Report 2

ASSOCIATION RULE MINING OF

NEW YORK CITY’S STOP-QUESTION-FRISK (SQF) DATA

By

PRAJWAL NAGARAJ

Abstract:

Association rule mining is a valuable technique for discovering interesting patterns and relationships within datasets. In this report, we explore association rules within the context of police encounters from the Stop, Question, and Frisk (SQF) program in New York City. By analyzing demographic data and identifying frequent itemsets, we aim to uncover insights that can inform law enforcement strategies and policies.

Data Preparation:

1. **Construction of Transaction Data Set:**
    - **We constructed the transaction dataset by encoding demographic information, including race and age, into binary dummy variables. Instances where arrests were made were filtered out to focus on other outcomes of police encounters.**

Modelling:

1. **Creation of Frequent Itemsets and Association Rules:**
    - **We applied the Apriori algorithm to identify frequent itemsets with a minimum support threshold of 0.2. Subsequently, association rules were generated with a minimum confidence threshold of 0.8.**
2. **Results Visualization:**
    - **We visualized the top 10 frequent itemsets by support using a bar chart to highlight the most common patterns in the dataset. Additionally, a scatter plot was created to visualize the relationship between support and confidence for the generated association rules.**

Evaluation:

1. **Most Interesting Findings:**
    - **The association rules generated provide insights into the relationships between demographic attributes such as race and age. For example, certain demographic groups may exhibit a higher likelihood of specific outcomes during police encounters. Further analysis of these patterns could contribute to the development of targeted intervention strategies and policies aimed at reducing disparities in law enforcement practices.**

1. **Support vs Confidence Graph**:
    - This scatter plot displays the relationship between support and confidence for various association rules derived from a dataset.
    - **Support** on the x-axis ranges from about 0.55 to 0.80. This metric measures the proportion of transactions in the dataset that contain the itemset.
    - **Confidence** on the y-axis ranges from about 0.80 to 1.00. This metric measures the likelihood that an item B is also bought if item A is bought, expressed as a conditional probability.

    - The plot shows a wide distribution of points, indicating varying levels of confidence for different levels of support. Generally, as support increases, the confidence of the rules also tends to be higher, though there are exceptions.

2. **Top 10 Frequent Itemsets by Support**:
    - This bar chart ranks the top 10 most frequent itemsets based on their support values.
    - The itemsets are combinations of different racial/ethnic group identifiers such as 'WHITE', 'BLACK', 'WHITE-HISPANIC', 'BLACK-HISPANIC', and 'ASIAN/PACIFIC ISLANDER'.
    - The support values are shown on the x-axis, and they range from about 0.1 to nearly 0.9, indicating the prevalence of these itemsets in the dataset.
    - The most frequent itemset, containing only 'BLACK', has the highest support, close to 0.9, suggesting it appears in nearly 90% of the transactions (or data entries).
    - Other combinations, such as 'WHITE-HISPANIC', 'BLACK', and 'WHITE-HISPANIC', also show high support, indicating these combinations are common in the dataset.

Analysis:

- The results suggest that certain racial/ethnic group combinations appear more frequently together in the dataset, which could be indicative of underlying patterns or associations in the data.
- The scatter plot of support vs confidence helps in understanding the strength and reliability of these associations. High confidence at high support levels suggests strong and reliable rules.
- These insights can be used for further analysis, such as demographic studies, targeted marketing, or policymaking, depending on the context of the data.