A CNN-LSTM framework for flight delay prediction

Abstract

Flight delay prediction has become one of the most critical topics in intelligent aviation systems due to its essential role in flight scheduling, airline planning, and airport operation. The accurate prediction of flight delays is very challenging because numerical factors will affect flight delays. Moreover, owing to the connectivity of the aviation system, flight delays also present complex spatial–temporal correlations, including the spatial correlations between airports along with the temporal correlations among timestamps. To address these challenges, we proposed a CNN-LSTM deep learning framework to consider the spatial–temporal correlations together with the extrinsic features for flight delay prediction. The CNN-LSTM model consists of a Convolution neural network (CNN) architecture to learn the spatial correlations followed by a Long short-term memory (LSTM) architecture to capture the temporal correlations. The spatial–temporal correlations obtained from the CNN-LSTM framework are then fused with the extrinsic features (e.g., airline issues, distance, schedule fly time, etc.) as inputs of the random forest (RF) model for flight delay prediction. The U.S. domestic flights in 2019 are collected from the Bureau of Transport Statistics to confirm the outperformance of the proposed model. The results show that the accuracy of the CNN-LSTM model reaches 92.39%. For the on-time samples, approximately 91% are correctly identified; for the delayed samples, the classification accuracy reaches 84% which exhibits better performance compared with several benchmark models. The created prediction model of this study could provide useful information for airport regulators in understanding the potential delays in advance and developing effective airport management strategies for improving airport on-time performance.

Introduction

The International Air Transport Association (IATA) reported that the annual compound average growth rate of air traffic demand is roughly 3.7%, and it is predicted that there will be up to 7.2 billion travelers globally in 2035, nearly double the 3.8 billion passengers who flew in 2016 (IATA, 2016). With limited airport expansion possibilities, the tremendous growth of the civil aviation industry will put a strain on the air transportation system and raise flight delays (Wandelt and Sun, 2015; Zhang and Mahadevan, 2019). For example, the average yearly delay rate for commercial flights in the US from 2016 to 2019 was around 20%, and the Federal Aviation Administration (FAA) report projects that flight delays will incur economic expenses of 33 billion in 2019 in comparison to 23.7 billion in 2016 (FAA, 2020). In addition to the direct financial expenses, flight delays will also raise customers' complaints and lead to disputes between customers and airlines and decrease the air traffic system's operational effectiveness (Wu, 2008). Due to these negative effects, flight delay prediction was proposed to track delays in real-time, improve flight planning to greatly reduce flight delays, and increase airport productivity (Thiagarajan et al., 2017).

In the last decades, numerous techniques such as statistical analysis and probabilistic models were applied for flight delay prediction (Wang et al., 2008, Zou and Hansen, 2014). However, these models struggle to deal with high-dimensional data and extract non-linear correlations (Yu et al., 2019). With the rapid development of machine learning techniques, especially their

benefits in managing huge data and identifying non-linear correlations (Kim et al., 2016, Güvercin et al., 2020), researchers have applied machine learning approaches, such as random forest (Rebollo and Balakrishnan, 2014, Li and Jing, 2021a, Li and Jing, 2021b), Deep belief network (DBN) (Yu et al., 2019), and Long short-term memory (LSTM) (Gui et al., 2019, Li and Jing, 2022) in flight delay prediction and have achieved great success.

Making an accurate and trustworthy prediction is challenging due to the complexity of the causes of flight delays. Although existing works have tried to consider a wide range of potential variables such as seasonal effects (Tu et al., 2008, Kim et al., 2016), airport infrastructure (Yu et al., 2019, Lambelho et al., 2020, Rodríguez-Sanz et al., 2021), operational times (Abdel-Aty et al., 2007, Belcastro et al., 2016), aircraft properties (Yu et al., 2019, Lambelho et al., 2020), and weather conditions (Pamplona et al., 2018, Li and Jing, 2022) to achieve high accuracy prediction of flight delays. However, airports in the aviation system are extensively connected rather than isolated, which indicates flight delays are not only related to its operated airport but also associated with its nearby airports (Jetzki, 2009). For example, an adverse weather event not only causes flight delays in a particular airport but will also propagate the delays into downstream airports due to the connected resources (e.g., aircrafts, crew members, and passengers) (Li and Jing, 2021a, Li and Jing, 2021b). In addition to delay propagation, the aviation system also experiences congestion spillover effects, whereby the congestion at local airports will affect the nearby areas (Redondi and Gudmundsson, 2016). Therefore, the congestion at a particular airport would not only increase the flight delays on it but would also spill the congestion and flight delays to its neighboring airports, which indicates a strong spatial correlation exists among nearby airports (Du et al., 2018). Moreover, future flight delays are related to both past and current flight delays and congestion (Pyrgiotis et al., 2013). An obvious example is that the same aircraft flies numerous flight legs in a day and the delay of one trip might affect the subsequent flights of the same aircraft (Kafle and Zou, 2016). When the airport is congested, the current operational capacity of the airport may not be able to arrange all flights to take off and land normally during that period, which could cause a backward shift in demand and further impair the smooth operation of subsequent flights (Zhang et al., 2019), indicating a strong temporal correlation. Except for the spatial and temporal correlations, some external factors, such as airline issues (Lambelho et al., 2020), and weather conditions (Li and Jing, 2022) will also affect the performance of the flight delay prediction model.

Despite the tremendous advancement and positive outcomes achieved by the current delay prediction models, existing works focus on considering more external features (e.g., flight schedule (Belcastro et al., 2016), surveillance-broadcast (ADS-B) messages (Gui et al., 2019), turnaround time (Guo et al., 2021), network topology properties (Li and Jing, 2022)) but still showing certain limitations due to the inherent limitations of spatial–temporal correlations. First, previous works have paid attention to the spatial features of the aviation network and employed the network theory (Li and Jing, 2022; Güvercin et al.,2021) and the k-means model (Rebollo and Balakrishnan, 2014, Gopalakrishnan and Balakrishnan, 2017) to extract the network effects. However, spatial features are not spatial correlations. The spatial features such as degree and betweenness centrality represent the degree of congestion of an airport and the level of independence of airports from each other, respectively. Although these variables can explain the characteristics of delays at the spatial level, they still cannot effectively indicate the specific amount of airports affected by other airports, i.e., whether flight delays are relevant at

the spatial level. Second, temporal characteristics are important for flight delay prediction (Abdel-Aty et al., 2007). Existing works have considered the temporal characteristics, such as weather conditions or congestion of airports at different times before flight departure or arrival (Belcastro et al., 2016, Gopalakrishnan and Balakrishnan, 2017, Li and Jing, 2021a, Li and Jing, 2021b). For example, Belcastro et al. (2016) considered the weather conditions at the origin airport from departure time back to 12 h before for flight delay prediction. However, incorporating those features into a prediction model not only increases the complexity of the model but also may hurt prediction accuracy due to the temporal correlation among features (Yu et al., 2019). Understanding the temporal correlations can therefore significantly enhance the effectiveness of data augmentation and prediction accuracy.

To solve the aforementioned issues, in this work, we propose a novel deep learning architecture called CNN-LSTM-Random Forest to deal with flight delay prediction in the context of extracting the spatial–temporal correlations. The CNN-LSTM-Random Forest is a two-stage model which contains an LSTM component, a joint Convolutional Neural Network (CNN) and the LSTM component, a feature fusion layer, and a random forest classifier. In the first stage, the CNN-LSTM-Random Forest learns the temporal correlations and the spatial–temporal correlations with an LSTM architecture and a CNN-LSTM architecture. The outputs are then fused with the external features as the inputs of random forest for flight delay prediction. The main contributions of this work are summarized as follows:

First, we map the airports to the map using each airport's geographic information and then use a grid-based segmentation method to create the spatial distribution information of flight delays and congestion of each airport in the aviation system. Second, we propose a novel framework based on CNN and LSTM units to extract the spatial–temporal correlations. The framework consists of two components: The LSTM-based architecture and the CNN-LSTM-based architecture, where the LSTM-based architecture is used to capture the temporal effect of the recent and current weather conditions on flight delays, and the CNN-LSTM-based architecture is used to capture the spatial–temporal effects of the recent flight delays and congestion on flight delays. Third, we conduct an extensive set of experiments on the U.S. domestic flights collected from the Bureau of Transportation Statistics (BTS) to validate our proposed model, and the results indicate the CNN-LSTM-Random Forest model outperforms the benchmark model, including the model without considering the spatial–temporal dependencies, only considering the spatial dependencies, considering temporal dependencies and both considering the spatial–temporal dependencies.

The rest of the article is structured as follows. The related research on the spatial–temporal determinants of flight delays and flight delay prediction is covered in Section 2. Section 3 introduces the data in detail and the proposed CNN-LSTM-Random Forest model. Section 4 reports the experimental findings. Section 5 highlights the limitations and upcoming projects.

Section snippets

Spatial-temporal dependencies of flight delays

Congestion is one of the major factors for flight delays at the busiest airports (Jacquillat & Odoni, 2015), and existing research has proven the existence of spatial and temporal correlations between airport delays and congestion (Cai et al., 2021, Shao et al., 2022). For example, Redondi and Gudmundsson (2016) explored the congestion spillover effects of Heathrow and Frankfurt airports on connection traffic and discovered that the congestion in local areas would spread to other regions. The

Data and methodology

In this section, we mainly introduce the data, associated features, and the proposed model named CNN-LSTM-Random Forest for flight delay prediction.

Experimental results

In this section, we mainly introduce the input features, evaluation metrics, and benchmark models and present the results of the proposed CNN-LSTM-Random Forest model in terms of prediction performance, and investigate the impact of the CNN architectures, and the input features on prediction accuracy.

Conclusions

Due to the ongoing operation of the aircraft and the connection of the airline network, relevant features such as airport delays and congestion present strong spatial–temporal correlations. The spatial–temporal properties of airports have posed great challenges to flight delay prediction. In this work, we present a CNN-LSTM-based architecture that could simultaneously consider the influence of spatial–temporal correlations for flight delay prediction. A grid-based segmentation is conducted to

CRediT authorship contribution statement

Qiang Li: Conceptualization, Methodology, Software, Writing – review & editing. Xinjia Guan: Methodology, Software, Writing – review & editing. Jinpeng Liu: Methodology, Software, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.