Analyzing the Traffic Crash Data of Chicago

Anisha Islam, Student ID: 1783191, ccid: aislam4

Problem The consequences of traffic accidents and crashes can be fatal and financially devastating to people. Analyzing the traffic crash data can help law enforcement authorities to identify the reasons and potential locations for traffic accidents, develop infrastructures, and allocate resources accordingly to reduce the severity of accidents. This project aims to analyze the traffic crash data of Chicago [1] and determine the effects of the contributing factors, such as speed limit, vehicle condition, weather condition, time of day, road type, road condition, and location. Using different clustering algorithms, we will extract relevant clusters and information to identify the common reasons and conditions for traffic accidents in Chicago.

Methodology We have identified the following tasks to achieve our goal:

1. **Literature Review:** The first step is to review two different kinds of research, traffic crash and applied clustering, to understand contemporary works.
2. **Preprocess the dataset and select useful features:** Secondly, we plan on preprocessing the collected dataset by filling in missing data, removing duplicates, encoding features, and extracting features that are most likely to affect the possibility of a traffic accident.
3. **Find correlation:** Next, we plan on applying statistical analysis to find the correlation between different factors to determine how the combination of these factors can affect traffic crashes.
4. **Select the clustering algorithms:** After that, we plan to select the clustering algorithms for our project based on the information from the class lectures and literature review.
5. **Comparison:** Then, we plan on experimenting with different clustering algorithms and comparing the performances of these algorithms using various cluster validation techniques.
6. **Presentation:** Finally, we will present our work.

Evaluation We will use the traffic crash dataset [1] from the Chicago Data Portal [2]. We will evaluate our clusters using different cluster validation algorithms introduced in class (i.e., Jaccard Coefficient, Adjusted Rand Index, Silhouette Width Criterion) and compare the performance of different clustering algorithms. We also plan on incorporating different cluster validation methods used in relevant works. We will conduct different experiments based on the following research questions:

1. **RQ1** What are the most accident-prone areas of the city?
2. **RQ2** When do the crashes happen more often?
3. **RQ3** What is the most common factor for road accidents?
4. **RQ4** Is there any correlation between different contributing factors?
5. **RQ5** Is there any pattern in traffic accidents considering different contributing factors?
6. **RQ6** Is there any outlier in the dataset?

Timeline We expect to implement the project by 9th April 2023 in order to schedule a meeting for the presentation before the end of the semester. A tentative timeline for the completion of the project is given below:

Week	Dates	Task
1	February 27 to March 5	Literature review
2	March 6 to March 12	Preprocess the dataset and select useful features
3	March 13 to March 19	Find correlation between different features by using statistical analysis
4	March 20 to March 26	Select the clustering algorithms and start extracting relevant information
5	March 27 to April 2	Compare the performance of different clustering algorithms
6	April 3 to April 9	Prepare the presentation and demo

References

[1] "Traffic crashes - crashes." https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if. Accessed: 2023-02-23.

[2] "Chicago data portal." https://data.cityofchicago.org/. Accessed: 2023-02-23.