Prediction of the Flight Delay Based on Neural Network and Support Vector Machine: A Comparative Study

Yuxin Huang

Tan Kah Kee College of Xiamen University, Xiamen, 363105, China

ABSTRACT

This paper training a model which aims to predict the flight's delay based on the ANN and SVM , which are current mainstream machine learning models. We use the dataset of the flight delay from Kaggle and divided it as three parts: 0.7:0.15:0.15, for training dataset, validation and test dataset. The mean relative error of the prediction model are 18.6% and 21.2% by ANN and SVM, and the time cost of predicting a single roughness value based on cutting parameters are about 100.4ms and 58.5ms with ANN and SVM. Based on the experimental results, these two models can accurately and efficiently predict the delay and the ANN model have a little better accuracy than the SVM.

KEYWORDS

Flight delay; Neural network; Support vector machine

DOI: 10.47297/taposatWSP2633-456910.20230402

1 Introduction

Commercial airlines understand delay as the time a flight is delayed. Therefore, delay can be expressed as the difference between the scheduled and actual time when an aircraft takes off or arrives. National regulators have many indicators related to flight delay tolerance thresholds. In fact, flight delays are an important topic in the air transportation system.

Flight delays have serious influence to various aspects of the aviation industry, mainly economically. When this happens, In order to arrive on time, passengers often plan to show up to their appointments hours in advance. But this behavior will lead to an increase in time costs. For the Airlines also be influence by additional operating costs, such as crew and aircraft being held up at airports, they also need to pay for the fines and suffer penalties. Furthermore, delays will increasing fuel consumption and gas emissions that can damage the environment.

Delays also have big influence on the airline marketing strategies. Airlines' frequent-flyer programs need the customer loyalty, but the decrease of the dependability will affectedly influence the consumer's choice. Estimates of flight delays can further airport and airline tactical and operational decisions, meanwhile, it can alert passengers that make them have opportunity to reschedule their plans. Commercial aviation collects large amounts of data every year and stores them in databases, with this data, they can better understand the entire flight ecosystem. In this context, data modeling is a trend to solve challenging problems related to big data.

In the literature [1], they compare several machine learning by using a public dataset named VRA (http://www.anac.gov.br/) to predict the flight delays, including k-nearest neighbor (K-NN) [2], support vector machine (SVM) [3], naive Bayesian classifier and random forest. The best best accuracy rate they achieved is 78.02%. However, route information (eg, traffic volume and size for each route) was not considered. Literature [4] established an artificial neural network to predict flight delays. They use a 4 hidden layers network and get the accuracy rate of 87%. Greater impact. However, this model does not take meteorological factors into account.

Therefore, this article will consider as many factors as possible that affect the delay time. Using the factor on the flight delay prediction model. Based on the current mainstream machine learning models: BP neural network and support vector machine for training, selecting the best predictive model for analyzing which factors most affect flight delays.

2 Methods

(1) Data collection

The dataset of the flight delay from Kaggle (https://www.kaggle.com/datasets/usdot/flight-delays?resource=download) are selected for the prediction model. The possible attributes influencing the delay includes scheduled departure time, scheduled arrival time, weather conditions, air temperature, departure location, arrival location, months, etc. We divided the dataset as three parts:0.7:0.15:0.15 as training dataset, validation dataset and test dataset.

(2) Data reprocessing

The neural network focus on the number calculation. Therefore, some nonnumerical attributes will be transferred to number values. Specifically, the longitude and latitude value will be used to denote the location of the departure and arrival location. Other attributes including departure and arrival time, and date which are originally denoted by number can be used directly as the input the prediction model. More details can be seen in Table 1.

Table 1. The example data reprocessing

Attributes	Month	Day of week	Scheduled Departure time	Scheduled Arrival time	Departure location		Arrival location	
					PEK		SH	
Original	6	4	1530	1729	longitude	latiude	longitude	latiude
Reprocessed	6	4	1530	1729	120	30	89	98

(3) Model training

1) Neural network

The structure ANN (artificial neural network) model can be seen in Figure 1(a). Neurons of input, hidden and output layer layers are full-connected. The calculation process of each neuron can be seen Figure 1(b). The training process of ANN model is shown in Figure 2.

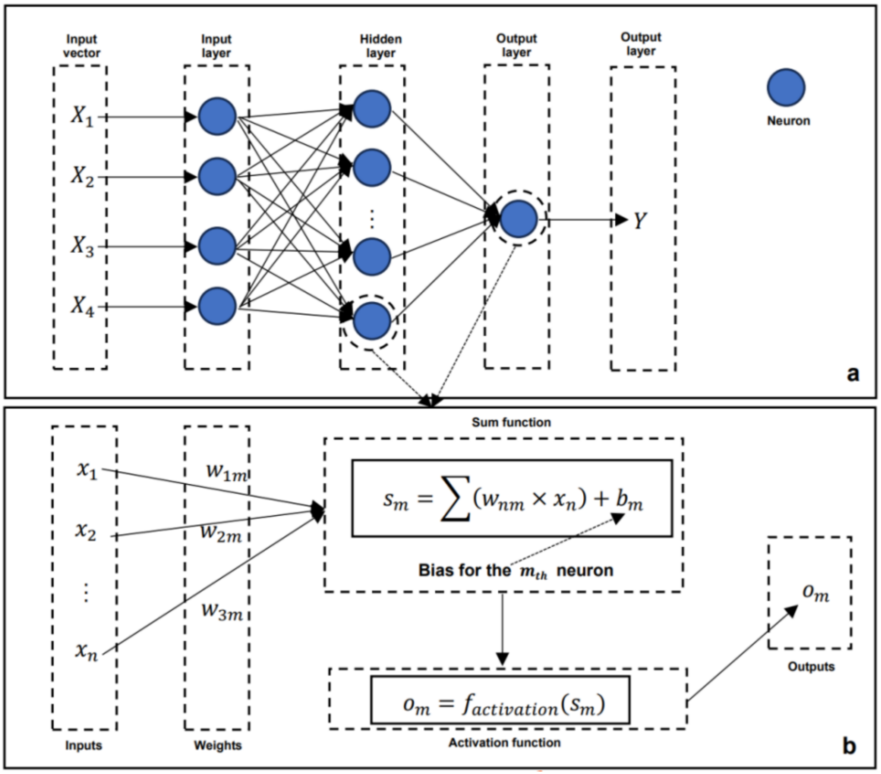

Figure 1. Structure of neural network model and calculation process of the neuron

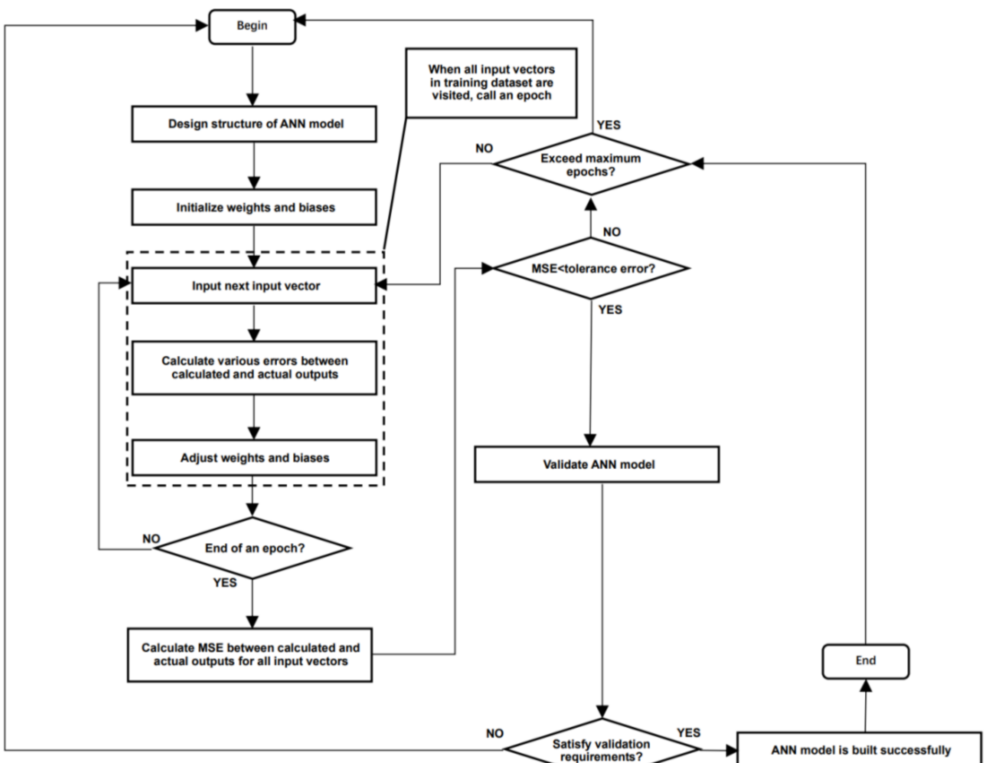

Figure 2. Training process of neural network model

The reprocess dataset from Section 2.2 is used to train, validate, and test the ANN flight delay prediction ANN model. Whether getting an accurate and reliable ANN model lies heavily on the model parameters' settings. Therefore, sufficient appropriate analysis and attempts about these parameters' settings are conducted in this paper. An example configuration of the ANN model is given in Table 2.

Table 2. Configurations of the ANN model

Parameter	Value
Number of the hidden layer	4
Number of the neurons in the input layer	4
Nunber of the neurons in the hidden layer	9
Number of the neurons in the output layer	1
Activation function of the hidden layer	Tan-Sigmoid
Activation function of the output layer	Linear
Goal mean squared error	1.5e-4
Maximum epochs	1000
Learning rate	Adaptative
Training algorithm	Levenberg-Marquardt Backpropagation [5]

2) Support vector machine

The classification principle of SVM is given in Equation (1).

$$W^T X + b \geq +1, y \Rightarrow +1 W^T X + b \leq -1, y \Rightarrow -1 \qquad (1)$$

where W, b, X and y are the weight matrix, bias, input vector and classification labe. The training process of SVM is shown in Equation (2). The training, validation and test dataset for this SVM are the same as those of the above ANN model.

$$wb^{i+1} = wb^i - \gamma \nabla_w J(X^i, wb^i, y^i) \qquad (2)$$

where wb is a matrix including all weights and bias, γ is the learning rate, J is the cost function, and Xi and yi are respectively the input vector and actual label of ith training data.

3 Experimental Results

Once the two models are trained successfully, the test dataset is used to validate the performance of the model. Part of the test data and predicted delay time by these two models are shown in Table 3 and Figure 3. The mean relative error based on Equation (3) between the predicted and actual delay of the prediction model are 18.6% and 21.2% by ANN and SVM, respectively. Besides, the time cost of predicting a single roughness value based on cutting parameters are about 100.4 ms and 58.5 ms with ANN and SVM under such a software environment with an i5 7500 CPU and 16GB 2400HZ RAM.

$$E_{mr} = \frac{\sum_{i=1}^{N_{test}} \left| \frac{t_{predict\_i} - t_{actual\_i}}{t_{actual\_i}} \right|}{N_{test}} \qquad (3)$$

where E_{mr} indicates the mean relative error, N_{test} is the number of the test data (or the size of the test dataset, N_{test} = 9 in this paper), $t_{predict\_i}$ and T_{actual_i} are respectively the predicted and actual delay for the i_{th} test data.

Table 3. Test data and predicted flight delay by ANN and SVM

Month	The day of week	Scheduled Departure time	Departure location		Arrival location		Scheduled Arrival time	Delay	Predicted delay by ANN	Predicted delay by SVM
1	2	0630	40.65236	-75.44040	42.55708	-92.40034	0828	10	12.6	12.8
3	4	1230	32.41132	-99.68190	35.21937	-101.70593	1554	-5	-3.1	-4.2

5	7	1750	35.04022	-106.60919	61.17432	-149.99619	2354	12	14.2	13.4
7	6	1900	45.44906	-98.42183	45.07807	-83.56029	2121	6	4.8	3.9
8	2	2130	31.53552	-84.19447	39.22316	-106.86885	0040	8	9.1	10.2
9	3	2300	41.25305	-70.06018	33.64044	-84.42694	0124	9	7.2	6.9
10	6	0300	31.61129	-97.23052	44.25741	-88.51948	0552	6	6.5	6.7
11	5	1430	40.97812	-124.10862	30.19453	-97.66987	1820	20	24.1	23.5
12	1	0430	40.65236	-75.44040	35.43619	-82.54181	0632	7	6.8	5.6

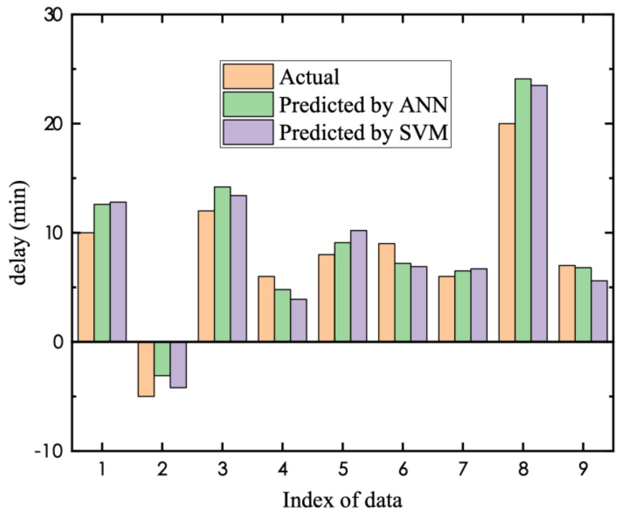

Figure 3. The comparison of actual and predicted flight delay

4 Conclusion

This paper proposes two models, ANN and SVM, for flight time delay prediction. Based on the experimental results, these two models can predict the delay accurately and efficiently and the ANN model's accuracy is a little better than that of the SVM one. Specifically, the mean relative errors of ANN and SVM are 18.6% and 21.2% within a certain test dataset, respectively. While the SVM model with average 58.5 ms time cost is more efficient than the ANN model with average 100.4 ms time cost.

About the Author

Yuxin Huang is undergraduate of Tan Kah Kee College of Xiamen University, and his research field is Artificial Intelligence.

References

[1] Moreira, L., Dantas, C., Oliveira, L., et al. (2018) On evaluating data preprocessing methods for machine learning models for flight delays. In: IJCNN. Rio de Janeiro. pp. 1-8.

[2] Zhang, S., Li X., Zong M., et al. (2017) Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, 29(5): 1774-85.

[3] Sun, J., Wu, Z., Yin, Z., et al. (2018) SVM-CNN-Based fusion algorithm for vehicle navigation considering atypical observations. IEEE signal processing letters, 26(2): 212-16.

[4] Pamplona, D.A., Li, W., Barros, A.G., et al. (2018) Supervised neural network with multilevel input layers for predicting of air traffic delays. In: IJCNN. Rio de Janeiro. pp. 1-6.

[5] Indolia, S., Goswami, A.K., Mishra, S.P., et al. (2018) Conceptual understanding of convolutional neural network-a deep learning approach. Procedia computer science, 132: 679-88.