ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 269-275
doi: 10.18178/wcse.2018.06.048

Predicting Peak Service Rate Based On Weather Impacts Using Machine Learning Techniques

Si Chen[+], Minghua Hu and Zheng Zhao

Collage of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract. As the air traffic congestion and large-scale flight delays become more and more serious, it is particularly important to predict the highest sustainable throughput grades of terminal which improved the effect of TFM. Current research has focused on predicting the impact of runway configurations on airport capacity. However, the selection of runway configuration does not take into account all the weather conditions that affect the terminal zone operation, and the transition from runway configuration to airport capacity is also a complex study. This article describes a methodology for predicting peak service rate based on the meteorological conditions directly. Two machine learning algorithms were introduced and evaluated for use in the process of developing the model. K-means algorithm is one of the unsupervised learning algorithms in machine learning filed which can be used for clustering. Random forest is a non-traditional machine learning algorithm, composed of many decision trees, which can be used for classification and regression. First, a k-means algorithm is applied to all days in 2016 of Guangzhou airport, resulting in the identification of 3 clusters that represent unique classifications of peak service rate that were historically implemented. Second, a forecast model based on the weather is developed by the application of random forest algorithm. The model provided 7 input features that describe the weather and evaluated the importance of each. Finally, after calculating the 5301 data from Guangzhou airport, the predict accuracy of model indicates that this methodology is feasible but still needs some improvements.

Keywords: peak service rate, weather, machine learning, K-means, random forest, prediction.

1. Introduction

Nowadays, the phenomenon of air traffic congestion and large-scale flight delays become more and more serious which have not only caused an increasing number of irregular flights, but also resulted in a decrease in the quality of civil aviation safety and service. Ultimately, this is bound to produce overloaded airspace where more number of aircraft go through the airspace than its capacity due to the limited resources of airspace system all over the world (Shin et al, 2013). In this study, we mainly focus on the terminal area which plays an important role during the whole aircraft operation. Terminal airspace refers to the area that is 50 nautical miles horizontally and18000 feet vertically from the major airports which is regarded as one of the most complex airspace (Chen et al, 2012). Therefore, follow-up operation will be disrupted by this situation which has been the reason for making terminal area as the main bottleneck detrimental to the response to the increase in throughput of the airport (Zuniga et al., 2011).

After a certain time period, a post-operations analysis will reveal the highest sustainable throughput values that the airport has been able to accommodate. This can be determined by looking at the traffic levels during typical busy-hour periods. This typical busy-hour indicator is called the Peak service rate. More accurate and real-time prediction of peak service rate has made a more reasonable and scientific basis for the implementation of traffic management decision-making which improved the effect of TFM. There are many factors that affect air traffic operation, of which weather is the key factor exerting direct and indirect influence on it. Not only wind speed, precipitation and visibility are included, but also ceiling, wind direction

[+] Corresponding author.
E-mail address: 603550521@qq.com

and other weather features have localized and temporary effects on operations of the airport (Kicinger et al., 2011). Low ceiling, big wind speed, thunderstorm and other hazardous weather conditions have negative impacts on terminal airspace operations which are all proved in previous work.

Up to now, there have been existing some methods and techniques to support the implementation of traffic management strategies, of which the applications of machine leaning perform especially well. Liu (2008) utilized K-means clustering to cluster the data into several clusters with the similarity between each observation based on the historical airport capacity profiles. Wang (2011) used the Support Vector Machine (SVM) and Ensemble Bagging Decision Tree (BDT) to solve the prediction of the airport arrival rate (AAR) with weather records and he found that BDT out-performed SVM. Dhal et al. (2013) introduced a multinomial Logistic Regression model on the problem of forecasting Airport Arrival Rates (AAR) over a full-day time horizon. The author applied just significant capacity drop instead of specific runway configurations of specific airport so that this methodology can be applied to any airports. Zhang et al. (2013) used hourly AAR records of three years as the training data to construct the regression model and then used data of the year next to that as the testing data to validate the regression model respectively. Despite the fact that there have been already some studies that use weather data to predict the operation of terminal areas, it is still difficult to estimate and forecast it accurately and real-time. From Dhal et al. (2013)'s research, there still are some challenges should be addressed as well, including the selection of appropriate meteorological attributes. Kenneth D. Kuhn (2016) also say that there are still several challenges here including the need to identify the appropriate machine learning algorithms to apply, to obtain appropriate raw data, and to evaluate results in a meaningful way.

In this study, we propose a prediction method of peak service rate based on weather impacts. We are committed to mining the relationship between multiple weather variable combinations and peak service rate as well as enhancing the reliability and accuracy of forecast. The paper is organized as follows. Section II introduces the way of processing the data to meet the requirement of model inputs. In Section III, we propose clustering and classification algorithms which are prepared to use. The results in Section IV describe how the algorithms perform with respect to the problem objective, as well as the results of some feature scoring and parameter settings work. Finally, some possibilities for future work and conclusions are described in Section V.

2. Data Pre-Processing

2.1. Meteorological Data

The most widely used weather data formats is Meteorological Terminal Aviation Routine Weather Report (METAR) which summarize observed airport weather conditions. METAR reports are issued hourly in order to ensure information keep up with changing weather conditions and because of that, the amount of data which we need is sufficient. The main source of METAR data is comprehensive historical archive online. In this paper, we developed a Matlab script to download the historical METAR information from www.ogimet.com to obtain 2016 annual weather data of Guangzhou Airport.

The collected weather data which we acquired from the website contains a large amount of data that we do not need as well as outliers such as missing values, noise value, and so on. And we also need to make sure the data set of model's input is in a numeric format. Therefore, data pre-processing is necessary. In this paper, we process the data mainly in two steps. First, we survey the aviation system literatures and subject matter experts to yield weather features that are accepted as important for describing weather conditions. Second, the data should be standardized. Since the change of the wind direction of the small wind speed does not affect the operation. In our case, varying wind direction and wind direction missing value with small wind speed are all replaced by 0. For the lack of visibility and cloud height, it is generally because the weather is good and does not affect the operation, so such missing value is replaced by the maximum value in the same category data records.

The weather vector, W, is then a 7-dimensional vector:

$$W = [DIR; SPD; GUS; CLG; VSB; SWT; PCP] \quad (1)$$

where, DIR is wind direction, SPD is wind speed, GUS is gust, VSB is visibility, CLG is cloud height, SWT is special weather types which represents the presence of thunderstorms, snow as well as some other adverse weather conditions, and PCP is precipitation.

2.2. Peak Service Rate

In the report issued by Euro-control, the historical flight data is utilized to determine the peak flow of the departure and arrival per unit time. This is defined as the peak service rate. As a pre-processing step, we have filtered the data to remove historical records between 21PM and 9AM in the local time. For the value of these hours is often low, rather than required by the environmental conditions, as there is no need to meet a higher traffic demand. So we cannot get the peak service rate at these hours. After deleting some of the useless data, we give a more clear definition of peak service rate. The research indicates 95% is the most appropriate envelopment interval, which means that the top 5% of the movement rate can be selected to be the peak service rate which constrained by certain combination of weather variables in our case. In accordance with the theory, we first count the frequency of each combination of meteorological variables and then find out the movement rate on the top 5% as the corresponding peak service rate. The value of peak service rate, R, can be expressed by the following formula.

$$\mathrm{R}_{w_i} = 95\%(F_{w_i}{}^{\max} - F_{w_i}{}^{\min}) + F_{w_i}{}^{\min} \tag{2}$$

Here R_{w_i} represents the value of peak service rate under the impact of w_i which is belong W. $F_{w_i}{}^{\max}$ is the maximum value of flow when the weather vector is w_i as well as $F_{w_i}{}^{\min}$ is the minimum value.

After the peak service rate in the case of each weather vector has been determined, our next task is to determine the peak service rate categories based on numerical thresholds which is the fundamental for the classification model. We choose the thresholds by using the K-means algorithm which will be described below. Once the numerical thresholds have been determined, the numerical peak service rate values, R, should be replaced by symbolic labels, C, indicating the different classifications.

3. Prediction Model

3.1. Peak Service Rate Clustering Model

To determine the peak service rate categories, a simple k-means clustering algorithm was employed. This algorithm usually uses the square error criterion, the cost function is as follows:

$$\mathrm{F}(\mathrm{C}) = \sum_{j=1}^{n}\sum_{i=1}^{m}(x_{ji} - c_{li})^2 \tag{3}$$

In the formula, x_{ji} is the value of the ith dimension of the object j, c_l is the center of the class closest to the object j, c_{li} is the value of the ith dimension of c_l. Briefly, this algorithm attempts to partition a set of n objects (peak service rate in our case) into K clusters. The purpose of it is to make the intra-cluster similarity is high while the inter-cluster similarity is small. Here the cluster similarity is measured with respect to the mean value of the objects in the cluster. In our paper, the Euclidean distance was utilized as the distance metric. One of the essential parameters of the k-means clustering algorithm is the number of clusters, K. Here the average Silhouette score is used to select the best value of K. The Silhouette score lies between -1and 1, among them, the value of 1 indicates that the object is within an appropriate cluster, the value of -1 indicates that the object would belong to a more suitable cluster, while a value of 0 indicates that the object is on the border of being placed in two different clusters. It is obvious that the average Silhouette score is a measure of how well a given object fits within the other one in the same cluster. The Silhouette score for object *i* can be written as below:

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \tag{4}$$

Here a_i is the average distance of object *i* to all other objects in the cluster that *i* belong to. Meanwhile, b_i is the lowest average distance between object *i* and the members of another cluster that *i* has been not assigned to.

3.2. Prediction Model Based on Weather

Numerous published research papers utilized decision tree models for prediction in this field. Based on this, we choose random forest which is improved from bagging as the classifying model of this paper. Specifically, random forest is composed of the combination of decision trees. Random forests have many advantages. First, random forest models outperform decision tree models, in terms of accuracy. This is because they are less likely to over fit observed data since they construct the models based on a sample of the observed data instead of the full set of that. Second, it is not necessary to have separate data sets for training and testing a random forest model. Because we can use part of the data to train a tree, and then use the ‘‘out of bag” data test that tree in an unbiased way. Third, the Mean Decrease Gini (MDG) is utilized when using random forest models so that the feature importance can be easily acquired. When every node in every decision tree split, the Gini impurity value which is defined as follow decreases. The more it decreases, the more useful the split is for prediction purposes.

$$i_c = 1 - \left(\frac{w_i n_{i,c}}{\sum_i w_i n_{i,c}} \right)^2 \tag{5}$$

The Gini impurity value is smaller, while the MDG is bigger which means that the feature used to make the split plays a more important role in these model.

Random forest constructs different training sets to increase the differences among the classification models so as to improve the extrapolation prediction ability of the combined classification model. By k-round training, the classification model sequence $\{h_1(X), h_2(X), ..., h_k(X)\}$ is obtained, and then they are used to form a multi-classification model system. The final classification result of the system is obtained by the simple majority voting method. Therefore, the final decision can be expressed by the following formula.

$$H(w_i) = \arg\max \sum_{i=1}^{m} I(h_i(w_i) = c_j) \tag{6}$$

Here $H(\bullet)$ represents the combined classification model, $h(\bullet)$ is a single decision tree classification model, c_j represents the target peak service rate classification, $I(\bullet)$ is an illustrative function.

4. Results

4.1. Identification of Inputs

We collect the flight data and METAR information for the whole year of 2016. After data pre-processing, the classification of peak service rate is required. The categories can be determined by using k-means clustering algorithm.

To determine a reasonable value of K for our project, the flight data in 2016 was repeatedly clustered into 2 through 10 different clusters to get the average silhouette score for each of the different number of clusters which is plotted in Fig. 1. The maximum average silhouette score was occurred when k=2. However, taking into account the reality, it is not appropriate to divide peak service rate into two categories. Finally, the movement rates are categorized into 3 levels (L, M, H) is given in Table 4.1 when the prediction accuracy of classification model and operation situation as well as the average silhouette score are all considered. Label L means that the level of peak service rate is low which shows the weather conditions are likely to cause serious traffic backlog. Label M means the weather has an impact on the operation, but not much. Label H shows that the weather has almost no effect on operation. The level of peak service rate support the implementation of traffic management strategies.

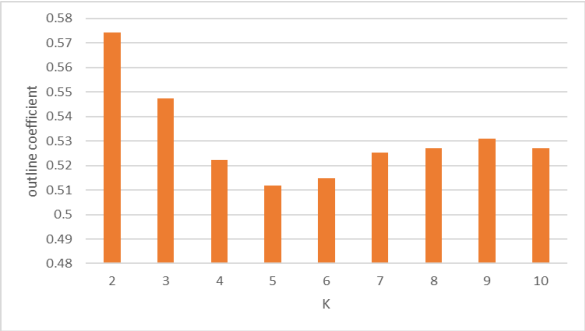

Fig. 1: Average silhouette score as a function of the number of clusters.

Table 4.1: The peak service rate levels at ZGGG

Value	Label
Peak service rate ⩽ 54	L
54 < Peak service rate⩽65	M
Movement rate > 65	H

4.2. Peak Service Rate Prediction

In this section, model is constructed to predict the peak service rate at Guangzhou Airport (ZGGG). There are three important parameters that need to be identified when using random forest algorithm. Mtry is the number of feature variables extracted when each node splits. Numerous experiments have shown that when using random forests for classification, if the total number of features is M, then the mtry's best value is suggested to be $\sqrt{M}$ (two in our case). Nodesize is the number of samples that leaf nodes contain. In general, nodesize takes a value of 1 when implementing random forest to classify. (one in our case) Out of bag (OOB) data is the characteristic of the random forest to estimate the error of the model. The magnitude of the error is associated with the number of decision trees. Too few trees are likely to increase the error, whereas the model may be more complex and less efficient owing to the excessive numbers of decision trees.

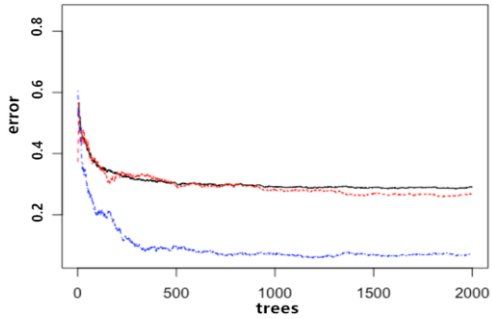

Fig. 2: The error of different numbers of decision trees.

In Figure 2, the error curves start to be stable with the number of decision trees between 500 and 1000, it may be more reasonable to build the random forest model composed of the number of decision trees in that range. (1000 in our case).

Table 4.2: The importance of weather features

Weather Feature	Importance
DIR	210.54
SPD	156.90
GUS	10.77
CLG	271.41
VSB	217.40
SWT	36.84
PCP	25.39

Among the meteorological features that exert influence on the peak service rate, not all of them play important roles in the prediction. Table 4.2 presents the importance of each feature in the model.

As seen in the figure, the cloud height (CLG) is the major contributor to the model. Additionally, the visibility (VSB) and wind speed (SPD) as well as wind direction (DIR) also play an important role. However, the special weather types (SWT) and precipitation (PCP) seems not essential to this model.

In our project, the total samples were divided into two groups using random sampling. Among them, 70% of the samples were used as the training samples to optimize the parameters and construct the model. And 30% of them were used as the test samples to examine the reliability of the model.

Table 4.3: Confusion Matrix of prediction

	H	L	M
H	881	19	391
L	0	33	19
M	106	12	129

By calculating 5301 data, the predict accuracy of model is 65.6%. As in the table 4.3, the number on the diagonal represents the classification to the right levels. For instance, 881 forecasts are classified correctly to the High level, whereas 19 and 391of the total are assigned to other levels by mistake. Similarly, only 129 forecasts of M are placed in the right level. Moreover, almost 40% of the results in Low level are misclassified into the Medium level.

5. Conclusion

A random forest classification model has been developed for prediction of peak service rate based on environmental parameters. Our main work here was on constructing and testing the model using historical data. Case studies of Guangzhou Airport, indicate that the utilized methodology is able to find the relationship between peak service rate categories (e.g., L, M, H) and the weather features. However, there are still several challenges must be addressed. The accuracy of prediction needs some improvements, choosing appropriate forecast tools for prediction, selecting suitable features are still challenges. And we also need to think about how to obtain numerical rather than categorical predictions of peak service rate.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China (U1333202).

7. References

[1] Shin, S., Nandiganahalli, J. S., & Hwang, I. (2013). Diagnostic Tool for Throughput FactorAnalysis in En-route Airspace. In 2013 Aviation Technology, Integration, and Operations Conference (p. 4339).

[2] Chen, J. T., Yousefi, A., Krishna, S., Sliney, B., & Smith, P. (2012, October). Weather avoidance optimal routing for extended terminal airspace in support of dynamic airspace configuration. In Digital Avionics Systems Conference (DASC), 2012 IEEE/AIAA 31st (pp.3A1-1). IEEE.

[3] Zuniga, C., Delahaye, D., & Piera, M. A. (2011, October). Integrating and sequencing flows in terminal maneuvering area by evolutionary algorithms. In Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th (pp. 2A1-1). IEEE.

[4] Kicinger, R., Cross, C., Myers, T., Krozel, J., Mauro, C., & Kierstead, D. (2011, August). Probabilistic airport capacity prediction incorporating the impact of terminal weather. In AIAA Guidance, Navigation and Control Conference (pp. 8-11).

[5] Liu, P. C. B., Hansen, M., & Mukherjee, A. (2008). Scenario-based air traffic flow management: From theory to practice. Transportation Research Part B: Methodological, 42(7), 685-702.

[6] Wang, Y. (2011, October). Prediction of weather impacted airport capacity using ensemble learning. In Digital Avionics Systems Conference (DASC), 2011 IEEE/AIAA 30th (pp. 2D6-1). IEEE.

[7] Dhal, R., Roy, S., Taylor, C., & Wanke, C. (2013, August). Forecasting weather-impacted airport capacities for flow contingency management: Advanced methods and integration. In AIAA Aviation Technology, Integration, and Operations Conference, Los Angeles, CA.

[8] Mukherjee, A., Grabbe, S. R., & Sridhar, B. (2013). Classification of Days Using Weather Impacted Traffic in the National Airspace System. Aviation Technology, Integration, and Operations Conference (Vol.390, pp.285-297).

[9] Kuhn K D. A methodology for identifying similar days in air traffic flow management initiative planning[J]. Transportation Research Part C Emerging Technologies, 2016, 69:1-15.