Breitengrad Regression eines Flugzeugs mithilfe von ADS-B und Maschinellen Lernen

Jan Zimbelmann[1]

Universität Leipzig
Machine Learning Group
Leipzig, Germany
pge13cfo@studserv.uni-leipzig.de

Zusammenfassung. Die Lokalisation von Flugzeugen ist essentiell für die Sicherung des Flugraumes. Für viele Flugzeuge ist das Automatic Dependent Surveillance–Broadcast (ADS-B), zum Zweck der Flugzeugüberwachung, verpflichtend. Hierbei wird ein Datensatz verschickt und von mehreren Sensoren empfangen. In dieser Arbeit wird dieser Datensatz vorverarbeitet und der Breitengrad rekonstruiert. Dazu werden zwei statistische, maschinelle Verfahren verwendet: Random Forest und Support Vector Regressor.

Schlüsselwörter: ADS-B, Maschinelles Lernen, Empirische Daten, Preprocessing, Random Forest, Support Vector Machine, Luftraumsicherung

1 Einleitung

ADS-B ist ein Datensatz zur Flugzeugüberwachung, welcher im Frequenzbereich von 1090 MHz automatisch übertragen wird. Für viele Flugzeuge ist es, unter gewissen Voraussetzungen [1], Pflicht, einen ADS-B Transponder installiert zu haben. Das ADS-B verschickt hierbei die Daten unverschlüsselt [2]. Es ist generell möglich, diese zu empfangen und zu verarbeiten. Darstellung hierzu in Abb. 1.

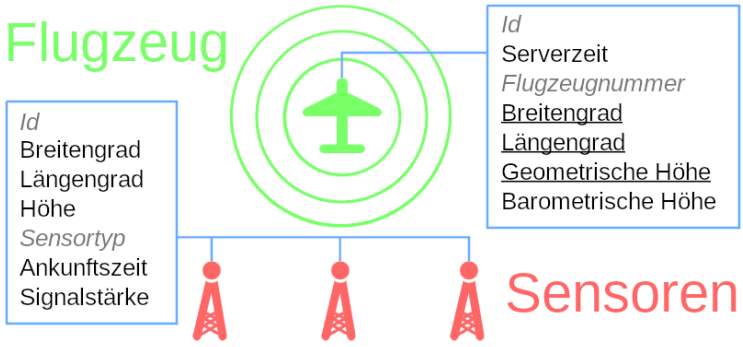

Abb. 1: Sensor und Flugzeug ADS-B Daten vom Opensky Network Datensatz

Die Position des Flugzeugs im ADS-B Datensatz basiert auf dem Global Positioning System (GPS). Diese Position kann beim Senden der Daten, aus technischen bzw. aufgrund schlechter Wetterbedingungen, nicht vorhanden oder fehlerhaft sein [2]. Ebenso kann der Datensatz durch einen Spoofing Angriff manipuliert worden sein [3].

Ziel dieser Arbeit ist es, mithilfe eines maschinellen Modells, den Breitengrad anhand eines vom Opensky Network zur Verfügung gestellten ADS-B Datensatz zu rekonstruieren. Es wird ein Verfahren vorgestellt, mit welchen der Random Forest und der Support Vector Regressor im Stande sind, eine Breitengrad Regression an dem ADS-B Datensatz durchzuführen. Beide Modelle werden mit zwei weiteren einfacheren Vorhersagemodellen verglichen, anhand zweier Metriken und mehreren Visualisierungen.

2 Stand der Technik

Die Federal Aviation Administration hat 2010 für den amerikanischen Luftraum, bei Flugzeugen unter bestimmten Bedingungen, den ADS-B als einen Standard für verpflichtend erklärt. Dieser ist bis 2020 nachzurüsten [1]. Diesbezüglich ergibt sich ein gesteigertes Interesse an der Verarbeitung großer Datensätze. Die Lokalisation von Flugzeugen mit anderen Methodiken, bspw. wie Radar, ist schon lange in Verwendung.

Die Lokalistation mit maschinellen Lernverfahren an den ADS-B Daten wurde bereits, mithilfe einer Support Vektor Machine und eines Deep Neural Networks, von Adesina et al. [2] untersucht. Hierbei kann die Position des Flugzeuges rekonstuiert werden. Die Genauigkeit hängt mit der Streuung der Breiten- und der Längengrade zusammen. Aufgrund dieser Korrelation, wird der Breitengrad genauer bestimmt als der Längengrad, da letzterer stärker gestreut ist.

Das Forschungsgebiet erstreckt sich weit über die Positionslokalisation hinaus. Däster et al. [4] haben mit Random Forests, Gradient Boost Trees und Multilayer Perceptrons eine Klassifikation von Flugzeugen in zivil oder militärisch erreichen können. Hierbei hat der Datensatz eine starke Schieflage von 49:1 der vorhandenen Daten, zugunsten der zivilen Daten. Dem wurde mit Oversampling beim Lernen entgegengewirkt, wobei der Fehler der beiden Kategorien somit stark vom Oversamplingverhältnis abhängt. Bei dieser Aufgabe ist allerdings eine Genauigkeit von 60% beider Kategorien bereits kein schlechtes Ergebnis.

3 Methodik

Jedes Signal im Datensatz wurde von 2 bis 13 Sensoren aufgenommen. Die Daten, die dadurch zur Verfügung stehen, können der Abb. 1 entnommen werden. Die Anzahl der Sensoren kommt hier noch als ein weiteres Feature hinzu. Der Datensatz enthält 2002847 Einträge, welche gemischt und dann im folgenden Kapitel verarbeitet werden. Im Anschluss werden die Maschinellen Modelle beschrieben und ihre Parameter offengelegt.

3.1 Datenvorverarbeitung

Um die Daten vorzuverarbeiten, werden zwei Theorien anhand von physikalischen Überlegungen präsentiert, anhand welcher das Lernen einer Position vorstellbar ist. Diese dienen lediglich als Grundlage für die Datenvorverarbeitung.

1. Eine Positionsbestimmung anhand von Multilateration, Beispiel: GPS
2. Eine Annahme und das Lernen von wiederholten Flugzeugrouten

Für eine Multilateration im dreidimensionalen Raum wäre es notwendig, die Position und Abstände von vier Sensoren zu haben. Rechts in Abb. 2 ist zu sehen, dass die Signalstärke eine große Varianz im bezüglich des Abstands hat. Somit kann die Serverzeit, als auch die Sensorzeit wichtig sein. Mit der Zeitdifferenz von Serverzeit und der Ankunftszeit an den Sensoren, kann bei einer elektromagnetischen Welle die Distanz ermittelt werden. Links in der Abb. 2 ist zu sehen, dass die barometrische Höhe stark mit der geometrischen korreliert. Somit ist die Position potentiell in einer Dimension, der Höhe, auflösbar. Daraus folgt ein zweidimensionales Problem und es ist als ausreichend angesehen, drei Sensoren zu berücksichtigen.

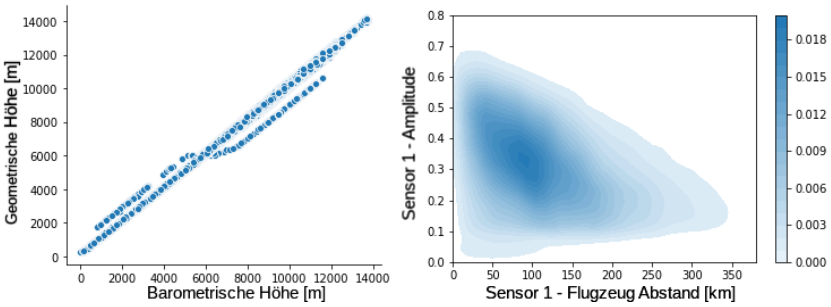

Abb. 2: Links: Korrelation zwischen der Geometrischen Höhe und der Barometrischen Höhe. Rechts: Korrelation zwischen dem Abstand vom Flugzeug zum ersten Sensor und der Signalstärke dieses Sensors.

Bei der Implementation ist jede Messung auf drei Sensoren mit den Sensortypen 'Radarcape' und 'GRX1090' gefiltert und reduziert. Letzterer zeigt noch wenige Ausreißer anhand der Amplitude, diese sind entfernt. Desweiteren ist die Amplitude beider Sensoren einzeln zwischen 0 und 1 normiert und im Anschluss sind Amplituden unterhalb von 0.02 ebenfalls entfernt. Die verwendeten Features sind mit der Farbe Schwarz in Abb. 1 gekennzeichnet. Die unterstrichenen Features sind die Zielparamter, wobei diese Arbeit sich nur auf den Breitengrad konzentriert. Somit ergibt sich mit jeweils 5 Features pro Sensor und 2 Features für das Flugzeug, eine Gesamtzahl von 17, plus einen Zielparameter. Ein Teildatensatz der Größe 100 000 wird im Verhältnis 3:2 in Train und Test umgewandelt. Somit soll eine eine ausreichende Größe zum Lernen garantiert werden. Im Anhang ist die Verteilung vom Train und Test Datensatz zu finden.

3.2 Maschinelle Lernmodelle

Bei dieser Arbeit wird sich auf zwei statistisch motivierte, maschinelle und überwachte Lernmodelle konzentriert: Random Forest (RF) und Support Vector Regressor (SVR). Grund für diese Modelle ist ein guter Umgang mit Daten mit einer hohen Anzahl an Features. Um die Genauigkeit dieser Modelle zu vergleichen, wird zum einen ein Dummy Estimator anhand des Mittelwerts verwendet und zum anderen ist ein Decision Tree (DT) implementiert.

Alle Modelle wurden mithilfe der Scikit Learn Bibliothek [5] implementiert.

- DT: $criterion : mse$, $maxdepth : -1$, $minsamplessplit : 2$
- RF: $estimators : 500$, $criterion : gini$, $maxdepth : -1$, $minsamplessplit : 2$
- SVR: $\gamma : 0.05$, $tol : 5e-4$, $C : 5$, $\epsilon : 0.01$, $cachesize : 400$

Diese Parameter wurden durch einfaches Trial and Error angepasst. Alle weiteren Parameter sind standardgemäß von Scikit Learn vorgegeben. Für den Support Vector Regressor wurden die Daten noch in einer Pipeline mit einer Standardskalierung verarbeitet.

4 Ergebnisse

Zu der Bewertung der Ergebnisse werden zwei Metriken verwendet:

- Mean Absolute Percentage Error (MAPE): $\frac{\sum |\frac{true-prediction}{true}|}{testsize} \cdot 100$
  kleine und große Fehler werden linear dargestellt
- Root Mean Squared Error (RMSE): $\sqrt{\frac{\sum (true-prediction)^2}{testsize}}$
  Abweichung der Daten werden berücksichtigt

Im Folgenden werden die beiden Fehlermetriken der einzelnen Methoden, sowie die Trainingsdauer gezeigt und der MAPE der einzelnen Methoden wird im Boxplot dargestellt:

RMSE	MAPE	time [s]	methods
1.900942	3.272851	<1.0	Mean
0.472857	0.618459	1.2	DT
0.324672	0.457171	330.0	RF
0.424596	0.605662	1111.9	SVR

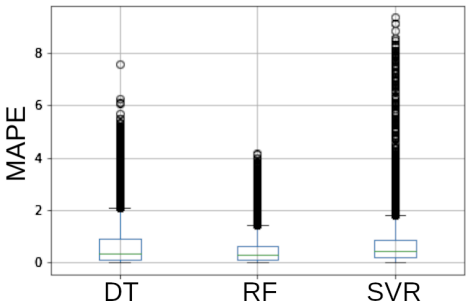

Abb. 3: Links: Fehlertabelle, Rechts: MAPE Boxplot

Die Breitengradvorhersage wurde anhand seines wahren Wertes für die beiden untersuchten Modelle visualisiert, sowie im Vergleich mit dem Dummy Estimator dargestellt. Des Weiteren wurden die Vorhersagen der gleichen Modelle, anhand weniger Datenpunkte beispielhaft verglichen, mit der Position der 3 Sensoren.

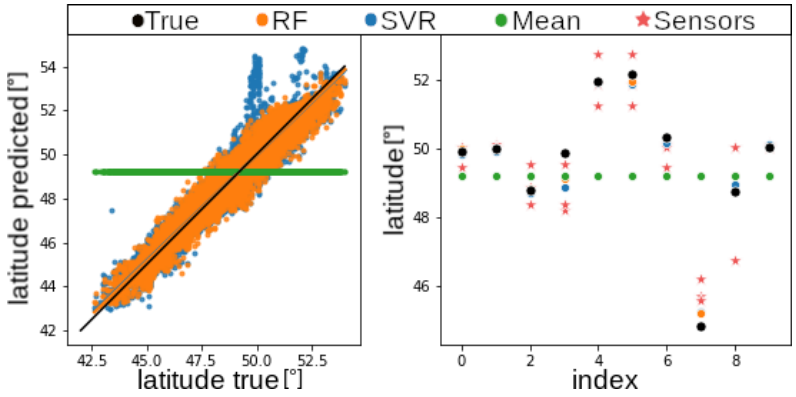

Abb. 4: Links: latitude true auf predicted plot, Rechts: Breitengradvorhersage und Sensor Position anhand von den ersten 10 Datenpunkten.

5 Diskussion

Die Fehlertabelle in Abb. 3 stellt dar, dass die beiden gewählten, maschinellen Modelle die Position besser als einen Mittelwert rekonstruieren können. Darüber hinaus sind beide ebenfalls besser als ein einfacher Entscheidungsbaum, wobei der SVR nur knapp besser abschneidet. Im Boxplot ist zu sehen, dass es beim SVR die größten Ausreißer gibt. Diese Ausreißer können links in der Abb. 4 auf der 'latitude true' Achse, bei ca. 49° herum beobachtet werden. Rechts ist zu sehen, dass die Modelle es berücksichtigen können, wenn der wahre Wert nicht zwischen den Sensoren liegt, wie es bei dem siebten Eintrag der Fall ist.

Die Modelle sind somit geeignet, eine Position vorherzusagen und es deutet darauf hin, dass die Modelle eine bestimmte Physik rekonstruieren können. Bei den beiden verwendeten statistischen Methodiken zeigt sich das Random Forest Modell als sehr brauchbar für eine Ortung des Flugszeuges anhand sehr großer ADS-B Datensätze. Der SVR weist allerdings kaum eine bessere Genauigkeit als der einfache Entscheidungsbaum auf und die Trainingszeit steigt wesentlich mit der Größe des Datensatzes. Weiterhin bleibt noch kritisch zu betrachten, dass der Datensatz sich auf einen Breitengrad um 47.5 ± 5 beschränkt und offen bleibt, ob somit das Modell universell anwendbar ist. Bei dieser Größenordnung lässt sich eine prozentuelle Abweichung in der ersten Nachkommastelle auf eine Abweichung von mehreren Kilometern umrechnen.

6 Schlussfolgerungen

Es konnte gezeigt werden, dass eine Positionsbestimmung des Breitengrads mithilfe der RF und der SVR prinzipell möglich ist. In Anbetracht der Zeiteffizienz und der Genauigkeit ist der RF auch noch für größere Datensätze geeignet. Der SVR skaliert mit den verwendeten Parameter nicht so gut mit der Datengröße.

Als alleinige Positionsbestimmung würde ein maschinelles Verfahren zur Flugsicherheit nicht ausreichen, da der Fehler im Kilometerbereich liegt. Die Anwendung könnte jedoch als eine Ergänzung zu den gängigen Methoden in Betracht gezogen werden. Dies ist der Fall, wenn die GPS Daten fehlen oder um einen ADS-B Datensatz auf seine Richtigkeit zu überprüfen ist, wie bei einem Spoofing Angriff.

7 Anhang

Die Verteilung der Flugzeug- und Sensorpositionen ist in der Abb. 5 als Boxplot dargestellt. Hierbei sind der Mittelwert und die Abweichungen ersichtlich.

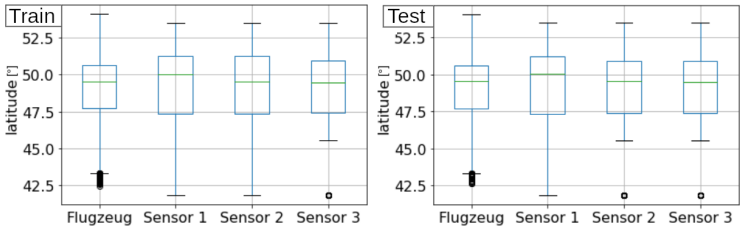

Abb. 5: Breitengrad Verteilung im Train (links) und Test (rechts) Datensatz

Literatur

1. F. A. Administration, "Revision to automatic dependent surveillance-broadcast (ads-b) out equipment and use requirements," *federalregister.gov*, pp. 1–8, 2019.
2. D. Adesina, O. Adagunodo, X. Dong, and L. Qian, "Aircraft location prediction using deep learning," in *MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM)*, pp. 127–132, 2019.
3. Z. Zhang, M. Trinkle, L. Qian, and H. Li, "Quickest detection of gps spoofing attack," in *MILCOM 2012 - 2012 IEEE Military Communications Conference*, pp. 1–6, 2012.
4. K. Dästner, E. Schmid, B. v. H. z. Roseneckh-Köhler, and F. Opitz, "Learning from ads-b data for real-time radar applications," in *2019 20th International Radar Symposium (IRS)*, pp. 1–10, 2019.
5. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.