2017

Analysis of Automatic Dependent Surveillance-Broadcast Data

Salcido, Ryan; Kendall, Anthony; Zhao, Ying

AAAI

Salcido, Ryan, Anthony Kendall, and Ying Zhao. "Analysis of automatic dependent surveillance-broadcast data." 2017 AAAI Fall Symposium Series. 2017.
http://hdl.handle.net/10945/62615

Deep Models and Artificial Intelligence for Military Applications:
AAAI Technical Report FS-17-03

Analysis of Automatic Dependent Surveillance-Broadcast Data

Ryan Salcido, Anthony Kendall, Ying Zhao
Naval Postgraduate School, Monterey, CA, USA
ryansalcido.rs@gmail.com, wakendal@nps.edu, yzhao@nps.edu,

Abstract

The primary research objective is to detect commercial flight baselines, anomalies and patterns using Automatic Dependent Surveillance – Broadcast (ADS-B) data to analyze aircraft tracks over a period time using Lexical Link Analysis (LLA) and visualizing with Google Earth and Maps. This research could potentially improve situational awareness for naval air warfare decision makers. LLA is a form of text mining showing relationships and associations with the given data. Because there is a large amount of daily ADS-B data, a Hadoop cluster is utilized for parallel processing and then LLA provides data visualizations, patterns and associations for profiling the aircraft based on the kinematic characteristics. Based on the correlation to the speed and altitude of aircraft and its country of origin, our results did identify unusual behavior for some aircrafts. When the kinematic and behavior patterns are discovered from historical ADS-B data, the resulted models can also be used to identify flying patterns and anomalies that can increase CTAP and the prediction accuracies of CID.

Introduction

The Naval Common Tactical Air Picture (CTAP) collects, processes, and analyzes data to provide situational awareness to air warfare decision makers. Accurate Combat Identification (CID) enables warfighters to locate and identify critical airborne objects as friendly, hostile or neutral with high precision. CID plays an important role in generating the CTAP and other combat systems by reducing the probability of fratricide.

Deep Analytics are critical to design futuristic CTAP and CID products such as adaptive, cooperative and learning combat systems along with authoritative data sources, standards and interoperability from sensors, platforms, and weapons for mission requirements. It is also imperative to test and adapt commercially available tools to satisfy the ongoing needs and requirements of the CTAP and CID.

Deep Analytics are abundant in the commercial world, but not appropriate to solve military applications such as CTAP/CID without modification or assisting technologies. Big Data sources for the CTAP and CID applications including disparate real-time sensors with extremely high velocity rates and large volumes which require information sharing as well as security across all domains make these applications more challenging when using commercial technology.

The Naval Postgraduate School (NPS) team has applied Big Data, Deep Learning and artificial intelligence methods to specifically examine open source Federal Aviation Association (FAA), Mode-S, and Automatic dependent surveillance – broadcast (ADS-B) for known US regions to identify kinematic signatures for commercial aircraft using commercially available Deep Analytics including various Deep Learning algorithms for pattern recognition and anomaly detection. The resulted models can be used as fingerprints or signatures to identify anomalous from normal commercial aircraft behavior.

The primary research objective is to detect commercial flight anomalies and patterns using Automatic Dependent Surveillance – Broadcast (ADS-B) data to analyze the tracks of aircraft over a period time using Lexical Link Analysis (LLA) and visualizing flight paths with Google Earth and Maps.

ADS-B Data

An aircraft can be identified by radar transponder identification, friend or foe (IFF) modes such as one, three, and five (military). This is done by Line of Sight (LOS) air traffic control ground radar stations. For improved (cooperative) surveillance for flight separation and control an aircraft can have an ADS-B Out system to broadcast its identification and location from the aircraft's GPS to LOS receiving ground stations and other aircraft (around a 150 mile range). An interesting note: an ADS-B receiver on Amazon costs about $100. The transmit frequency is generally 1090 MHz but there are slight differences in some countries. There is also *ADS-B In* that allows ground stations to broadcast to planes weather data and other needed information.

The stations are not in the middle of the ocean and according to the FAA, improved oceanic surveillance could be the next opportunity for ADS-B. As part of a project called Advanced Surveillance Enhanced Procedural Separation (ASEPS), the FAA is analyzing two methods of improving surveillance coverage in oceanic service volumes: Space-Based ADS-B reports and more frequent Automatic Dependent Surveillance – Contract (ADS-C) reports.

According to International Civil Aviation Organization (ICAO), a.k.a, the international "FAA", states that notable outcomes of using ADS-B include a new frequency allocation for space-based ADS-B reception, enabling tracking of aircraft globally including remote and polar regions.

FAA is likely to make a decision in 2018 to fully employ ADS-B and implement the related infrastructure in 2020. The Aireon company in cooperation with FAA has already done a successful test of a space based ADS-B using Low Earth Orbit (LEO) Iridium satellites and Space X launches will add more satellites (Collins, 2017; Grush, 2015; Hillenbrand, 2017). According to a report, by 2030 oceanic traffic will double and the current way of reporting an aircraft's position/altitude via high frequency (HF) radio when oceanic just cannot handle the traffic. ADS-B will allow more aircrafts to be closer together to accommodate for the increased traffic. ADS-B provides an additional return on investment (ROI) by allowing in-trail procedures so aircraft can save fuel/time by flying more direct (NextGen: In-Trail Procedures, 2016).

ADS-B data can be useful in developing decision support systems and providing Business Intelligence (BI) in assisting airport management. For example, an interesting non-track use of ADS-B data analyzed arrivals at Schiphol, Amsterdam (Zmarrou, 2015).

The NPS team downloaded historical ADS-B JSON files (taken every minute) dating back to June 2016 (Dan, Data Field Descriptions, n.d.). There are about 1,440 (6MB each) files produced each day. Over a year, the data would be over 3TB or about a half million JSON files at 6MB for each one. The focus of this paper is to discuss how to analyze this Big Data set using Big Data processing capabilities such as a Hadoop cluster and Deep Learning of pattern and anomaly detection algorithms such as Lexical Link Analysis (LLA). When the kinematic and behavior patterns are discovered from historical data ADS-B, the resulted models can be also used to identify flying patterns and anomalies to enhance the CTAP and the predictive accuracies of CID.

ADSB Exchange provides unfiltered ADS-B data with many aircraft characteristics not important to data mining and therefore the first step filters the data to only those attributes useful to the research objectives such as position, speed, and registration number.

The NPS team first wrote a Python script to extract the required fields from the initial JSON data. Because the initial JSON data is a Big Data, the Python script was designed to be executed in the Big Data software such as Hive and Spark alongside a Hadoop Distributed File System (HDFS).

- The Python script organizes the data based on its unique registration number and time it was tracked to allow aircrafts to be monitored throughout the day. The data is stored in a tab-separated values text file to be used for data mining. The following fields of ADS-B are initially considered for the research objective above (Dan, Data Field Descriptions, n.d.):
    - Reg: Aircraft registration number
    - FSeen: Date/time receiver first started tracking aircraft on particular flight
    - TSecs: Number of seconds aircraft is tracked
    - Trak: Aircraft's angle from ground
    - Alt: Altitude (in feet)
    - Lat: Latitude
    - Long: Longitude
    - Species: Aircraft type
    - Mil: Is aircraft operated by the military?
    - Cou: Country aircraft is registered to
    - HasPic: True if aircraft has a picture
    - Interested: Is aircraft marked interesting?
    - Gnd: Is aircraft on the ground?
    - Trt: Transponder type
    - Engines: Number of engines the aircraft has
    - Icao: Hexadecimal identifier for aircraft
    - Man: Manufacturer's name
    - Type: Aircraft model's ICAO type code
    - OpIcao: ICAO code of operator
    - ΔV: change in Velocity
    - ΔH: change in Heading (direction)
    - ΔA: change in Altitude
- The Python script is also designed to filter the data further based on a specified radius. This will limit the amount of aircrafts that are shown to have a more accurate understanding of what is occurring in that area.

Implications with ADS-B Data

ADS-B data contained a large number of attributes associated with each aircraft and those not important to the research were filtered out by creating a script based on which attributes were desired.

Also, filtering was used to insure integrity and Information Assurance (IA) of ADS-B data essential to research credibility. Since the data is sent from the aircraft thousands of feet in the air to ground stations, one of the first steps of the research was to use the ADSB Exchange's data field descriptions webpage to determine which fields were reliable (Dan, Data Field Descriptions, n.d.). From examining the documentation, some of the descriptions emphasized how certain fields are user-reported, meaning that the pilot manually enters in the information. Although the pilot is most likely entering correct data, this exemplifies how some fields may be of questionable validity.

Another IA concern is based on ADS-B not automatically encrypting the data and therefore data transmitted to the ground stations could be compromised and inaccurate.

Overall, ADS-B is a reliable source for tracking commercial aircraft but analysis could help insure data integrity.

Utilizing the Hadoop Cluster

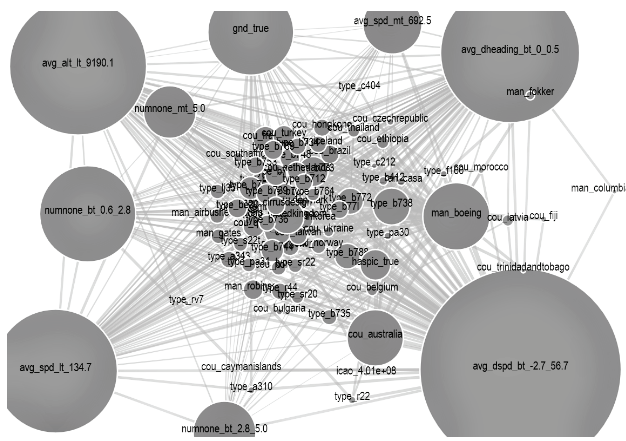
The Hadoop cluster system is designed to store and analyze large data sets in a distributed computing environment (Rouse, 2017). This means that the Hadoop cluster provides the ability to process the large amounts of ADS-B data by distributing pieces of the job to multiple nodes within the cluster. The Hadoop cluster has more computational capabilities when compared to an average computer. The data that is processed using Hadoop is stored in HDFS to provide a reliable storage solution to the large amount of processed data. This allows Hadoop to be robust and efficient in its big data processing.

Since the script is programmed using Python, Hadoop streaming will be used to execute the script on the cluster. Hadoop streaming requires a Java archive (JAR) file, a mapper and reducer script, the path to where the input files are stored, and where to store the output files. Hadoop streaming uses the mapper script to process each line in the JSON file and prints the contents to standard output. Then, the reducer script takes the standard output from the mapper script and reduces the set of values that share a common key. In this case, the key would be the aircraft's registration number. The reducer script then prints its' contents to standard output which is saved in a text file specified in the command. The next step after completing the job on the cluster is to do Lexical Link Analysis.

ADS-B Data with Lexical Link Analysis

ADS-B data is filtered down to a more useful data set that can used for data mining and for Lexical Link Analysis (LLA). LLA visualizes the associations among different flight characteristics of the data to detect outliers that are considered anomalies. Through the LLA visualization tool, the aircraft data can be visualized based on relationships between lexical features. These visualizations can then be filtered based on keywords and the amount of links between the entities (Zhao, MacKinnon, & Gallup, System Self-Awareness and Related Methods for Improving the Use and Understanding of Data within DoD, 2011; Zhao, MacKinnon, & Gallup, Big Data and Deep Learning for Understanding DoD Data, 2015).

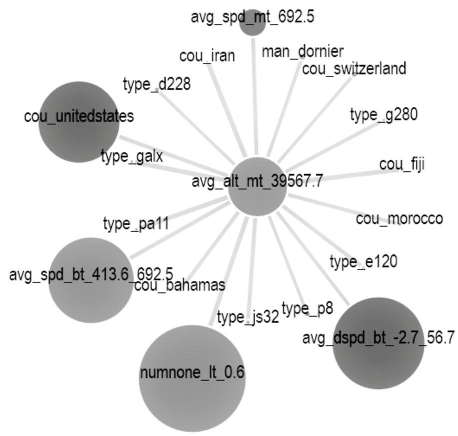
Analyzing the ADS-B data in LLA proved to be a useful tool in detecting flight patterns and anomalies. The data set analyzed in LLA was over a period of one hour with five-minute increments. In one hour, there were over 7,000 unique aircrafts tracked with ADS-B. Prior to analyzing the data in LLA, one initial test was to determine if the data made sense under normal circumstances. For example, Figure 1 illustrates the patterns and associations such as having an altitude less than a certain value where each link represents a pattern. For example, node "avg_alt_lt_9190.1" linking to "avg_spd_lt_134.7" means that the planes that have an average altitude less than 9190.1 feet are also likely to have the average speed less than 134.7 miles per hour. This makes sense since the lower flying airplanes likely fly slower. The likelihood of this pattern is represented in the thickness of the edge between the two nodes. Also, node "gnd_true" linking to node "avg_alt_lt_9190.1" shows that it makes sense because it associates the aircraft being on the ground to having an average altitude less than 9190.1 feet.



[Figure 1] An example of a word cloud visualization in Lexical Link Analysis. The different bubbles represent attributes with certain values that appear often in the data. The thickness of the link is the probability of two associated attributes.

The other part of data analysis with LLA was to detect any anomalies. Based on the correlation to the speed and altitude of aircraft and its country of origin, our results did identify unusual behavior for some aircraft.

Figure 2 shows the various word pairs associated with having an average altitude greater than 39657.7 feet. An aircraft flying this high is unusual because it's close to the limit for commercial planes. However, this visualization shows how some countries such as Iran, Morocco, and United States are associated with having planes flying higher than this threshold. As a result, this can become suspicious because not many planes can go higher than 39,000 feet.



[Figure 2] The picture above is another visualization from Lexical Link Analysis with ADS-B data. This one shows all the associations between planes that have an average altitude greater than 39567.7 feet.

Overall, the objective of analyzing ADS-B data in LLA and deriving useful information was successful because we noticed flight paths' patterns as well as anomalies and verifying the data did make sense. Further research will be conducted on analyzing the patterns and associations on the whole Big Data set.

Limitations when Performing Lexical Link Analysis

LLA is a useful tool for this research project because of its capabilities in visual large data sets and organizing it based on commonalities. However, since the tool is web-based, it has limitations to how much data it can process at once. The web-based LLA is capable of handling big data, but it is difficult to determine the maximum amount of input the browser can handle. The browser runs extremely slow when uploading an unfiltered ADS-B file to the collaborative agent that is used for LLA. It becomes difficult to analyze the data to check for associations and then make modifications to filter the data. Unfortunately, there is not a solution to this because of the browser limitations.

Google Earth Application Programming Interface

[Figure 3] The picture above shows one visualization of Google Earth using KML. Each number on the map is a randomly generated unique ID representing an aircraft.

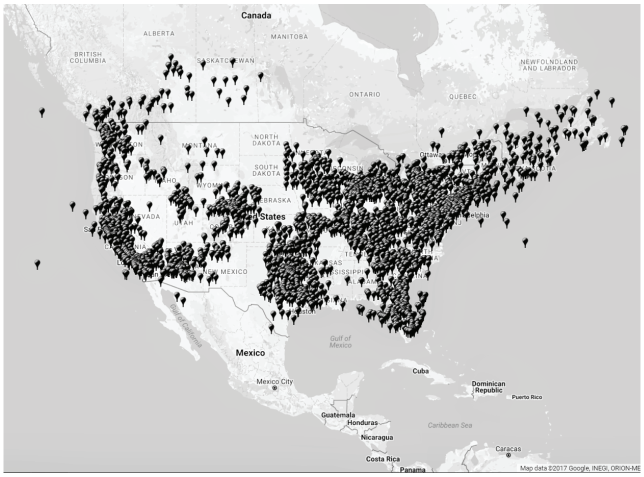
In the beginning of the research, the Google Earth Application Programming Interface was used to visualize ADS-B data. Google Earth utilizes a Keyhole Markup Language (KML) to display the data on the globe. KML is useful as a Big Data aggregator because of its scalability. Google Earth allows warfighters to collect and analyze accurate CID to have the ability to locate an aircraft on the globe. In addition, it is important to air combat personnel to provide positional awareness of other aircrafts whether they are friendly or hostile with the use of the Naval CTAP. Google Earth also highlighted areas of sparse and dense aircraft traffic and the ability to track planes over a period of time.

Google Maps Application Programming Interface

Google Maps was another tool that was used to visualize a subset of the ADS-B data. This is helpful because it allows the data to be visualized on a map with each aircraft displayed at a specific timestamp based on its latitude and longitude. This is beneficial to the research objective because one of the end goals is to visualize flight paths throughout the day and determine any suspicious aircrafts around the world.

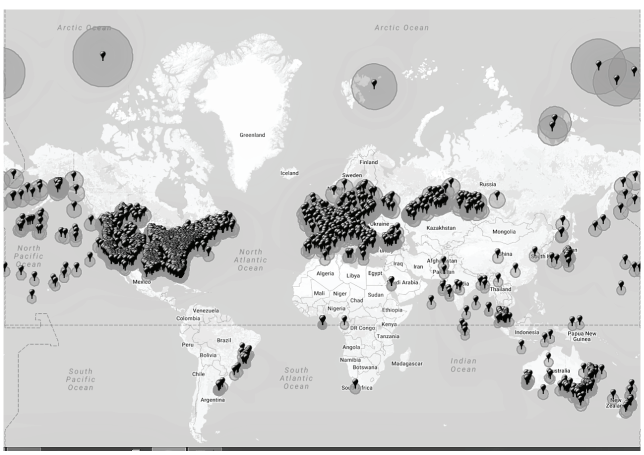
This application can be used for military personnel because each marker on the map includes the aircrafts' characteristics. For example, it indicates the type, country of registration, and manufacturer for each aircraft. This information can then be used to determine which aircrafts are considered allies or not. Additionally, since it is difficult for one person to track hundreds of aircrafts at once, reducing the data set to focus only on enemy planes can help alleviate the amount of planes that are shown.

Figures 4 and 5 represent some examples of the visualizations when using the Google Maps API.



[Figure 4] In the picture above, each black marker on the map indicates an aircraft that was being tracked at a specific time. This map focuses on the North American continent to provide a more concise representation of the data.

Furthermore, Figure 4 shows the thousands of aircrafts that were tracked through ADS-B ground radar stations in the United States. These aircrafts can range from commercial to military aircrafts from around the world.



[Figure 5] This screenshot represents a worldview of ADS-B data. In this picture, the black markers represent aircrafts and the circle around each one provides a rough estimate of where a radar station is located.

When compared to Figure 4, Figure 5 provides a radius of 200 nautical miles, or approximately 230 miles, to where the ground radar station is that picked up the aircraft. This was helpful during research because it highlights certain areas where there should be flight traffic, but do not show up in the ADS-B data. These findings are helpful because it documents where holes in the data are due to the lack of ADS-B radar station coverage. Although this is outside the scope of the research project, it can be used in a future research project to strategically place ground stations in areas without ADS-B coverage.

Conclusion

We showed the initial results of analyzing the ADS-B data with the following conclusion:

- The initial phase of determining the uses of ADS-B showed how the data can be processed through a Python script and run a Hadoop cluster for faster and more efficient results. The Hadoop cluster provided a single storage system for the hundreds of files that are processed each day.
- Lexical Link Analysis was proved to be an important part of the research because it helped detect anomalies within the data set by analyzing the aircrafts that had more associations and patterns in the data sets. It is an unsupervised learning tool since some of the patterns, associations and anomalies were previously unknown.
- Furthermore, we found that there were flights that only got detected by a radar station once and transmitted nonsensical position coordinates. In this case, some of the positions automatically sent by the aircraft are mapped to be in the middle of the ocean where there is not a ground radar station within range. This can be viewed as an outlier in the data that can then be further analyzed to understand what happened.
- We also showed that Google Maps was beneficial because it showed areas with and without coverage around the world. Google Maps indicated where the most traffic was at and have the ability to monitor traffic around certain areas.

Further research can be conducted to determine how to live-stream the ADS-B to process on a Hadoop cluster. This can help collect more data that can be used for Lexical Link Analysis and Google Maps and analyze the patterns and associations between aircrafts.

Overall, the research project was successful in providing results of how ADS-B data can be utilized for military or commercial use. When the kinematic and behavior patterns are discovered from historical ADS-B data, the resulted models can be also used to identify flying patterns and anomalies that can increase CTAP and the prediction accuracies of CID.

References

Collins, M. (2017, January 18). *First Space-Based ADS-B Satellites in Orbit*. Retrieved from AOPA: https://www.aopa.org/news-and-media/all-news/2017/january/18/first-space-based-ads-b-satellites-in-orbit

Dan. (n.d.). *ADS-B Exchange - World's largest co-op of unfiltered flight data*. Retrieved from ADS-B Exchange: https://www.adsbexchange.com/

Dan. (n.d.). *Data Field Descriptions*. Retrieved from ADS-B Exchange: https://www.adsbexchange.com/datafields/

Dan. (n.d.). *Serving the Flight Tracking Enthusiast*. Retrieved from https://www.adsbexchange.com/

Grush, L. (2015, January 30). *To Keep Airplanes On The Map, Track Them From Space*. Retrieved from Popular Science: http://www.popsci.com/tracking-airplanes-space-may-finally-become-reality

Hillenbrand, J. (2017, April 26). *Federal Aviation Administration and Aireon*. Retrieved from Aireon: https://aireon.com/2017/04/26/federal-aviation-administration-aireon-announce-space-based-ads-b-nextgen-technology-flight-test-success/

NextGen: In-Trail Procedures. (2016, September). Retrieved from Federal Aviation Administration: https://www.faa.gov/nextgen/snapshots/stories/?slide=52

Penchikala, S. (2015, January 30). *Big Data Processing with Apache Spark - Part 1: Introduction*. Retrieved from InfoQ: https://www.infoq.com/articles/apache-spark-introduction

Rouse, M. (2017, March). *Big Data Analytics*. Retrieved from Search Business Analytics - Tech Target: http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Schafer, M., Strohmeier, M., Lenders, V., Martinovic, I., & Wilhelm, M. (2014, April). *Bringing up OpenSky: A Large-scale ADS-B Sensor Network for Research*. Retrieved from Information Processing in Sensor Networks: https://www.cs.ox.ac.uk/files/7238/ipsn2014.pdf

Zhao, Y., MacKinnon, D. J., & Gallup, S. P. (2011, September). *System Self-Awareness and Related Methods for Improving the Use and Understanding of Data within DoD*. Retrieved from ASQ: http://asq.org/qic/display-item/index.html?item=33878

Zhao, Y., MacKinnon, D., & Gallup, S. (2015). *Big Data and Deep Learning for Understanding DoD Data.*

Zmarrou, H. (2015, April 16). *ADS-B Data Analysis*. Retrieved from RPubs: https://rpubs.com/zmarrou/73169