APEM journal
Advances in Production Engineering & Management **ISSN 1854-6250**
Volume 16 | Number 3 | September 2021 | pp 285–296 Journal home: apem-journal.org
https://doi.org/10.14743/apem2021.3.400 Original scientific paper

Using the gradient boosting decision tree (GBDT) algorithm for a train delay prediction model considering the delay propagation feature

Zhang, Y.D.[a,*], Liao, L.[a], Yu, Q.[a], Ma, W.G.[a], Li, K.H.[b]

[a]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, P.R. China
[b]School of management, Xihua University, Chengdu, P.R. China

A B S T R A C T

Accurate prediction of train delay is an important basis for the intelligent adjustment of train operation plans. This paper proposes a train delay prediction model that considers the delay propagation feature. The model consists of two parts. The first part is the extraction of delay propagation feature. The best delay classification scheme is determined through the clustering method of delay types for historical data based on the density-based spatial clustering of applications with noise algorithm (DBSCAN), and combining the best delay classification scheme and the k-nearest neighbor (KNN) algorithm to design the classification method of delay type for online data. The delay propagation factor is used to quantify the delay propagation relationship, and on this basis, the horizontal and vertical delay propagation feature are constructed. The second part is the delay prediction, which takes the train operation status feature and delay propagation feature as input feature, and use the gradient boosting decision tree (GBDT) algorithm to complete the prediction. The model was tested and simulated using the actual train operation data, and compared with random forest (RF), support vector regression (SVR) and multilayer perceptron (MLP). The results show that considering the delay propagation feature in the train delay prediction model can further improve the accuracy of train delay prediction. The delay prediction model proposed in this paper can provide a theoretical basis for the intelligentization of railway dispatching, enabling dispatchers to control delays more reasonably, and improve the quality of railway transportation services.

A R T I C L E I N F O

Keywords:
Train delay prediction;
Actual train operation data;
Delay type identification;
Delay propagation feature extraction;
Density-based spatial clustering of applications with noise (DBSCAN);
k-nearest neighbor (KNN);
Gradient boosting decision tree (GBDT);
Random forest (RF);
Support vector regression (SVR);
Multilayer perceptron (MLP)

**Corresponding author:*
ydzhang@swjtu.edu.cn
(Zhang, Y.D.)

Article history:
Received 24 July 2021
Revised 25 October 2021
Accepted 28 October 2021

1. Introduction

With the development of railway network and the growth of passenger travel demand, the utilization rate of railway lines is getting higher and higher. Under the premise of ensuring the safety of train operation, ensuring the punctuality is the key of railway transportation to improve the quality of service. Once the train is delayed, the dispatchers must use dispatch adjustment methods reasonably and scientifically [1]. The accurate prediction of train delays can assist dispatchers to make scientific decisions, and even realize the intelligent dynamic adjustment of train operation plans.

The traditional train delay prediction mainly relies on the work experience and operating skills of dispatchers. Due to the uncertainty of train delays, this method is difficult to reasonably predict the delay before it occurs. On the other hand, the primary delay caused by the interference of external factors will produce a domino effect-like delay propagation effect on the line, which will lead to the secondary delay. However, it is very limited to rely on the experience of dispatchers to predict the secondary delay. With the development of railway informatization, on the basis that the actual operation data of trains can be fully collected and fully processed, the application of big data and machine learning to train delay prediction has important reference value for the development of intelligent dispatching and command work [2]. The machine learning method is based on the actual train operation data and do not require relevant details within the system. The actual data can reflect the relevant factors and their interactions of delays. This method is conducive to revealing the occurrence and propagation of train delays.

This paper proposes a train delay prediction model that considers the delay propagation feature. When the model uses machine learning methods for delay prediction, the delay propagation feature is added to improve the prediction accuracy. The main contributions of this paper include: (1) Design the clustering method of delay types for historical data based on density-based spatial clustering of applications with noise (DBSCAN) and the classification method of delay types for online data based on k-nearest neighbor (KNN); (2) According to the determined delay type, the delay propagation factor is used to quantify the delay propagation relationship and construct as the horizontal and vertical delay propagation feature; (3) Construct a gradient boosting decision tree (GBDT) model to complete the prediction of train delays according to the train operation status feature and the delay propagation feature.

This paper is organized as follows. Section 2 summarizes the current research on train delay prediction. Section 3 describes the problems to be solved. Section 4 describes the overall structure of the train delay prediction model proposed in this paper and describes the design principles of each part in detail. In Section 5, an example is analyzed based on the actual data of train operation. Section 6 summarizes the work of this paper.

2. Related work

The input feature set of the machine learning model will affect the performance of the model. So the selected input feature should have the greatest impact on the output results.

It is a routine consideration to use information related to train characteristics and train status as the input feature set of the prediction model, because these are factors that directly affect train delays. Oneto *et al.* [3] took the section running time, working day/non-working day feature as the input feature set. Wang and Zhang [4] identified the number of delayed trains at each station, the total value of delays of each station and the total value of each train's delays as factors affecting train delays. Shi *et al.* [5] considered the current station and train codes when establishing the feature set. The related historical value of train operation is also related to the delay prediction. Nair *et al.* [6] considered the historical delay value and historical running time of the train at the station as the input feature of model. But the train runs according to the predesigned train operation diagram, therefore, the planned values related to the operation diagram such as the planned running time [7], the planned stopping time of trains and the planned running interval between trains [8, 9] should also be taken into consideration.

To improve the quality of prediction, some studies have begun to consider other factors besides the feature of train operation status. Tang *et al.* [10] used the primary delay time, the number of affected trains and the delay causes as independent variables. Zhang *et al.* [11] based on the secondary delay data, considered the impact of the preceding train on the current train and constructed a model input feature set. Hu *et al.* [12] used a hierarchical clustering algorithm to analyze the delayed trains and based on the results of 4 types of delayed train sequences made subsequent delay time predictions. Zeng *et al.* [13] used cause analysis to infer the delay propagation chain, integrated the train event information with the primary delay and secondary delay information contained in prediction model.

The above literature review shows that when establishing the input feature of prediction model, most studies consider the train operating status information, and also the factors related to the data characteristics and the predicted output. But few studies consider the delay propagation relationship in the delay prediction. In terms of prediction models, using machine learning models to predict delays has a better fitting effect than traditional statistical models [14, 15]. The machine learning model can realize the prediction of delay more accurately and quickly, and at the same time, the output can be stabilized in the case of a large amount of data. Decision tree model [16], random forest model [6, 7, 9, 13], neural network model [8, 17] and SVR model [10] are now widely used in train delay prediction research.

Therefore, this article will establish a GBDT-PF model considers the delay propagation between trains and the delay propagation of the train itself. The model can identify the delay types and obtain the delay propagation relationship. The delay classification scheme can enable dispatchers to understand the law of the occurrence of primary delays, and the delay propagation relationship has a direct impact on the subsequent train delays, so the prediction model that considers the delay propagation relationship can obtain higher accuracy.

3. Problem statement

The train operation process in the railway network can be expressed as a collection of a series of events and processes [18]. The dependency between events and processes can be represented by timed event graphs. Fig. 1 is the use of time event graph to show the operation status of each train. According to the time of day, each train is arranged vertically $(1,2,3,\cdots,i)$ and each station is arranged horizontally $(1,2,3,\cdots,s)$.

In Fig. 1, the node $t_{i,s}$ represents the arrival or departure event of train i at station s, the weight of the node $D_{i,s}$ represents the delay value of train i at station s, the directed arc $arc\big(t_{i,s},t_{i_{+1},s}\big)$ represents the operation process that train i and train $i+1$ run to station s. the weight of the directed arc $w\big(t_{i,s},t_{i+1,s}\big)$ represents the running interval between train i and train $i+1$ at station s. When $t_{i,s}$ represents the departure event, the directed arc $arc\big(t_{i,s},t_{i,s_{+1}}\big)$ represents the running process of train i from station s to station $s+1$, the weight of the directed arc $w\big(t_{i,s},t_{i,s_{+1}}\big)$ represents the running time of the corresponding train i from station s to the station $s+1$. When $t_{i,s}$ represents the arrival event, the directed arc $arc\big(t_{i,s},t_{i,s_{+1}}\big)$ represents the stopping process of train i at station s, and the weight of the directed arc $w\big(t_{i,s},t_{i,s_{+1}}\big)$ represents the stop time of train i at station s.

According to Fig. 1, when $D_{i,s}$ is primary delay, the horizontal propagation of the delay occurs through the directed arc $arc\big(t_{i,s},t_{i+1,s}\big)$, and it affects the next train. Vertical propagation of the delay occurs through the directed arc $arc\big(t_{i,s},t_{i,s+1}\big)$, and it affects the train itself.

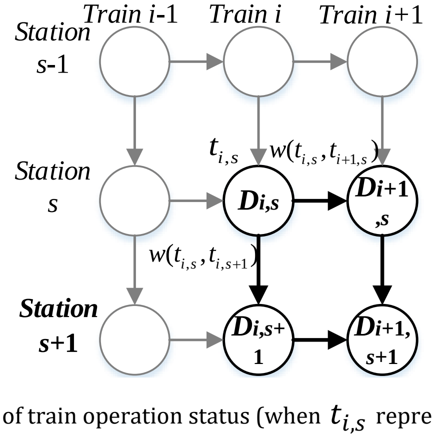


Fig. 1 Time event graph of train operation status (when $t_{i,s}$ represents a departure event)

The delay value of the train should be related to the historical value and has nothing to do with whether the train will be delayed in the future. For the node $t_{i,s}$, there are only two nodes directly related to it, the node $t_{i\text{-}1,s}$ in the horizontal direction and the node $t_{i,s\text{-}1}$ in the vertical direction. These two nodes also represent the delay propagation between trains and the delay propagation of the train itself. Therefore, given the train i and the station s, considering the train operating state factors Z_i, the delay propagation factor of the train itself $P_{i,s-1}$ and the delay propagation factor between the trains $P_{i-1,s}$, a nonlinear function is learned to complete the prediction of $\widehat{D}_{i,s}$:

$$\widehat{D}_{i,s} = f\left(Z_i, P_{i,s-1}, P_{i-1,s}\right) \tag{1}$$

Among them, $\widehat{D}_{i,s}$ is the delay time of the arrival or departure of train i at station s. Z_i is a feature set related to $D_{i,s}$ based on historical data, $P_{i,s-1}$is a feature set related to the vertical propagation of delays, $P_{i-1,s}$ is a feature set related to the horizontal propagation of delays, and f is the establishment machine learning model.

4. Methodology

Fig. 2 shows the structure of the GBDT-PF model. The feature extraction part is to complete the construction of the input feature set. Then the GBDT model is trained to realize the prediction of $\widehat{D}_{i,s}$. For the value $\widehat{D}_{i,s}$, it is necessary to confirm the type of the two nodes $t_{i,s-1}$ and $t_{i-1,s}$, but since the original data set has no related records for the type of delay, before constructing the feature set, the first task is to classify the delay types in the historical data, then construct a data set that can be used for delay type identification, so as to carry out the subsequent delay prediction.

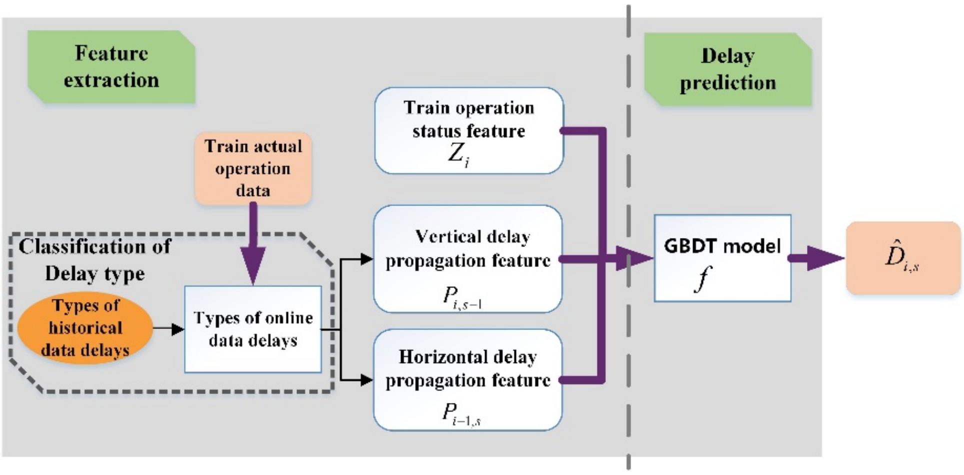

Feature extraction
Delay prediction
Train operation status feature Z_i
Train actual operation data
Classification of Delay type
Types of historical data delays
Types of online data delays
Vertical delay propagation feature $P_{i,s-1}$
Horizontal delay propagation feature $P_{i-1,s}$
GBDT model f
$\hat{D}_{i,s}$

Fig. 2 The GBDT-PF model

4.1 Cluster analysis of delay types for historical data based on DBSCAN

Using DBSCAN algorithm to complete the cluster analysis of delay types for historical data is shown in Fig. 3. For $D_{i,s}$, according to the characteristics of the primary delay and the secondary delay, the input feature set of the cluster analysis is determined as shown in Table 1.

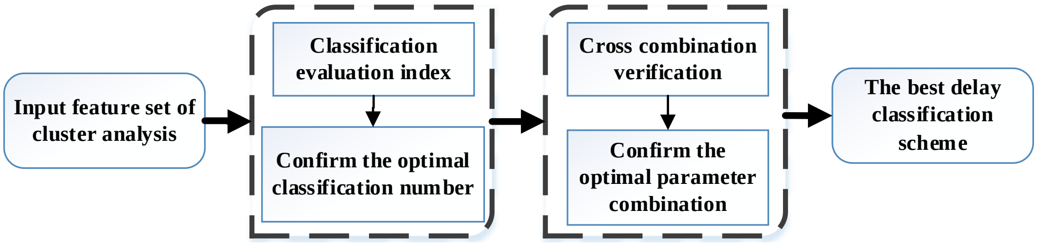

Input feature set of cluster analysis
Classification evaluation index
Confirm the optimal classification number
Cross combination verification
Confirm the optimal parameter combination
The best delay classification scheme

Fig. 3 The process of cluster analysis of delay types for historical data based on DBSCAN

Table 1 The input feature set of the cluster analysis

Feature	Symbol	Meaning
Input feature set of cluster analysis	$Tdelay_{i,s}$	The delay value of train i at station s, a positive value means delay, a negative value means early, on time means the value is 0
	$LTdelay_{i,s-1}$	The delay value of train i at station $s-1$, a positive value means delay, a negative value means early, on time means the value is 0
	$TTdelay_{i-1,s}$	The delay value of train $i-1$ at station s, a positive value means delay, a negative value means early, on time means the value is 0
	$LTdflag_{i,s-1}$	The delay sign of train i at station $s-1$
	$TTdflag_{i-1,s}$	The delay sign of train $i-1$ at station s
	$LTdiff_{s,s-1}$	The deviation of running time of train i from station $s-1$ to station s
	$TTdiff_{i,i-1}$	The deviation of running interval between train i and train $i-1$ at station s

The best delay classification scheme is an important basis for the classification of online data delay categories and has a direct impact on the subsequent prediction accuracy. Therefore, different classification schemes need to be compared to select the best scheme. For the DBSCAN algorithm, it is necessary to confirm the optimal classification number and the optimal combination of parameters (radius and minimum sample size). Table 2 shows the internal indicators that can complete the evaluation of the classification number.

Table 2 Cluster evaluation indicators

Index	Variable symbol	Correlation
Silhouette Coefficient	M_{1c}	Positive correlation
Calinski Harabasz Score	M_{2c}	Positive correlation
Davies-Bouldin Index	M_{3c}	Negative correlation

In order to compare the effectiveness of different delay classification schemes, each evaluation index is standardized. $M_{nc}^{'}$ represents the standardized result of the nth index. The range of the standardized indicators $M_{nc}^{'}$ is 0-1. For further comparison, the final weighted evaluation index is calculated by Eq. 2, the final weighted evaluation index is a negative correlation index.

$$M_c = \sum_{n=1}^{3} \alpha_n \cdot M_{nc}' \tag{2}$$

Herein: M_c is the final weighted index, α_n is the weight of the nth index, and $\sum_{n=1}^{3} \alpha_n = 1$, $0 \le \alpha_n \le 1$.

The best parameter combination will be determined by cross validation. Establish the clustering model under the corresponding parameter combination, and output the number of classifications, the number of abnormal points, and the number of sample points in each category. There are two principles for determining the optimal parameter combination. First, select a parameter combination with a small number of abnormal points. Second, choose a parameter combination with a relatively reasonable distribution of the number of sample points in each category.

4.2 Classification of delay types for online data based on KNN

After obtaining the best delay classification scheme, construct a data set containing delay type labels and complete the training of the delay classification model. The delay classification model will be completed using KNN algorithm. In KNN classification model, the K value has impact on the accuracy of prediction. Therefore, this paper uses the 10-fold cross validation method to determine the best K value. Fig. 4 shows the process of the classification method of delay type for online data.

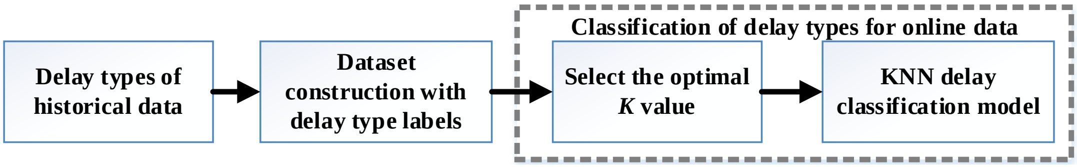

Classification of delay types for online data
Delay types of historical data
Dataset construction with delay type labels
Select the optimal K value
KNN delay classification model

Fig. 4 The process of the classification of delay types for online data based on KNN

4.3 Feature extraction

The train operation status feature Z_i are variables related to the node $t_{i,s}$. First, the station information and arrival/departure information of the node should be considered. Secondly, variables related to $t_{i,s-1}$ and $t_{i-1,s}$ should be considered, including the delay value, the running interval between trains, and the train running time between stations, etc.

In order to better characterize the delay propagation relationship, this paper defines the delay propagation factor of each node. The calculation method is as shown in Eq. 3, where $factor_{i,s}$ is delay propagation factor of the node $t_{i,s}$, $D_{i,s}^n$ is the delay value of nth node in the delay propagation chain. In particular, for nodes that are early or punctual, the delay propagation factor is 0, and for nodes that are primary delay, the delay propagation factor is 1.

$$factor_{i,s} = \frac{D_{i,s}^n}{\sum_{n=1}^{N} D_{i,s}^n} \tag{3}$$

In the time event graph, each node has horizontal delay propagation and vertical delay propagation, the delay propagation factor is also different in the two directions, so there should be delay propagation factor in both horizontal and vertical directions corresponding to each node. The vertical delay propagation feature $P_{i,s-1}$ is related to the delay propagation of the train itself, so it should be related to the vertical delay propagation factor of $t_{i,s-1}$. The horizontal delay propagation feature $P_{i-1,s}$ is related to the delay propagation between trains, so it should be related to the horizontal delay propagation factor of $t_{i-1,s}$. Finally, the three type of input features of the GBDT model are shown in Table 3.

Table 3 Input feature set of GBDT model

Feature	Symbol	Meaning
Z_i	Sta_s	Station number of the sth station
	$A\_flag_{i,s}$	Arrival/departure sign of train i at station s
	$LTdelay_{i,s-1}$	The delay value of train i at station $s-1$, a positive value means delay, a negative value means early, on time means the value is 0
	$LTplan_{s,s-1}$	The planned running time of the train i from station $s-1$ to station s
	$TTdelay_{i-1,s}$	The delay value of the train $i-1$ at station s, a positive value means delay, a negative value means early, on time means the value is 0
	$TTplan_{i,i-1}$	The planned running interval between the train i and train $i-1$ at station s
$P_{i,s-1}$	$factor_{i,s-1}$	Delay propagation factor of $t_{i,s-1}$ (vertical)
$P_{i-1,s}$	$factor_{i-1,s}$	Delay spread factor of $t_{i-1,s}$ (horizontal)

5. Experiment and simulation

5.1 Data description

This paper uses the actual operation data of the 1H passenger express of the West Coast Main Line (WCML) railway in the United Kingdom. There are 37 stations on the route, and the time span is from June 1, 2017 to June 30, 2017, with a total of 47,693 records. The original data records information such as the train number, station number, actual or planned operating time, etc. Data from 75 % of the original data for model training and remaining 25 % of the original data for model testing.

5.2 Determine the type of delay

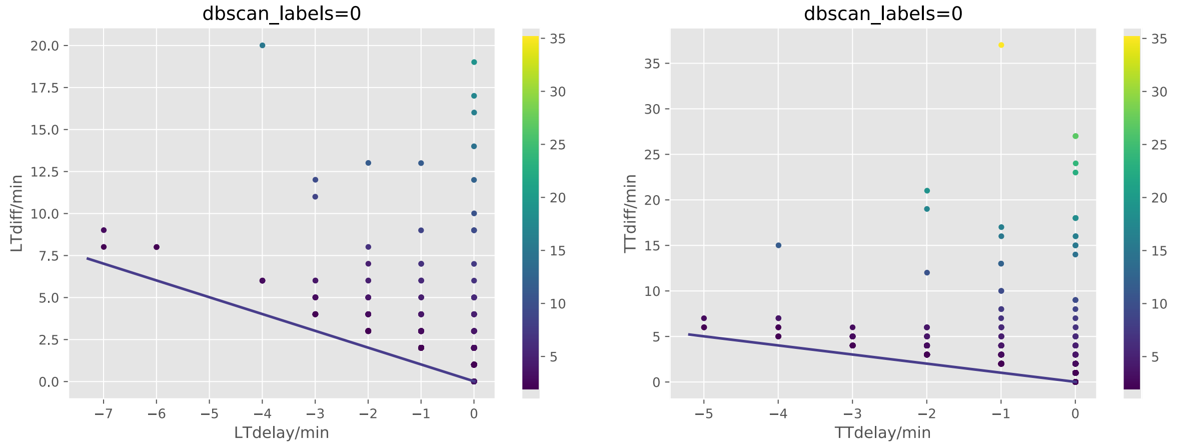
Use the weighted evaluation index M_c to determine the optimal classification number within the range of the classification number 2-10. The evaluation indicators under different classification numbers are shown in Table 4. When classification number is 9, M_c has a minimum value, so the best classification number is 9. In the range where the radius value is 0.05-3 and the minimum sample size value is 2-20, the DBSCAN model is constructed by cross validation, and get 21 groups of parameter combination results with the classification number of 9. According to the principle of selecting parameter combinations in section 4.1, the radius is 2.15, and the minimum sample size is 4.

According to the classification scheme, the delay type of each category is determined in a visual form. In the visual analysis, the analysis is carried out from the vertical and horizontal directions of the train delay propagation. Fig. 5 to Fig. 8 are the visual analysis diagrams of category 0-3, $Tdelay_{i,s}$ is represented by the color of the point in the scatter diagram.

The visualization of category 0 is shown in Fig. 5. In the vertical direction, the value of $TTdiff_{i,i-1}$ is greater than or equal to the value of $LTdelay_{i,s-1}$. In the horizontal direction, the value of $TTdiff_{i,i-1}$ is greater than or equal to the value of $TTdelay_{i-1,s}$. So the train was not affected by the delay propagation in both directions. The train was delayed during operation or was delayed at this station, so it was primary delay.

Table 4 The weighted evaluation index values under different classification numbers

Classification number c	M_c	Classification number c	M_c	Classification number c	M_c
2	0.682	5	0.257	8	0.195
3	0.839	6	0.274	9	0.015
4	0.315	7	0.209	10	0.046

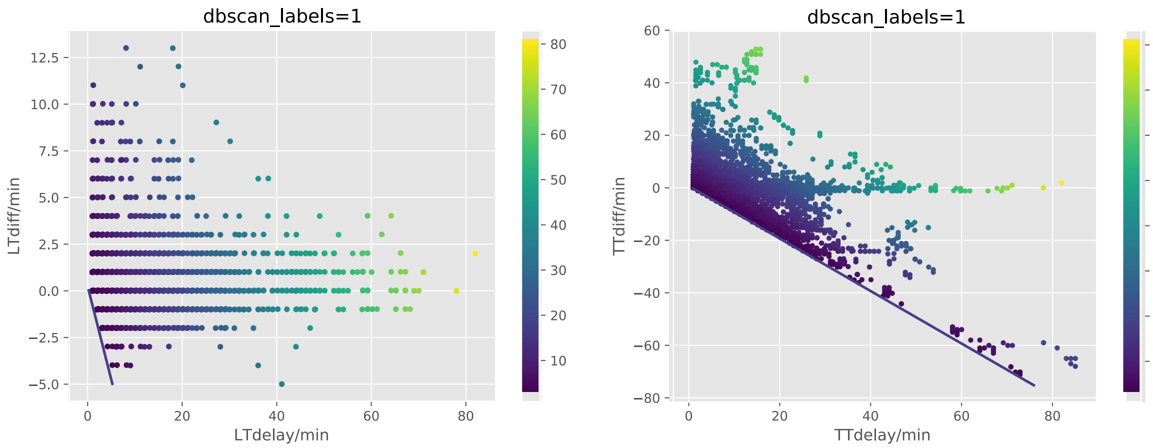


a) Scatter diagram of $LTdelay_{i,s-1}$ and $LTdiff_{s,s-1}$ b) Scatter diagram of $TTdelay_{i-1,s}$ and $TTdiff_{i,i-1}$

Fig. 5 Visualization of Category 0

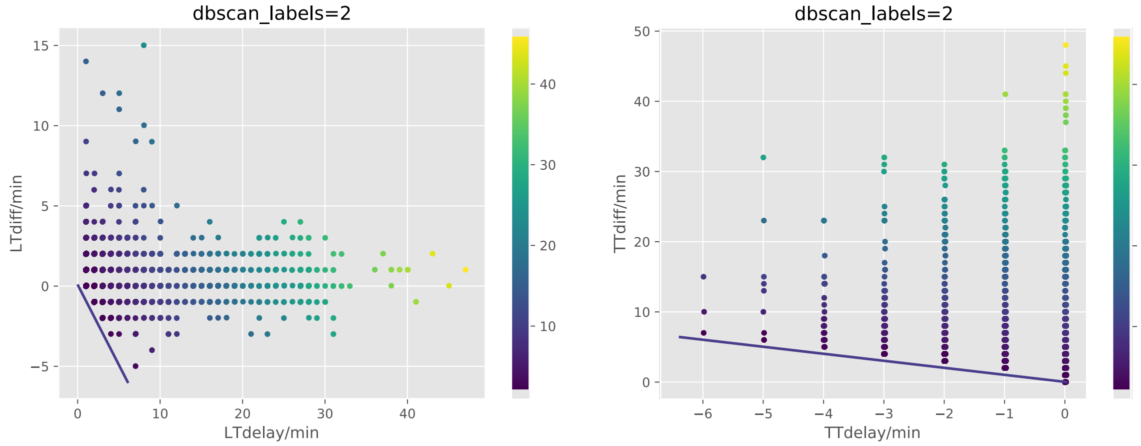
The visualization of category 1 is shown in Fig. 6. The delay of this train was affected by the delay propagation in both the horizontal and vertical directions, so it was secondary delay. The visual analysis results of categories 4, 5, 7, and 8 are the same as category 1. The visualization of the category 2 is shown in Fig. 7. The delay propagation occurred in the vertical direction, and the delay of the train at the previous station had an impact on this station, but there was no delay propagation in the horizontal direction, so it was secondary delay. The visual analysis results of the category 6 are the same as category 2. The visualization of category 3 is shown in Fig. 8. Contrary to category 2, category 3 is related to delays in horizontal directions, so it was secondary delay.

Finally, according to the characteristics of various types of delays in different directions, all data can be finally integrated into 4 types. The characteristics of the 4 types of delays are shown in Table 5, where delay type 1 is the primary delay, and the other three types are secondary delay. According to Table 5, the calculation methods of the delay propagation factors of the four types are different. Delay type 1 starts to propagate from the current node in the horizontal and vertical directions, and should be regarded as the primary delay in both directions. Delay type 2 are affected by the front nodes in both directions and should be calculated as secondary delay in both directions. Delay type 3 and delay type 4 are only affected by the front nodes in one direction, so it is regarded as secondary delay in one direction and the primary delay in the other direction.



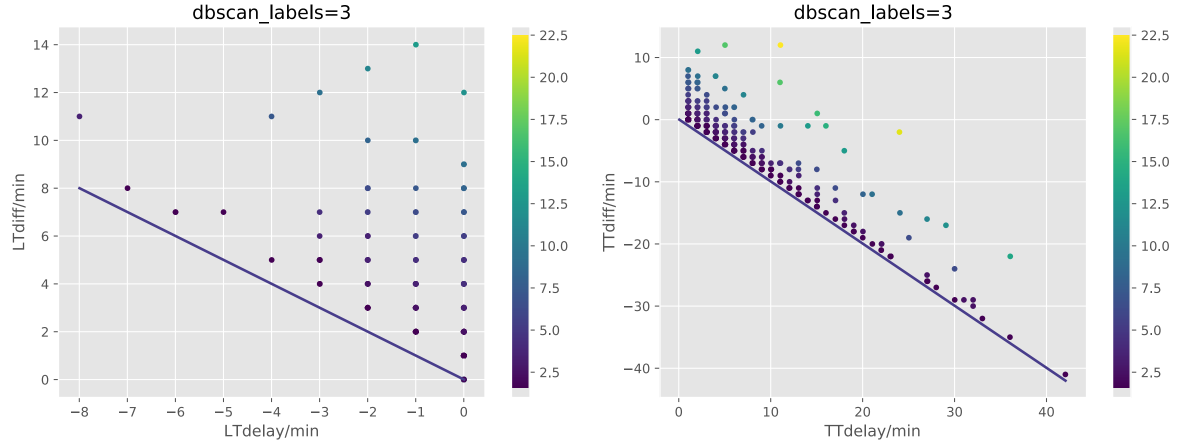
a) Scatter diagram of $LTdelay_{i,s-1}$ and $LTdiff_{s,s-1}$ b) Scatter diagram of $TTdelay_{i-1,s}$ and $TTdiff_{i,i-1}$

Fig. 6 Visualization of Category 1



a) Scatter diagram of $LTdelay_{i,s-1}$ and $LTdiff_{s,s-1}$ b) Scatter diagram of $TTdelay_{i-1,s}$ and $TTdiff_{i,i-1}$

Fig. 7 Visualization of Category 2



a) Scatter diagram of $LTdelay_{i,s-1}$ and $LTdiff_{s,s-1}$ b) Scatter diagram of $TTdelay_{i-1,s}$ and $TTdiff_{i,i-1}$

Fig. 8 Visualization of Category 3

Table 5 Results of four types of delays

Delay type	Category	Horizontal direction	Vertical direction	Result
1	0	No delay propagation	No delay propagation	Primary delay
2	1, 4, 5, 7, 8	Delay spread horizontally	Delay spread vertically	Secondary delay
3	2, 6	No delay propagation	Delay spread vertically	
4	3	Delay spread horizontally	No delay propagation	

5.3 Model performance

On the basis of determining the delay type, construct a training set containing the label of the delay type, and determine the K value of the KNN classification model to be 4. In order to verify the performance of GBDT-PF model, this paper will compare the random forest (RF), support vector regression (SVR) and multilayer perceptron (MLP) that are widely used in train delay prediction, and will also verify the importance of the delay propagation feature. The optimal parameter combination of four models is shown in Table 6. This paper uses three evaluation indicators include root mean square error (RMSE), mean absolute error (MAE) and coefficient of determination (R2) to evaluate the parameter combination. Table 7 shows the RMSE, MAE and R2 of each model.

Table 6 The optimal parameter combination of four models

Model	Parameter	The value under different feature combinations	
		Z_i	Z_i,$P_{i,s-1}$,$P_{i-1,s}$
GBDT	learning_rate	0.49	0.06
	n_estimators	96	91
	min_samples_split	300	284
	min_samples_leaf	2	3
	max_depth	5	17
	max_feature	5	3
RF	n_estimators	80	90
	max_features	4	6
	max_depth	7	9
SVR	C	3.2	2.1
	loss	epsilon_insensitive	epsilon_insensitive
MLP	hidden_layer_sizes	(20,20,20)	(80,80,80)

Table 7 The index values of each model under different feature combinations

Model	Input feature	RMSE	MAE	R2
RF	Z_i	1.53354	0.51600	0.94404
	Z_i,$P_{i,s-1}$,$P_{i-1,s}$	1.44952	0.44150	0.95000
SVR	Z_i	1.67973	0.79100	0.93286
	Z_i,$P_{i,s-1}$,$P_{i-1,s}$	1.58231	0.62100	0.94042
MLP	Z_i	1.69797	0.64000	0.93140
	Z_i,$P_{i,s-1}$,$P_{i-1,s}$	1.66474	0.60150	0.93405
GBDT	Z_i	1.40097	0.40600	0.95330
	Z_i,$P_{i,s-1}$,$P_{i-1,s}$	1.35860	0.37400	0.95608

After adding the delay propagation feature, the index value of each model is optimized. Therefore, considering the impact of the delay propagation in the delay prediction can improve the prediction accuracy. Among the four models, the GBDT model with delay propagation feature performs better on three indicators. Fig. 9 shows the distribution of the prediction errors of each model.

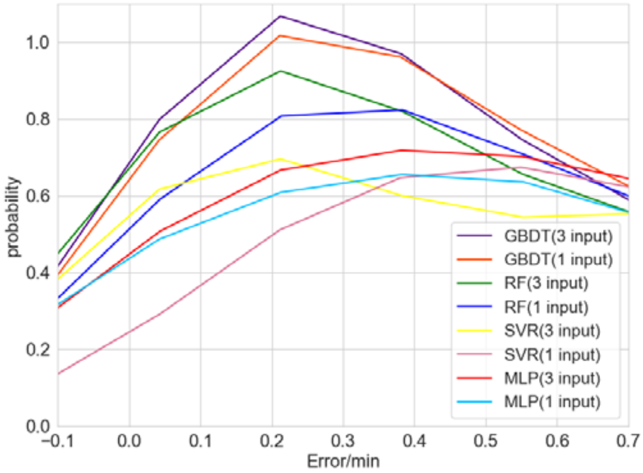


Fig. 9 Error distribution of each models under different feature combinations

It can be seen from Fig. 9 that the peak value of the model error distribution curve containing the delay propagation feature is closer to the vertical axis, indicating that the overall error is smaller. The error distribution curve of the GBDT model with delay propagation feature is closest to the vertical axis in all models, so its overall error is the smallest and the prediction accuracy is higher.

5.4 Model simulation

In order to display the prediction results of the delay prediction model more intuitively, this paper uses the PYQT5 package to complete the simulation of the model on the Pycharm software. The simulation program design process completed by PYQT5 is shown in Fig. 10.

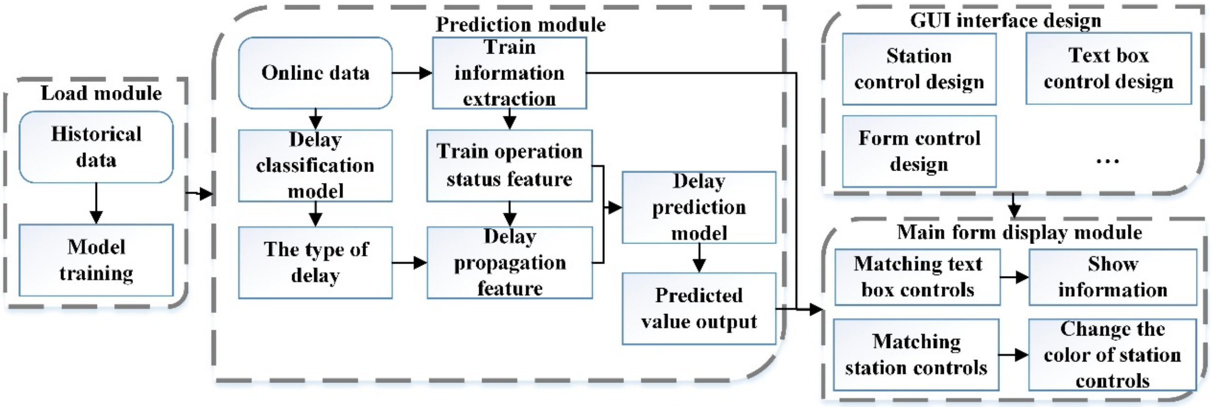
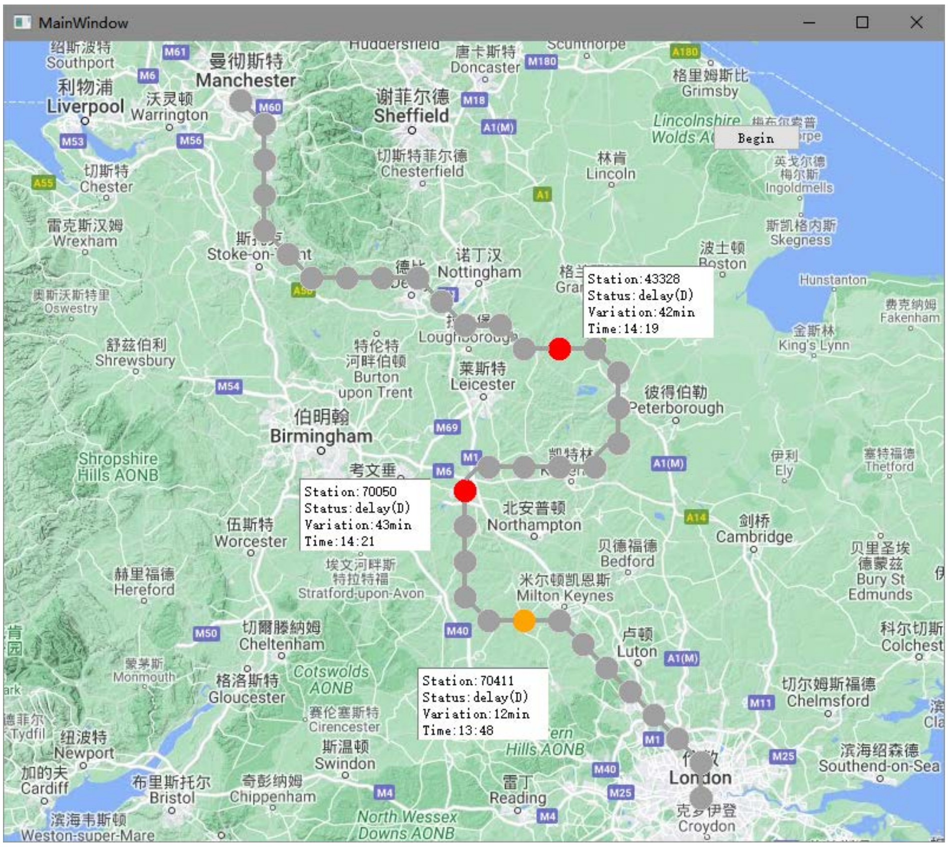


Fig. 10 The design process of simulation program for train delay prediction based on PYQT5



Station:43328
Status:delay(D)
Variation:42min
Time:14:19

Station:70050
Status:delay(D)
Variation:43min
Time:14:21

Station:70411
Status:delay(D)
Variation:12min
Time:13:48

Fig. 11 Train delay prediction simulation interface

There are two main functions of the simulation program: (1) It can display the train information, station information, departure or arrival time in real time. At the same time, different colors are used to indicate the current degree of delay. Green means the train is on time, blue means the train is early, yellow means the train is delayed within 5 minutes, orange means 5-15 minutes delay, and red means more than 15 minutes delay; (2) According to train operation data and prediction model, display the operation status of each train on the line dynamically. According to the simulation interface, the prediction results can be viewed in real time, and the delay propagation phenomenon can be observed. Fig. 11 shows the operation of the three delayed trains on the line through simulation interface.

6. Conclusion

This paper proposes a GBDT-PF model that considers the delay propagation feature. The effectiveness of the method is evaluated by taking the train operation data of the British WCML line as an example, and the following conclusions are drawn:

- Based on the characteristics of primary delay and secondary delay in the delay propagation, using DBSCAN algorithm to design a clustering method of delay types for historical data, through this method, the delays can be finally divided into four categories. The four types of delays have obvious characteristics in the vertical and horizontal direction. And according to the best delay classification scheme, the KNN algorithm is used to design the classification method for online data to identify the type of delay.
- Based on the results of the identification of delay types, the delay propagation relationship is quantified by the delay propagation factor and used as the input feature of the GBDT model. According to the experimental comparison results, when predicting train delays, considering the delay propagation feature can further improve the prediction accuracy.

With the development of railway informatization, based on the comprehensive collection of actual train operation data, the dispatching and commanding of railway trains will also be more intelligent. The delay prediction model proposed in this paper can provide delay prediction data for intelligent dispatch and make the dispatching and command work more efficient.

Acknowledgement

This work was supported by Sichuan Science and Technology Program (2021YJ0070).

References

[1] Huang, P., Peng, Q., Wen, C., Yang, Y. (2018). Random forest prediction model for Wuhan-Guangzhou HSR primary train delays recovery, *Journal of the China Railway Society*, Vol. 40, No. 7, 1-9.

[2] Wen, C., Li, Z., Huang, P., Tian, R., Mou, W., Li, L. (2019). Progress and perspective of data-driven train delay propagation, *China Safety Science Journal*, Vol. 29, No. S2, 1-9, doi: 10.16265/j.cnki.issn1003-3033.2019.S2.001.

[3] Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., Anguita, D. (2018). Train delay prediction systems: A big data analytics perspective, *Big Data Research*, Vol. 11, 54-64, doi: 10.1016/j.bdr.2017.05.002.

[4] Wang, P., Zhang, Q.-P. (2019). Train delay analysis and prediction based on big data fusion, *Transportation Safety and Environment*, Vol. 1, No. 1, 79-88, doi: 10.1093/tse/tdy001.

[5] Shi, R., Xu, X., Li, J., Li, Y. (2021). Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization, *Applied Soft Computing*, Vol. 109, Article No. 107538, doi: 10.1016/j.asoc.2021.107538.

[6] Nair, R., Hoang, T.L., Laumanns, M., Chen, B., Cogill, R., Szabó, J., Walter, T. (2019). An ensemble prediction model for train delays, *Transportation Research Part C: Emerging Technologies*, Vol. 104, 196-209, doi: 10.1016/j.trc.2019.04.026.

[7] Li, Z.-C., Wen, C., Hu, R., Xu, C., Huang, P., Jiang, X. (2020). Near-term train delay prediction in the Dutch railways network, *International Journal of Rail Transportation*, Vol. 9, No. 6, 520-539, doi: 10.1080/23248378.2020.1843194.

[8] Huang, P., Wen, C., Fu, L., Lessan, J., Jiang, C., Peng, Q., Xu, X. (2020). Modeling train operation as sequences: A study of delay prediction with operation and weather data, *Transportation Research Part E: Logistics and Transportation Review*, Vol. 141, Article No. 102022, doi: 10.1016/j.tre.2020.102022.

[9] Gao, B., Ou, D., Dong, D., Wu, Y. (2020). A data-driven two-stage prediction model for train primary-delay recovery time, *International Journal of Software Engineering & Knowledge Engineering*, Vol. 30, No. 7, 921-940, doi: 10.1142/S0218194020400124.

[10] Tang, Y., Xu, C., Wen, C., Li, Z., Song, S. (2019). Support vector regression models for delay time predicting considering high-speed rail facility failure, *China Safety Science Journal*, Vol. 29, No. S2, 18-23, doi: 10.16265/j.cnki.issn1003-3033.2019.S2.003.

[11] Zhang, Q., Chen, F., Zhang, T., Yuan, Z.M. (2019). Intelligent prediction and characteristic recognition for joint delay of high speed railway trains, *Acta Automatica Sinica*, Vol. 45, No. 12, 2251-2259, doi: 10.16383/j.aas.c190188.

[12] Hu, R., Xu, C., Feng, Y., Wen, C., Wang, Q. (2019). Prediction of different types of train delay of Guangzhou-Shenzhen high-speed railway, *China Safety Science Journal*, Vol. 29, No. S2, 181-186, doi: 10.16265/j.cnki.issn 1003-3033.2019.S2.030.

[13] Zeng, Y., Chen, F., Jin, B. (2019). A prediction model for timetable delays in dispatching area using neural network, *Railway Standard Design*, Vol. 63, No. 3, 148-153, doi: 10.13238/j.issn.1004-2954.201812160002.

[14] Milinković, S., Marković, M., Vesković, S., Ivić, M., Pavlović, N. (2013). A fuzzy Petri net model to estimate train delays, *Simulation Modelling Practice and Theory*, Vol. 33, 144-157, doi: 10.1016/j.simpat.2012.12.005.

[15] Lessan, J., Fu, L., Wen, C. (2019). A hybrid Bayesian network model for predicting delays in train operations, *Computers & Industrial Engineering*, Vol. 127, 1214-1222, doi: 10.1016/j.cie.2018.03.017.

[16] Pullagura, L., Katiravan, J. (2019). Train delay prediction using machine learning, *International Journal of Engineering and Advanced Technology (IJEAT),* Vol. 9, No. 2, 1312-1315, doi: 10.35940/ijeat.A2088.129219.

[17] Huang, P., Wen, C., Fu, L., Peng, Q., Tang, Y. (2020). A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems, *Information Sciences*, Vol. 516, 234-253, doi: 10.1016/j.ins.2019.12.053.

[18] Hansen, I.A., Goverde, R.M.P., van der Meer, D.J. (2010). Online train delay recognition and running time prediction, In: *Proceedings of 13th International IEEE Conference on Intelligent Transportation Systems,* Funchal, Portugal, 1783-1788, doi: 10.1109/ITSC.2010.5625081.