CORRELATION ANALYSIS OF TAICHUNG TRAFFIC ACCIDENT DATABASE

YI-FAN CHEN, PING-HE YEH, YI-SHAN CHENG, PO-CHUN HU

Abstract :

This study analyzes Taichung's traffic accident database, utilizing methods like Recursive Feature Elimination and Weka's attribute evaluation. Through K-modes clustering, we identified key factors contributing to accidents and proposed targeted improvement measures. The findings offer valuable insights for enhancing road safety and accident prevention in Taichung.

Keyword:
K-modes; traffic accident analyze

1. Introduction

In recent years, there has been a surge in traffic accidents. Although law enforcement and government efforts to improve roads are criticized by the public for the regulations they have formulated as highly imperfect. It is only through a scientific analysis of accident hotspots, including factors such as road conditions, and driver states in order to identify the underlying causes of accidents that we can genuinely reduce the frequency of accidents.

We implemented K-modes clustering to group the data and, from the clustering results, analyzed potential road conditions contributing to accidents, such as weather, road type, and accident location. Based on these conditions, we further proposed several improvement measures.

2. Method

We use the traffic accident dataset1 from the Taichung City Police Department from July 2022 to June 2023 which is available on the government open data platform. There are a total of 55 attributes, with "Accident Type and Pattern" being the target attribute. We attempted to filter attributes using Recursive Feature Elimination (RFE)[1] on 55 attributes, resulting in only 5 attributes remaining. However, the performance in the subsequent process was unsatisfactory. Additionally, we explored attribute evaluation using Weka[2]. We retained the top-performing 9 attributes, but the performance is still not satisfactory. Considering the results, all attributes have some impact on the target attribute. Therefore, we are considering using all attributes for association analysis.

After selecting attributes, We integrated all the datasets, totaling 115,392 records. After removing data with missing values, we were left with 61,286 records, each record has a following accident type, totaling 34 types of accident. We converted numeric data into nominal format for further processing and subsequently mapped the target attribute.

We applied K-modes[3] clustering to group all accident data into 10 clusters. Subsequently, we conducted an analysis of the top three clusters with the highest membership to identify relationships among accidents sharing similar attributes and explore potential causes behind these accidents. Similarly, we employed K-modes clustering to group individual accidents and analyzed the causes of single accidents.

3. Result & Analyze

Figure 1 shows 10 clusters without the attribute “accident type”. In most clusters, car and car accidents are the major types of accidents.

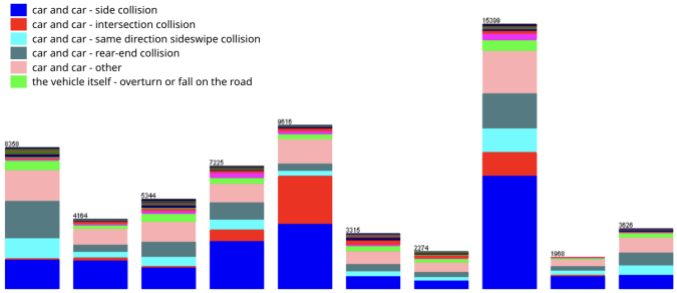

Figure 1

[1]**D**atasetlink: https://reurl.cc/lgV6l6

We found that cluster 0, cluster 2, and cluster 4 are similar in several key attributes, as shown in Figure 2. However, the distribution of accident types in these three environments is quite different, we further analyze these three situations.

	weather	light	road condition	road type	accident location	road defect	obstacle
0	sunny	daytime	dry	urban area	**straight road (without fast or slow lanes)**	none	none
2	sunny	**daytime**	dry	urban area	**three-way intersection**	none	none
4	sunny	**lighted at night**	dry	urban area	**four-way intersection**	none	none

Figure 2

In cluster 0, most accidents are rear-end and side collisions. This may be related to good daytime visibility, dry road conditions, and the characteristics of straight roads, which can lead to drivers becoming more relaxed and consequently more likely to become distracted, speed, and engage in other behaviors. Installing additional traffic signs and strengthening enforcement against speeding and distracted driving may reduce the occurrence of accidents.

Similar to Cluster 0, Cluster 2 with these types of conditions may cause side-collisions due to not checking rearview mirrors while turning or following too closely behind the vehicle in front. Designating a specific lane for turning vehicles may solve this problem.

Unlike the first two cases, Cluster 4 had the incident occurring at night. We assume that even with street lighting, visibility is still relatively poor compared to daytime, which could lead to drivers noticing pedestrians too late to avoid and subsequently side-colliding with other vehicles. Installing additional street lights and separating the timing for pedestrians and vehicles to cross intersections may improve this issue.

We further clustered the three types of accidents. The first one is "Accidents involving overturning and falling off the road" with a total number of 3327. There are 672(22%) cases with clear weather, lighted at night, three-way intersections, near crossroads, no traffic signals, and no lane dividers. Based on the conditions, it can be assumed that accidents are more likely to occur at three-way intersections within alleyways. Adding a speed hump before these alleyway intersections can be considered.

The second is "Pedestrian struck while crossing the road" with a total number of 1400. A cluster with 251(18%) members under the conditions of daytime, four-way intersection, within an intersection, no traffic signals, and no lane dividers. It can be assumed that pedestrians are more likely to be struck while crossing alleys without traffic lights. So installing traffic lights can be considered.

Next, "Collisions in the same direction" with a total number of 6051. There are 1052 accidents that happened with condition daytime, in three-way intersections, within an intersection, with no lane dividers, and no traffic signals. it can be assumed that overtaking in some alleys can lead to accidents. Installing "no overtaking" signs within those road segments can be considered.

4. Conclusion

In our report, we explored various methods for attribute selection, including RFE and attribute evaluation in Weka. Additionally, we utilized K-modes to categorize the data into 10 clusters. We conducted an analysis of the results after clustering to identify attributes correlated with accidents. Furthermore, we delved into potential underlying causes and proposed several improvement measures.

References

[1] https://www.analyticsvidhya.com/blog/2023/05/recursive-feature-elimination/.

[2] https://weka.sourceforge.io/doc.dev/weka/attributeSelection/ClassifierAttributeEval.html

[3] https://www.analyticsvidhya.com/blog/2021/06/kmodes-clustering-algorithm-for-categorical-data/