IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.047 ⁜ Vol. 10, Issue 5, May 2022

DOI: 10.17148/IJIREEICE.2022.10584

FLIGHT DELAY PREDICTION USING MACHINE LEARNING

Sarah Ajmeria[1], Srushti V[2], Prof. Kavitha S Patil[3], S Haripriya[4]

Dept. Information Science Engineering, Atria Institute of Technology, Bangalore, India

Abstract: Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks. Compared with the previous schemes, the proposed random forest-based model can obtain higher prediction accuracy (90.2% for the binary classification) and can overcome the overfitting problem.

Keywords: Flight Delay Prediction, Random Forest

I. INTRODUCTION

Air traffic load has experienced rapid growth in recent years, which brings increasing demands for air traffic surveillance systems. Traditional surveillance technology such as primary surveillance radar (PSR) and secondary surveillance radar (SSR) cannot meet requirements of the future dense air traffic. Therefore, new technologies such as automatic dependent surveillance broadcast (ADS-B) have been proposed, where flights can periodically broadcast their current state information, such as international civil aviation organization (ICAO) identity number, longitude, latitude and speed.Compared with the traditional radar-based schemes, the ADSB-based scheme is low cost, and the corresponding ADS-B receiver (at 1090 MHz or 978 MHz) can be easily connected to personal computers . The received ADS-B message along with other collected data from the Internet can constitute a huge volume of aviation data by which data mining can support military, agricultural, and commercial applications. In the field of civil aviation, the ADS-B can be used to increase precision of aircraft positioning and the reliability of air traffic management (ATM) systems . For example, malicious or fake messages can be detected with the use of multilateration (MLAT) , allowing open, free, and secure visibility to all the aircrafts within airspace . Thus, the ADS-B provides an opportunity to improve the accuracy of flight delay prediction which contains great commercial value.

The flight delay is defined as a flight taking off or arriving later than the scheduled time, which occurs in most airlines around the world, costing enormous economic losses for airline companies, and bringing huge inconvenience for passengers. According to the civil aviation administration of China (CAAC), 47.46% of the delays are caused by severe weather, and 21.14% of the delays are caused by air route problems. Due to the airline company or technical problems, air traffic control and other reasons account for 2.31% and 29.09%, respectively. Recent studies have been focused on finding a suitable way to predict the probability of flight delay or delay time to better apply air traffic flow management (ATFM) to reduce the delay level. Classification and regression methods are two main ways for modeling the prediction model. Among the classification models, many recent studies applied machine learning methods and obtained promising results .

A public dataset named VRA was used to compare the performance of several machine learning models including k-nearest neighbors (K-NN) , support vector machines (SVM) , naive Bayes classifier, and random forests for predicting flight delay, and achieved the best accuracy of 78.02% among the four schemes. However, the air route information (e.g., traffic flow and size of each route) was not considered in their model, which prevents them from obtaining higher accuracy. D. A. Pamplona et al. built an artificial neural network with 4 hidden layers, and achieved the highest accuracy of 87%; their proposed model suggested that the day of the week, block hour, and route has great influence on the flight delay. This model did not consider meteorological factors, so there is room for improvement. Y. J. Kim et al. proposed a model with two stages.The first stage is to predict day-to-day delay status of a specific airport by using a deep RNN model, where the status was defined as an average delay of all flights arriving at each airport. The second stage is a layered neuron network model to predict the delay of each individual flight using the day-to-day delay status from the first stage and other information. The two stages of the model achieved accuracies of 85% and 87.42%, respectively. This study suggested that the deep learning model requires a great volume of data. Otherwise, the model is likely to end up with poor performance or over fitting.

DOI: 10.17148/IJIREEICE.2022.10584

II. LITERATURE SURVEY:

In [1], Automatic dependent surveillance-broadcast (ADS-B) is an air traffic control system in which aircraft transmit their own information (identity, position, velocity etc.) to ground sensors for surveillance scope. The tracking of the different sensors' clocks by the use of time difference of arrival of ADS-B messages is proposed to check the veracity of the position information contained in the ADS-B messages. The method allows detecting possible on-board anomalies or the malicious injection of fake messages (intrusion) without the use of the multilateration (or any other) location algorithm. It follows that it does not need the inversion of the location problem (usually strong nonlinear and ill-posed), and, contrary to the multilateration, it works also with less than four sensors.

In [2] A novel air-to-ground (ATG) communication system, which is based on adaptive modulation and beam forming enabled by automatic dependent surveillance-broadcast (ADS-B) and multilateration techniques, is presented in this paper. From an aircraft geolocation perspective, the proposed multilateration technique uses the time-difference-of-arrival (TDOA), angle-of-arrival (AOA), and frequency-difference-of-arrival (FDOA) features within the ADS-B signal to implement the hybrid geolocation mechanism. Moreover, this hybrid mechanism aims for the optimal selection of multilateration sensors to provide a precise aircraft geolocation estimate by minimizing the geometric dilution-of-precision (GDOP) metric and imparts significant resilience to the current ADS-B-based geolocation framework to withstand any form of attack involving aircraft impersonation and ADS-B message infringement. From an ATG communication perspective, the ground base stations can use this hybrid aircraft geolocation estimate to dynamically adapt their modulation parameters and transmission beampattern in an effort to provide a high-data-rate secure ATG communication link. Additionally, we develop a hardware prototype that is highly accurate in estimating AOA data and facilitating TDOA and FDOA extraction associated with the received ADS-B signal. This hardware setup for the ADS-B-based ATG system is analytically established and validated with commercially available universal software-defined radio peripheral units. This hardware setup displays 1.5° AOA
These days we tend to face such a large amount of issues in agriculture fields, concerning irrigation and maintaining the rate of productivity. These requirements have found many issues, such as lack of communication systems and the massive distances to electricity offer points. To beat these issues, estimation accuracy, whereas the simulated geolocation accuracy is approximately 30 m over 100 nautical miles for a typical aircraft trajectory. The adaptive modulation and beamforming approach assisted by the proposed GDOP-minimization-based multilateration strategy achieves significant enhancement in throughput and reduction in packet error rate.

[3] With the growth of air transport, the air traffic control needs to enforce the Communication navigation surveillance air traffic management (CNS-ATM) because this is the back bone of the air operation in any country. This system has the responsibility of guaranteeing air safety and management of the national air space (NAS) that nowadays needs to increase the flight density to respond to the demand. To accomplish this, new technologies like air dependent surveillance broadcast (ADS-B) have been used to increase the accuracy and time response of data air surveillance sensor integration of sensor location and the reliability of ATM systems. CNS-ATM systems for surveillance and control of aircrafts have been mainly used in primary and secondary radars to calculate the aircraft position through signal delay or time difference between transponder pulses. The accuracy of each sensor depends on internal and external factors such as frequency, power, target distance, noise, maintenance, and others. When an aerodyne is detected by multiple sensors, it could create a multiple track in a geographic and temporal space where the aircraft will be possibly flying. This space depends on radar update time, aerodyne speed, and the accuracy of each sensor, and it is difficult to know where the aircraft really is. This work proposes a technique based on ADS-B for making an error calculation of each sensor in a fusion system, using business intelligence techniques for understanding the error condition of each sensor in a geographical area. Based on results, we propose a technique that could make an error correction to avoid phase shifts between sensors. The information of this data study was used for statistical calculation values such as variance and standard deviation. For fusion accuracy improvement, three steps have been proposed in this research. First, the use of the radar error by region and statistical values by calculating the Kalman filters for each sensor to reduce the internal.

[4] Air delay is a problem in most airports around the world, resulting in increased costs for airlines and discomfort for passengers. Air Traffic Flow Management (ATFM) programs were implemented with the main objective to reduce the delay levels in the whole air transportation sector. The question is to find a suitable way to predict possible delay scenarios to better apply ATFM measures. The present work seeks to enrich the academic literature on the subject and aims to present the application of Artificial Neural Networks (ANN) to a prediction model of delays in the air route between São Paulo (Congonhas) - Rio de Janeiro (Santos Dumont). The configuration of ANN exerts a great influence on its predictive power. To better adjust the parameters of the proposed ANN and for the hyper parameterization of the

DOI: 10.17148/IJIREEICE.2022.10584

network to occur, the Random Search technique is used. By using the recall, precision and F Score metrics in the performance measurement, the prediction results show satisfactory results in the case study.

[5] Supervised machine learning algorithms have been used extensively in different domains of machine learning like pattern recognition, data mining and machine translation. Similarly, there have been several attempts to apply the various supervised or unsupervised machine learning algorithms to the analysis of air traffic data. However, no attempts have been made to apply Gradient Boosted Decision Tree, one of the famous machine learning tools to analyze those air traffic data. This paper investigates the effectiveness of this successful paradigm in the air traffic delay prediction tasks. By combining this regression model based on the machine learning paradigm, an accurate and sturdy prediction model has been built which enables an elaborated analysis of the patterns in air traffic delays. Gradient Boosted Decision Tree has shown a great accuracy in modeling sequential data. With the help of this model, day-to-day sequences of the departure and arrival flight delays of an individual airport can be predicted efficiently. In this paper, the model has been implemented on the Passenger Flight on-time Performance data taken from the U.S. Department of Transportation to predict the arrival and departure delays in flights. It shows better accuracy as compared to other methods.

[6] Flight delays cause various inconveniences for airlines, airports, and passengers. According to data provided by the Brazilian National Civil Aviation Agency (ANAC), between 2009 and 2015, about 22% of domestic flights made in Brazil were delayed by more than 15 minutes. The prediction of these delays is fundamental to mitigate their occurrence and optimize the decision-making process of an air transport system. Particularly, airlines, airports, and users may be more interested in when delays are likely to occur than the accurate prediction of the absence of delays. This paper focuses on the unbalanced distribution of the classes of delay (presence and absence) by performing an experimental evaluation of several preprocessing methods for the development of machine-learning flight delay classification models. Those models were built from a dataset that integrates national flight operations with meteorological conditions of airports. Our results indicate the models that applied the balancing techniques performed much better in predicting the occurrence of delays, getting about 60% of hits.

In [7] This paper presents a new class of models for predicting air traffic delays. The proposed models consider both temporal and spatial (that is, network) delay states as explanatory variables, and use Random Forest algorithms to predict departure delays 2–24 h in the future. In addition to local delay variables that describe the arrival or departure delay states of the most influential airports and links (origin–destination pairs) in the network, new network delay variables that characterize the global delay state of the entire National Airspace System at the time of prediction are proposed. The paper analyzes the performance of the proposed prediction models in both classifying delays as above or below a certain threshold, as well as predicting delay values. The models are trained and validated on operational data from 2007 and 2008, and are evaluated using the 100 most-delayed links in the system. The results show that for a 2-h forecast horizon, the average test error over these 100 links is 19% when classifying delays as above or below 60 min. Similarly, the average over these 100 links of the median test error is found to be 21 min when predicting departure delays for a 2-h forecast horizon. The effects of changes in the classification threshold and forecast horizon on prediction performance are studied.

In [8] High arrival delay at major airports tends to propagate and generate secondary delay through the National Airspace System (NAS). In the United States, it is widely believed that the major culprits for delay throughout the NAS are the three New York commercial airports – Newark (EWR), LaGuardia (LGA), and John F. Kennedy (JFK). Various estimates of the extent to which the New York airports impact the delay in the NAS have been reported over the years. Yet there is no thorough investigation into the mutual relationship between delays at New York and non-New York airports. In this paper, we take two different approaches to quantify the impact of the three New York airports on delay throughout the NAS. First, we estimate and apply an econometric model using a large historical dataset. The other model is the FAA SWAC model that simulates flights and tracks the daily performance of the system. The counterfactual scenarios in these two models are adjusted to be comparable to each other. There is disparity between the results of the two different models, suggesting the simulation model might not capture all the factors that cause arrival delay. Still both results conclude that the portion of delay in the system caused by New York airports is much less than publicized estimates. Combining econometric and simulation models to address questions of this nature appears to be a promising approach.

In [9], K nearest neighbor (kNN) method is a popular classification method in data mining and statistics because of its simple implementation and significant classification performance. However, it is impractical for traditional kNN methods to assign a fixed k value (even though set by experts) to all test samples. Previous solutions assign different k values to different test samples by the cross validation method but are usually time-consuming. This paper proposes a

DOI: 10.17148/IJIREEICE.2022.10584

kTree method to learn different optimal k values for different test/new samples, by involving a training stage in the kNN classification. Specifically, in the training stage, kTree method first learns optimal k values for all training samples by a new sparse reconstruction model, and then constructs a decision tree (namely, kTree) using training samples and the learned optimal k values. In the test stage, the kTree fast outputs the optimal k value for each test sample, and then, the kNN classification can be conducted using the learned optimal k value and all training samples. As a result, the proposed kTree method has a similar running cost but higher classification accuracy, compared with traditional kNN methods, which assign a fixed k value to all test samples. Moreover, the proposed kTree method needs less running cost but achieves similar classification accuracy, compared with the newly kNN methods, which assign different k values to different test samples. This paper further proposes an improvement version of kTree method (namely, k*Tree method) to speed its test stage by extra storing the information of the training samples in the leaf nodes of kTree, such as the training samples located in the leaf nodes, their kNNs, and the nearest neighbor of these kNNs. We call the resulting decision tree as k*Tree, which enables to conduct kNN classification using a subset of the training samples in the leaf nodes rather than all training samples used in the newly.

In [10] Modern intelligent transport systems focus on the integration of multiple sensors to obtain hybrid navigation schemes. A key issue of a hybrid scheme is distribution of the information sharing coefficients (ISCs) of subsystems and the fusion of parallel multiple observations of navigation sensors. Recently, deep learning methods, particularly convolutional neural networks (CNNs), have achieved great success in image processing tasks. However, there has been limited work in using deep learning for multisensor-based integrated navigation solutions. In this letter, we propose an ensemble learner-based classification and information fusion method, in which estimation error covariance matrices provided by local adaptive filters are used as input for the classifier, and the triple numbers of ISCs are determined by the proposed scheme. The results validate the effectiveness of the proposed scheme, in which the adequately trained ensemble learner can detect the degradation of a subsystem that may suffer atypical observations or faults and consequently can adjust the corresponding ISC in real time.

III. SYSTEM ARCHITECTURE DIAGRAM

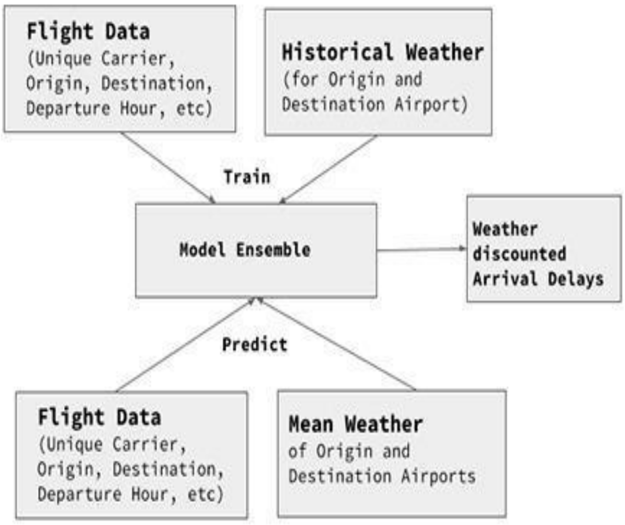
Flight Data
(Unique Carrier, Origin, Destination, Departure Hour, etc)
Historical Weather
(for Origin and Destination Airport)
Train
Model Ensemble
Weather discounted Arrival Delays
Predict
Flight Data
(Unique Carrier, Origin, Destination, Departure Hour, etc)
Mean Weather
of Origin and Destination Airports

Figure 1: Flight Delay Prediction Architecture Diagram

IV. CONCLUSION

In summary, the random forest-based architecture presented better adaptation at a cost of the training accuracy when handling the limited dataset. In order to overcome the over-fitting problem and to improve the testing accuracy for multi-categories classification tasks, our future work will focus on collecting or generating more training data, integrating more information like airport traffic flow, airport visibility into our dataset.

V. EXPECTED OUTPUT

DOI: 10.17148/IJIREEICE.2022.10584

We predict the delay of the flight by giving the parameters such as arrival and departure time, the origin and destination airport details, the weekday of the departure, aircraft type and manufacturer details. Based on the training data, the model will predict how many minutes the flight might be
delayed.

VI. REFERENCE PAPERS

[1] M. Leonardi, “Ads-b anomalies and intrusions detection by sensor clocks tracking,” IEEE Trans. Aerosp. Electron. Syst., to be published, doi:10.1109/TAES.2018.2886616.

[2] Y. A. Nijsure, G. Kaddoum, G. Gagnon, F. Gagnon, C. Yuen, and R. Mahapatra, “Adaptive air-to-ground secure communication system based on ads-b and wide-area multilateration,” IEEE Trans. Veh. Technol., vol. 65, no. 5, pp. 3150–3165, 2015.

[3] J. A. F. Zuluaga, J. F. V. Bonilla, J. D. O. Pabon, and C. M. S. Rios, “Radar error calculation and correction system based on ads-b and business intelligence tools,” in Proc. Int. Carnahan Conf. Secur. Technol., pp. 1–5, IEEE, 2018.

[4] D. A. Pamplona, L. Weigang, A. G. de Barros, E. H. Shiguemori, and C. J. P. Alves, “Supervised neural network with multilevel input layers for predicting of air traffic delays,” in Proc. Int. Jt. Conf. Neural Networks, pp. 1–6, IEEE, 2018.

[5] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, “A statistical approach to predict flight delay using gradient boosted decision tree,” in Proc. Int. Conf. Comput. Intell. Data Sci., pp. 1–5, IEEE, 2017.

[6] L. Moreira, C. Dantas, L. Oliveira, J. Soares, and E. Ogasawara, “On evaluating data preprocessing methods for machine learning models for flight delays,” in Proc. Int. Jt. Conf. Neural Networks, pp. 1–8, IEEE, 2018.

[7] J. J. Rebollo and H. Balakrishnan, “Characterization and prediction of air traffic delays,” Transp. Res. Part C Emerg. Technol., vol. 44, pp. 231– 241, 2014.

[8] L. Hao, M. Hansen, Y. Zhang, and J. Post, “New york, new york: Two ways of estimating the delay impact of new york airports,” Transp. Res. Part ELogist. Transp. Rev., vol. 70, pp. 245–260, 2014.

[9] ANAC, “The Brazilian National Civil Aviation Agency.” anac.gov, 2017. [online] Available:http://www.anac.gov.br/.

[10] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient knn classification with different numbers of nearest neighbors,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 5, pp. 1774–1785, 2017.

[11] J. Sun, Z. Wu, Z. Yin, and Z. Yang, “Svm-cnn-based fusion algorithm for vehicle navigation considering atypical observations,” IEEE Signal Process. Lett., vol. 26, no. 2, pp. 212–216, 2018.

[12] Y. J. Kim, S. Choi, S. Briceno, and D. Mavris, “A deep learning approach to flight delay prediction,” in Proc. Digit. Avion . Syst. Conf., pp. 1–6, IEEE, 2016.

[13] Y. Cong, J. Liu, B. Fan, P. Zeng, H. Yu, and J. Luo, “Online similarity learning for big data with overfitting,” IEEE Trans. Big Data, vol. 4, no. 1, pp. 78–89, 2017.

[14] F. Tang, Z. M. Fadlullah, B. Mao, and N. Kato, “An intelligent traffic load prediction-based adaptive channel assignment algorithm in sdn-iot: A deep learning approach,” IEEE Internet Things J., vol. 5, pp. 5141– 5154, Dec 2018.

[15] N. Kato, Z. M. Fadlullah, B. Mao, F. Tang, O. Akashi, T. Inoue, and K. Mizutani, “The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective,” IEEE Wireless Commun., vol. 24, pp. 146–153, June 2017.

[16] J. Wang, J. Liu, and N. Kato, “Networking and communications in autonomous driving: A survey,” IEEE Commun. Surveys Tuts., vol. 21, pp. 1243–1274, April 2019.

[17] Y. Kawamoto, H. Nishiyama, N. Kato, F. Ono, and R. Miura, “Toward future unmanned aerial vehicle networks: Architecture, resource allocation and field experiments,” IEEE Wireless Commun., vol. 26, pp. 94–99, February 2019.