methodology

In this part, we mainly describe the methodology we used in this project. It contains three parts: Hierarchical Clustering, Predicting with Naive Bayes and Visualizing.

4.1 Hierarchical Clustering

In data mining and statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. Here is an example of Hierarchical Clustering using distance matrix:

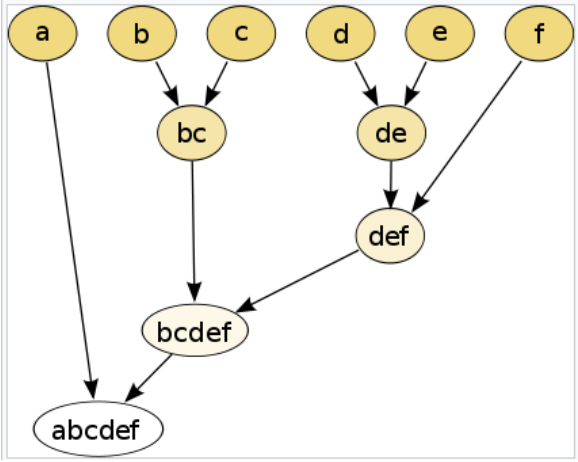


Figure 1: Example of Hierarchical Clustering

4.2 Naive Bayes

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

The basic Bayes theorem is:

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \tag{1}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature x_i is conditionally independent of every other feature x_j for $j \neq i$, given the category C. This means that:

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k) \tag{2}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is:

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{3}$$

where the evidence Z is a scaling factor dependent only on $x_1, \ldots, x_n$, but we haven't use it in our model.

4.3 D3.js

D3.js (or just D3 for Data-Driven Documents) is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG, HTML5, and CSS standards. It is the successor to the earlier Protovis framework. In contrast to many other libraries, D3.js allows great control over the final visual result.

D3.js is used on hundreds of thousands of websites. Some popular uses include creating interactive graphics for online news websites, information dashboards for viewing data, and producing maps from GIS map making data. In addition, the exportable nature of SVG enables graphics created by D3 to be used in print publications.

Embedded within an HTML webpage, the JavaScript D3.js library uses pre-built JavaScript functions to select elements, create SVG objects, style them, or add transitions, dynamic effects or tooltips to them. These objects can also be widely styled using CSS. Large datasets can be easily bound to SVG objects using simple D3.js functions to generate rich text/graphic charts and diagrams. The data can be in various formats, most commonly JSON, comma-separated values (CSV) or geoJSON, but, if required, JavaScript functions can be written to read other data formats.

5 Implementation

5.1 Clustering

In order to analyze population distribution, we apply hierarchical clustering to the dataset with location information and population. We try to discover adjacent states whose population are very similar within the group. By doing this, all states are divided into several groups, and it is more obvious to analyze the residence of Brazilians in United States rather than on the individual state level.

Original data :
It is obvious that the scale of 3 attributes are different which may affect the clustering quality. Thus,

Altitude	Longitude	Population
43.452492	-71.563896	2371.0

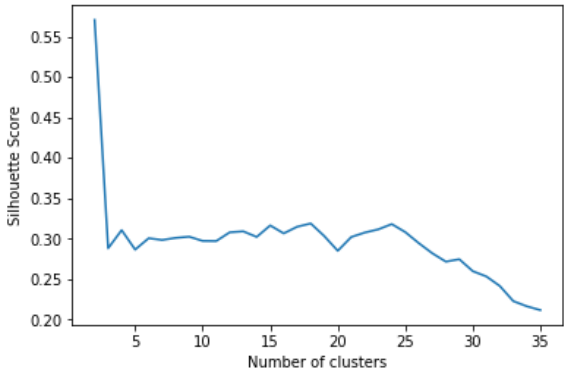
Figure 2: Getting the number of clusters

we scale population first by preprocessing function provided by sklearn.
Then, we perform a hierarchical clustering, still with the hierarchical function provided by sklearn.

Altitude	Longitude	Population
6.63533713e-01	1.14018416e+00	-3.30309750e-01

Figure 3: Getting the number of clusters

To find a best clustering number, we evaluate clustering from 5 to 35, and find 18 is the best clustering number. The metric to evaluate the clustering results is the Silhouette Score.




```
The best numbers of hierarchical clusters: 18
```

Figure 4: Getting the number of clusters

The clustering result is shown as below, each line stands for a cluster:

1. Massachusetts
2. Florida
3. California
4. Michigan, Wisconsin, Minnesota
5. Wyoming, Iowa, South Dakota, Nebraska
6. Montana, North Dakota
7. Oregon, Washington, Idaho
8. Alaska
9. New Mexico, Arizona
10. Utah, Colorado, Nevada
11. Missouri, Oklahoma, Arkansas, Kansas
12. Louisiana, Mississippi, Alabama
13. North Carolina, Georgia, Tennessee, South Carolina
14. Texas
15. Connecticut, New Jersey, New York
16. Rhode Island, New Hampshire, Vermont, Maine
17. Delaware, Virginia, Kentucky, Pennsylvania, Indiana, Maryland, West Virginia, District of Columbia, Illinois, Ohio
18. Hawaii

The visualized result is as below:
The result basically satisfies our clustering goal: divide states into groups according to their similarity in location and population. If they are adjacent and have similar population, we divide them into the same group. There are some individual states consists of a group on its own. It means these states are not like their neighbors in population. For example, California itself is a group, and it is to say that California has a population that is distinguished from its neighbors. Oppositely, states in the middle area form groups with their neighbors since the population are very similar for these state, and it may be a result of the fact that Brazilians are less willing to live in these area compared to more economic developed area such as Massachusetts and California.

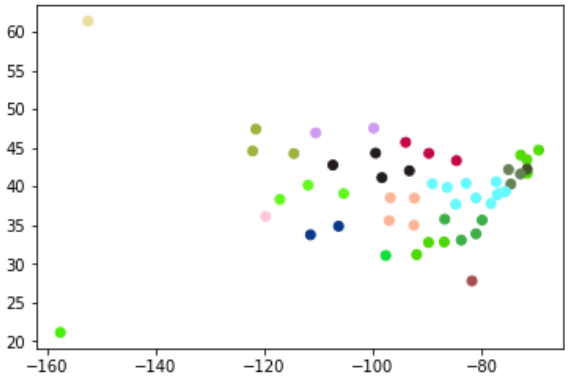


Figure 5: Clustering result

5.2 Predicting

To bring our project with some use in real life, we add the function that predicting where an individual might show up. In oder to do that, we applied a simple Naive Bayes to the data.
The general procedures are as below:

1. As states in the data description, we have re-calculate the data and change the data from 'shown-in-numbers' into 'shown-in-possibilities'.

2	nan	nan	nan	Alabama	Alaska	Arizona
3	Population	Est.	362,398	835	318	2,696

Figure 6: Data with numbers

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Alabama	Alaska	Ar
0	Population	Est.	362398	0.002304096600974619	0.0008774882863591963	0.

Figure 7: Data with possibilities

2. We had an input reading file written in python, which stores every input into a list.
3. Then according to the inputs, we can allocate those attributes in the possibility version data, hence we can calculate the overall possibilities that the individual might appears in each state and out put them as a result list.
4. Finally, we can get the max value from the list and return the name of the state according to the index of the max value in the result list.

The result basically satisfies our approach of this method, the reason why we choose to use the Naive Bayes is that the data is a census data so it is relatively small, and it is near impossible for us to find and scrap data containing personal information, so we applied this simple but useful method to do the prediction work.

```
[1.0125207735549334e-15,
 0.0,
 2.073766481248225e-11,
 4.005164446365018e-15,
 4.617395956502173e-06,
 6.39879058491171e-11,
 1.0083605681951121e-07,
 6.5918464738360774e-15,
 1.5683192818382063e-13,
 0.0002912756878982186,
 3.2332348693377676e-08,
```

Figure 8: Part of the Result List

5.3 Visualizing

In this part, we have used the data processed in previous part. And the default output is kind of a heat map of America on the population attribute of each state. And we have separated the attributes in the data into several groups. Then we can plot the charts of a group of attributes in a state while the cursor of the mouse is moved onto the area of the state in the browser. Here is an example of what it looks like and this is showing the attribute group 'Age Distribution'.

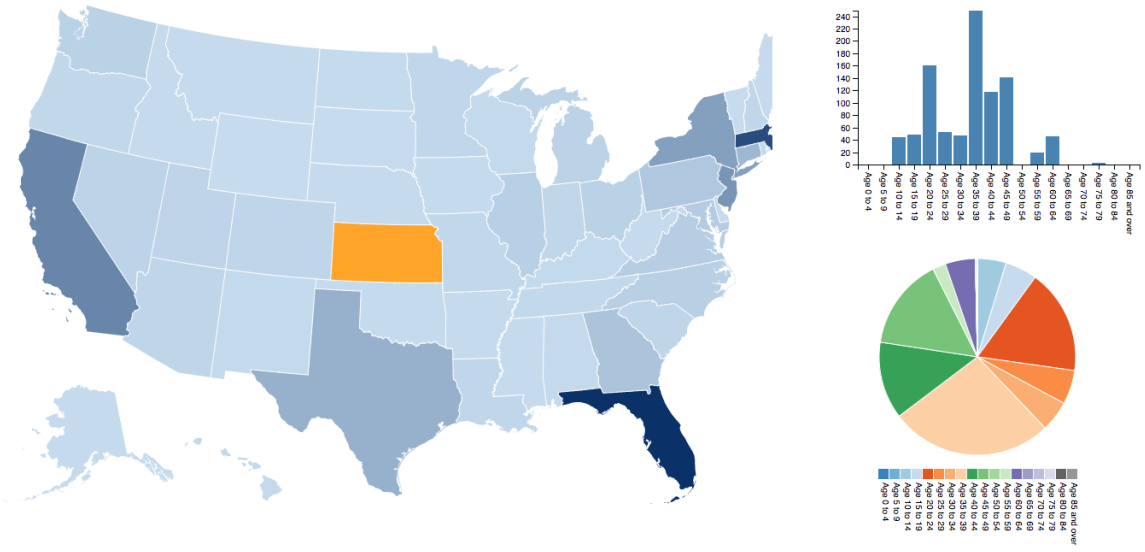


Figure 9: Example of Visualization

6 Conclusion and Future Works

6.1 Conclusion

As mentioned before, we have three goals in this project: clustering, classification and visualization. Basically, we have completed all of them.

1. For clustering, the areas consists of similar states are successfully detected. It is obvious to conclude that main states the most Brazilians live in are California, Massachusetts, and

Texas. The conclusion is reasonable considering these states are more economics developed than neighbor areas.

2. For prediction, due to the lack of individual data, we are not able to evaluate the prediction accuracy. Theoretically, it is able to perform prediction roughly.
   We think we could do better if we have more data like some data of individual information, then we can use some other models for the prediction because finally we can train them and pick the best result from various methods.
3. For data visualization, the visualized data shows the distribution of Brazilians in United States straightforward. Visualization successfully helps the display of data, and it helps to convey more information than original data.

6.2 Future Works

In the future, the clustering would be performed with more features, such as weather, economic status and etc..The clustering would contain more informations. Besides, if we obtain individual features, we are able to evaluate our model.

References

[1] https://d3js.org/