Implicit Bias Detection in Policing Practice

Zairan Xiang, zaxiang@ucsd.edu

Halıcıoğlu Data Science Institute & Department of Philosophy, University of California - San Diego

UC San Diego

Research Overview

Bias in real-world criminal data often leads to serious ethical challenges in the practice of predictive policing using machine learning models. The biased algorithm often reinforces any racial, gender, or other kinds of inequality that it picks up from real-world crime data as it produces biased predictions. The purpose of this research is to explore whether unsupervised machine learning methods can reveal biases in past policing practices in order to avoid feeding the biased data into predicted policing models.

Model

I introduce an improved algorithm based on k-means with better initializing techniques and a modified loss function with the support of feature reweighting.

- The original SSE loss function using Euclidean distance

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- the modified SSE loss function with the added penalty

$$d'(p,q) = \sqrt{\sum_{z=1}^{Z}(\frac{1}{\alpha_z}\sum_{i=1}^{n}(q_i^z - p_i^z)^2)}$$

The added penalty α would ensure the model pays equal amounts of attention to each feature and not overlooking numerical features that only have 1 dimension. With the support of features reweighting and a more efficient cluster initialization, the basic idea of the modified k-means is as follows in Algorithm 1:

Algorithm 1 K-means Algorithm

Input Data vector $\{x_n\}_{n=1}^N$, number of cluster K

$\{\mu_k\}_{k=1}^K \leftarrow$ **k-means++** $(\{x_n\}_{n=1}^N, k)$ ▷ Use K-Means++ to initialize centroids
while max iteration **do**
  for $n \leftarrow 1...N$ **do** ▷ new cluster assignment based on distance for each observation
    $r_n^{'} \leftarrow min_k$ Modified_SSE $(x_n - \{\mu_k\}_{k=1}^K)$
  end for
  for $k \leftarrow 1...K$ **do** ▷ update centroids based on current assignment
    $\mu_k \leftarrow mean\{x_n | r_n^{'} = k\}$
  end for
  if $\{r_n^{'}\}_{n=1}^N = \{r_n\}_{n=1}^N$ **then**
    Break ▷ early stopping
  end if
end while

Output cluster assignments $\{r_n\}_{n=1}^N$ for each data point, and k cluster centroids point $\{\mu_k\}_{k=1}^K$

Fig. 1: Modified K-Means Algorithm

By applying this improved clustering method, we should expect a uniform race distribution among clusters that also matched the overall distributions if the data are not overly related to race. Otherwise, we should expect to see the clusters to be highly related to protected personal traits like gender and race if any obvious discrimination based on these traits is present in past policing practices.

Clusters Results

Applying the modified k-means model with k=5 on LAPD arrest data from 2010 to 2019 and plotting the cluster assignments with dimensional reductions in Figure 2. From the graph, we can see some pretty clear clustering.

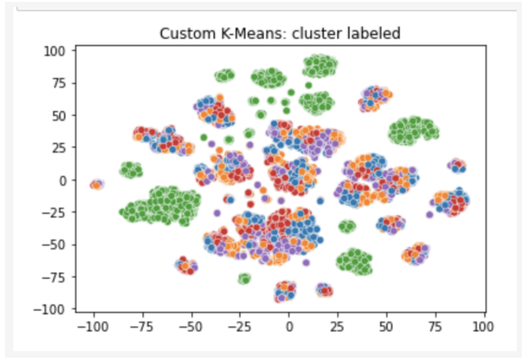

Fig. 2: Clusters

Analysis on Racial Distribution

In Figure 3, the red bar represents the percentage of certain races across all data, and the dark and light green bars represent the percentage of certain races within each cluster. The plot only shows four race groups that have a percentage over 1% in the dataset, which are Black (B), Hispanic/Latin/Mexican(H), White (W), and Other (O). By plotting the race distribution within 5 clusters and comparing it with the overall race distribution across all clusters, we can see from Figure 3 that for the middle cluster, the race distribution shows very different from the other clusters. With the arrest incident in cluster 2, the data contains more Black arrest cases and fewer White arrest cases compared to the true racial distribution in the dataset.

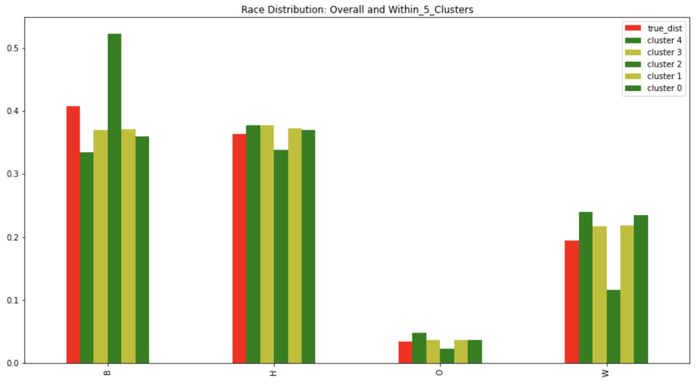

Fig. 3: Clusters

Discussion

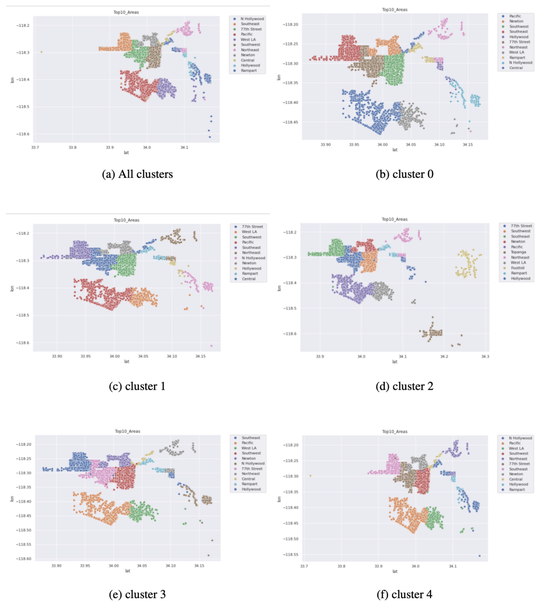

Fig. 4: Geographic Information of Arrest Incidents in Each Cluster

Cluster 2 is the only cluster that contains data from Foothill and Topanga (the far right data points on the map), which both have large proportions of White residents. The inclusion of the neighborhoods from the Foothill and Topanga areas might tell some interesting information since Foothill and Topanga shows opposite racial distribution compared to what it has in cluster 2. We can also see that except for cluster 2, the other four clusters share similar geographic information (with cluster 4 having an outlier), which is expected and matched the racial distribution information in Figure 3.

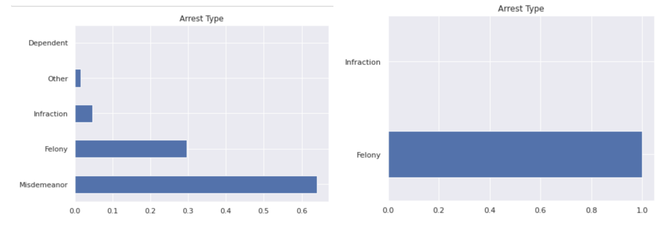

Fig. 5: Arrest Type Distribution

It also seems obvious that the arrest type is the dominant feature in the clustering: with the centroid in cluster 2 highly leaning toward the Felony arrest type. As shown in Figure 6 with a comparison with overall arrest type distribution, cluster 2 only contains almost only Felony incidents. We can conclude that racial information is highly related to arrest type in LAPD criminal data.