Aircraft Collision Avoidance System (ACAS)

Research on the main reasons for near midair collisions

Team: Wei Wei, Lingyi Gu, Haoyu(Jerry) Wu

1. Abstract

In the US, the VFR (Visual Flight Rules) allows pilots almost unlimited freedom to fly anywhere without filing a flight plan. In 2016, there were 179 reported near midair collisions, which led to critical consequences. Our main objectives are to generate insight from the narratives of collision reports that help researchers reinforce their collision avoidance strategy to alleviate the situation. The two specific questions we would like answer are as follows:

1. What are the major reasons for midair and near midair collisions?
2. How are the altitude and the relative position of the aircraft related to these major reasons? For example, an air collision happened near the aircraft may due to a wrong decision made by the traffic alert and collision avoidance crew, while a collision occurred far from the aircraft my because of pilot's operation error.

2. Data Description

2.1 Data Retrieval

We retrieved our data from the following two sources.

1. **NASA's Aviation Safety Reporting System**: We downloaded the collision data in New England area from the system and mainly utilized the narratives and summaries of each report, along with the altitude and the relevant distance between aircrafts. (via downloading CSV files)
2. **The U.S. National Transportation Safety Board**: We also scraped informative reports on aviation incidents, which serve as a useful aid to supplement our understanding of the data sourced from the ASRS database. (via **web scraping**: since the website is written in JavaScript, we use Selenium to simulate browser actions and scrape data from it)

2.2 Data Cleaning and Preprocessing

We have done the following to clean up the raw data downloaded or scrapped from the website.

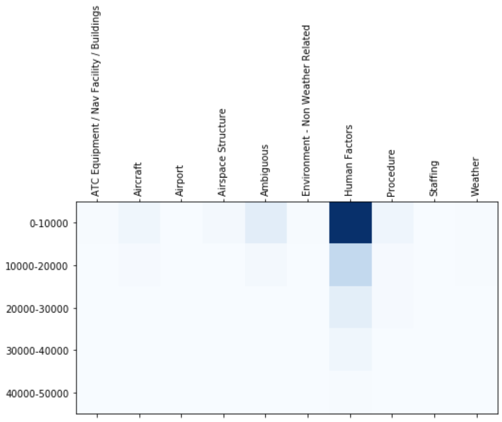
1. **Stemming**: eliminate noises in the text
   Since we are mainly focusing on the narrative and synopsis in the report for analysis, we eliminate some noises in the text by running the snowball stemmer on the text fields to get a more accurate result later in the principal component analysis (PCA).
2. **Stop words**: remove common English words (eg. we, are)
   This is another step we have done to eliminate the noises.
3. Combine two datasets.

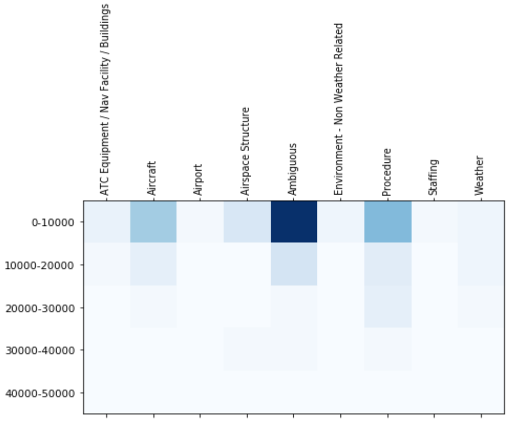
2.3 Preliminary Analysis

By looking at the data at first glance, there are many factors can cause the mid-air collisions. To give us an intuitive idea which factor we should specifically look for in our narrative analysis, we created the following **heat map** for preliminary analysis.

Figure 1. Human factors included Figure 2. Human factors excluded

Primary problems vs. Altitude (A darker color indicates that more incidents are reported)

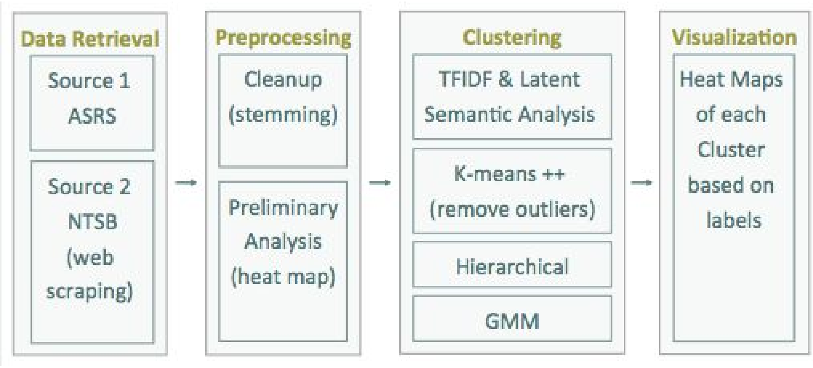




From the graph above, we can see that **human factors** seem to be a dominant reason for collisions regardless of the altitude. Other factors such as aircraft, airspace structure, procedure and some ambiguous reason also count. This can give us some insights into our first question, which is that **human factor is the major reasons for midair and near midair collisions**. Therefore, in our further analysis, we want to specifically look for what kind of actions or attributes of these major factors, or a combination of them at different altitude would cause the collision.

3. Approaches

The following graph shows an overview of our approach.



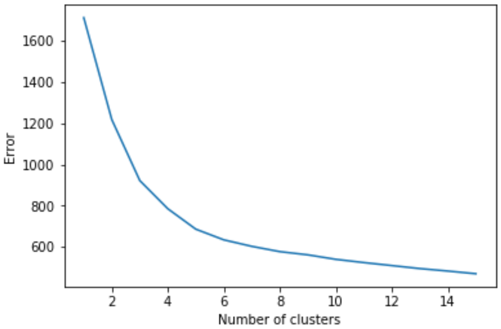
Data Retrieval
Source 1 ASRS
Source 2 NTSB (web scraping)
Preprocessing
Cleanup (stemming)
Preliminary Analysis (heat map)
Clustering
TFIDF & Latent Semantic Analysis
K-means ++ (remove outliers)
Hierarchical
GMM
Visualization
Heat Maps of each Cluster based on labels

3.1 Latent Semantic Analysis (with PCA)

We utilized PCA to extract the feature space from the narratives of the incident report and find the most significant mentioned terms like “visual.” We also use SVD to compress the matrix to a low dimension and normalize the data. After performing PCA and SVD, we scale the altitude and relevant distance up because we want to cluster based on these two features, but we also take the terms into account.

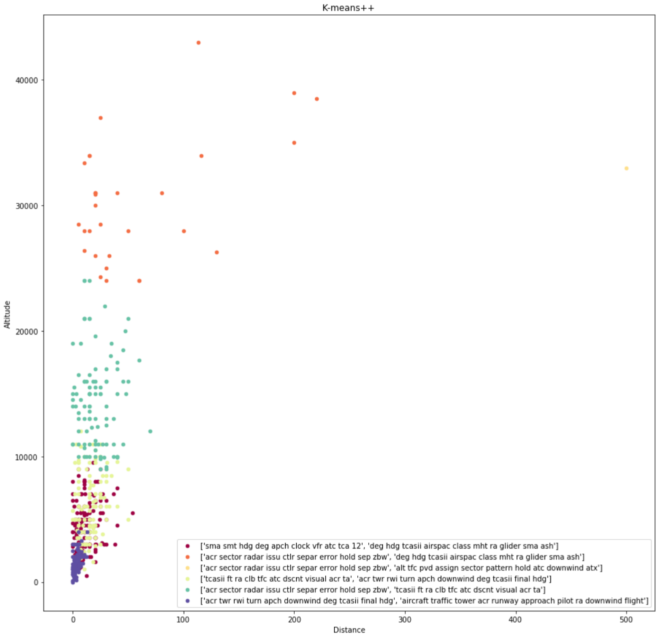
3.2 Clustering

3.2A Determine the number of cluster k



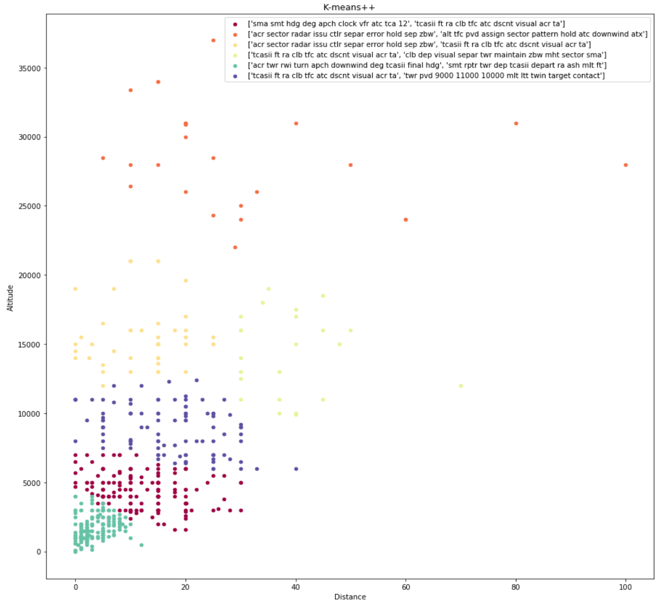
From the graph on the left, we can see that the error reach below 800 after 6 clusters and does not have a huge change after this number. Therefore, we choose 6 as the optimal number of clusters.

3.2B Run clustering algorithms



From the graph on the left, we can see that there is a outlier cluster that only contains one data point. Therefore, we need to remove the outlier to get a better result.

3.3 Remove outliers



To make our cluster more accurate and reliable. We have removed top 5 outliers from each cluster. The resulting graph (right) looks more stable.

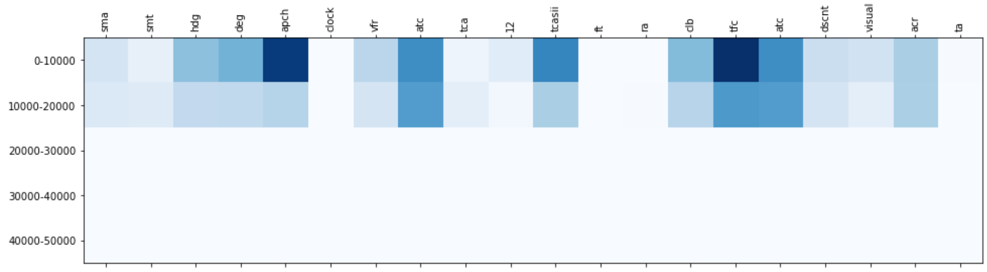
4. Experiments

4.1 K-means++ vs. Hierarchical vs. GMM

Because we have examined these three metrics in a past homework and have a good result, we use the same metrics for clustering. All three metrics give us satisfying results, but K-means ++ has the highest Silhouette Score, so we chose it for further analysis.

4.2 A heat map of each cluster based on labels

Clustering gave us an intuition of how different combination of factors at each different altitude, which may lead to a collision. By plotting a heat map for each cluster, we can observe the number of occurrence of these words and determine the important ones.



Labels vs. Altitude (A darker color indicates that the term has more occurrences in the report)

Take the above heat map as an example, atc (Air Traffic Control), hdg (Heading Mode: the autopilot keeps the nose of the airplane pointed at the magnetic heading bug), degree of turning and also visual may be important factors in causing the collision of altitude in range (0-2,0000)

5. Conclusions

Here are the conclusions we have drawn by examining the clustering results and experimental data. To answer our two questions: What are the major reasons for midair and near midair collisions? How are the altitude and the relative position of the aircraft related to these major reasons?

1. The lower the altitude, collision is more likely to occur because aircrafts are taking off or landing.
2. Among low altitude collision (<20,000) during climbing and descending, the parameters which may cause the collision are: downwind, tower, true airspeed, aircraft turning, heading mode, traffic control, visual and collision happens frequently among single-engine piston.
3. Among high altitude collision (20,000 ~ 50,000), collisions are likely happen during sector (a portion of an itinerary) or climb, and associates more with the arrangement of the traffic control center and airport. Visual and radar may also play an important role here.

6. Future Work

1. **Outlier detection improvement**
2. **Text vectorization improvement** (e.g. remove numbers)
   As you may notice our heat maps includes some words or numbers which does not make sense, this is due to the size of the narrative and make it difficult for us to ensure all our LSA components make senses and relate to the collisions. One way to improve manually pick the important factors in contributing to the collisions and cluster the narratives based on the similarity between these factors. However, we are not experts on air collisions and have limited time, this can be potentially improved in the future.
3. **Analyze based on phase rather than single word**
   Rather than analyzing on single word, we can also focus on phases because single word sometimes does not give us an intuitive finding. For example, it is hard to understand what the single word "failure" stands for without some context to supplement it. It could a system, radar or a human's failure.

7. Reference

1. NASA Aviation Safety Reporting System
2. National Transportation Safety Board (NTSB)
3. CS506-Computational-Tools-for-Data-Science

8. GitHub Repository

For source codes and more visualizations, please kindly refer to our github repository @ https://github.com/lingyigu/acas