Using Data Mining/Machine Learning to Auto-categorize ASRS Database

38

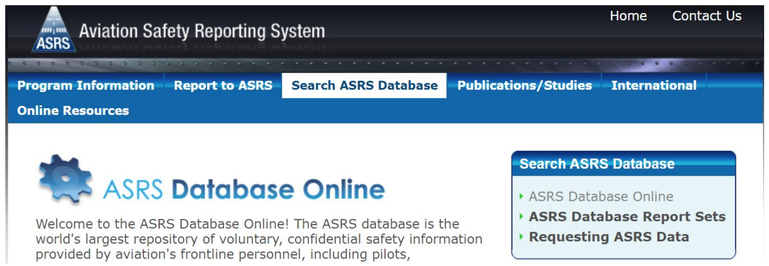
STUDENTS: Chase Whyte, Chia-Lin Liu, Wei Da Chen

Abstract

- Auto-categorizing aviation safety reports from ASRS database into 8 common occurrences defined by CICTT categories
- A tool which is compatible with all the aviation database
- Achieving 90% accuracy using dictionary vectors and other features

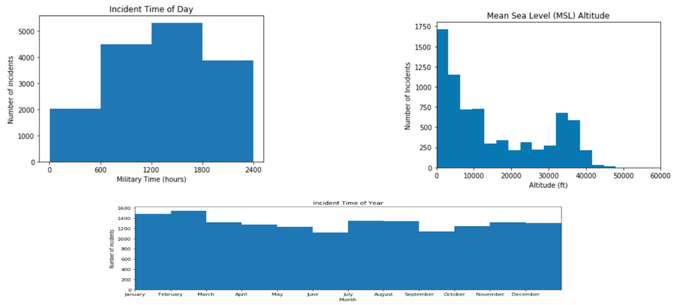
Introduction

- The ASRS database is the world's largest repository of voluntary, confidential safety information provided by aviation's frontline personnel, including pilots, controllers, mechanics, flight attendants, and dispatchers.
- CICTT categories are maintained by ICAO Taxonomy Team and provide a common standard for labeling flight incidents according to 34 categories. These categories were narrowed down to the most representative categories in the database.



- 1450 aviation incidents were hand-labeled according to four main categories: ATM, ADRM, TURB/WTHR, SYS
- Procedure extended to include 8 categories: AMAN, ADRM, CABIN, MAC, NAV, FOS, SYS, WTHR

Data Visualization



Data Preprocessing

- Data Cleansing: collecting 28 features by removing sparse features (more than 10% missing values)
- One-hot Encoding / Binary Encoding for multi-label features
- Label Encoding for categorical features
- Numerical Normalization for numeric values

Keyword Extraction

- The textual data fields in the ASRS database contain descriptions of an incident written by anyone who reports the event, hence filtering of the data is required before applying it to our model.
- Pre-processing techniques implemented are punctuation and stopwords removal, lemmatization and keeping verbs and nouns which are more informative.
- Keywords are extracted using TF-IDF on uni and bi grams for each categories and grouped into a bag of word.
- Binary encoded the whole corpus using the bag of word generated.

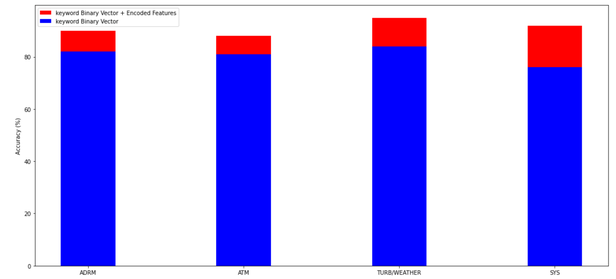
Word cloud generated using TF-IDF

- Dictionary generation for each category using the method mentioned above.
- Having a representative set of words in the dictionary is crucial to the prediction accuracy, so the process of dictionary generation and fine-tuning parameter is repeated till reasonable accuracy is achieved on the validation set.
- Classify on un-labelled data using the model generated and manually inspected the quality of the prediction on unseen results.

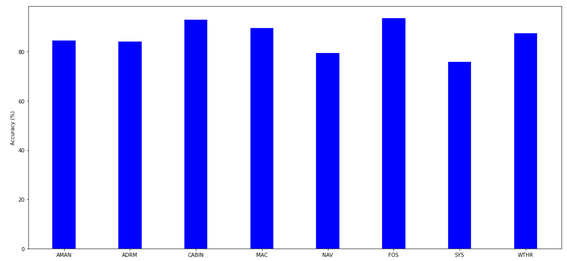
KNN & Results

- Using K-nearest neighbors approach as our main algorithm
  - Advantages: intuitive, naive, universal consistency
  - Disadvantages: large computation complexity when input has high dimensions
- Input: keyword generated binary vectors + encoded other features
- Output: Multilabel multiclass CICTT categories (binary vectors)
- Results

ADRM, ATM, TURB/WEATHER, SYS (4 main categories)
Overall Accuracy: 82% (keyword binary vector) / 91% (+encoded features)



AMAN, ADRM, CABIN, MAC, NAV, FOS, SYS, WTHR (8 common categories
Overall Accuracy: 86%



Future Work, References, and Acknowledgments

- Use different machine learning classifiers to compare performance
- Further automate the whole process by writing a script that has a GUI display as well as a dynamically display information based on the user's input.
- Extend the script to work with other aviation database.

Faculty: Payman Arabshahi, Arindam Kumar Das
Industry Sponsor: Tak-Kei Lee, Karina Cuadrado, John Dong
TA: Shruti Misra, Brandon Yee

[1] International Civil Aviation Organisation (ICAO), and Commercial Aviation Safety Team (CAST). "AVIATION OCCURRENCE CATEGORIES." *Aviation Occurrence Categories Definition*, Common Taxonomy Team, Oct 2011, https://www.icao.int/APAC/Meetings/2012_APRAST/OccurrenceCategoryDefinitions.pdf.

[2] Mariana Carmona."ASRS Database." *Aviation Safety Reporting System*, National Aeronautics and Space Administration (NASA), https://asrs.arc.nasa.gov/search/database.html.

ELECTRICAL & COMPUTER ENGINEERING
UNIVERSITY of WASHINGTON

ADVISORS: Tak-kei Lee, Karina Cuadrado, John Dong

SPONSORS: Boeing