A structured data preprocessing method based on hybrid encoding

Chang Liu, Liu Yang and Jingyi Qu*
Tianjin Key Laboratory of Advanced Signal Processing, Civil Aviation University of China, Tianjin, 300300, China
* Corresponding author’s e-mail: jyqu@cauc.edu.cn

Abstract. With the rapid development of civil aviation transportation industry, the passenger throughput in civil aviation is increasing, while the problem of flight delays is becoming more and more serious. For flight delay prediction under big data, deep learning methods can be applied to make high-precision predictions. Since data preprocessing is one of the most important parts, the method based on hybrid encoding is proposed in this paper. Firstly, the flight and meteorological data are fused with the associated primary key, since weather data has a greater impact on flight delay. Then, the fused data is encoded according to different data types. Min-Max encoding is used for continuous features, and CatBoost encoding is adopted for discrete features respectively. Finally, the data set which has been preprocessed can be put into the deep convolutional neural network ResNet to verify the effect. The experimental results show that the prediction accuracy rate of flight delay level can reach 94.02% on the structured data set after hybrid encoding.

1. Introduction

The development of aviation transportation industry is very rapid. With the support of big data storage technology, civil aviation industry has accumulated massive flight data and related meteorological data. These data provide reliable data sources for the field of big data scientific. Data preprocessing is one of the most important steps to make full use of data. It can not only improve the quality of data, but also shorten the calculation process and effectively improve the accuracy and performance of flight delay prediction. Therefore, it is really important to obtain high-quality data sets after processed.

Data is generally divided into structured data and unstructured data. The processing technology of unstructured data has been very mature in the field of deep learning. For example, for video data processing, a Tensor RPCA (Robust principal component analysis) method is proposed in literature[5], which separates the images in video according to low rank and sparsity, makes model learning faster and better. At the same time, with the development of deep learning framework, structured data can be processed by a specific data processing library, without the help of feature engineering and specific knowledge in this field, the workload will be greatly reduced.

In view of the preprocessing problem of structured data such as flight data, scholars at home and abroad have used many methods. Sun Choi et al. have transformed category variables into numerical variables [1], and standardized the data. Khanmohammadi Sina et al. normalized delay data by dividing them to max values [2]. Luo et al. used the phase space reconstruction method in chaos theory to map the nonlinear sequence to the high dimensional space [3], effectively obtained the dynamic characteristic information in the nonlinear sequence, and used the normalization method of norm to preprocess the data. Cheng et al. segmented the time data to optimize the characteristic data [4]. However, the data...
processed by these methods are not under the same dimension, and the coding method is relatively single, which is not targeted to different types of data.

2. Introduction of data sets
Data preprocessing is very important for the training results of the model. The data sets used in this paper is the flight and meteorological data of Shanghai Hongqiao Airport, which is provided by the China Civil Aviation Administration of East China.

2.1. Flight Data
Planned flight data from May 2018 to May 2019 is used, and several functions of this data set are shown in Table 1.

Table 1. Introduction features of flight data.

| Feature            | Example | Meaning                     |
|--------------------|---------|-----------------------------|
| PlanDestAirport    | ZGSZ    | Planned arrival airport     |
| PlanAircraft       | A321    | Planned aircraft type       |
| TailNumber         | B8540   | Aircraft tail wing          |
| TaskType           | S       | Task type                   |
| GuaranteeType      | S       | Types of safeguards         |
| GuaranteeLevel     | NML     | Security level              |

2.2. Meteorological Data
Meteorological data are observations from 2018 to 2019, updated hourly. It contains observation station location and meteorological information, in which location information contains airport name, so we can find the corresponding observation station of airport, get the corresponding meteorological data of airport, and provide the possibility for subsequent data fusion. Table 2 shows the meteorological data properties field.

Table 2. Introduction features of weather data.

| Feature   | Example      | Meaning                  |
|-----------|--------------|--------------------------|
| Record_time| 201901200002 | Data and time            |
| 01_35R_TP | 5.8          | Temperature              |
| 01_35R_RD | 76           | Relative humidity        |
| 02_34L_FP | 1029.06      | Field pressure           |
| 01_35R_DT | 1.91         | Dew point                |
| 01_35R_CSP | 1029.67     | Modified Sea Pressure    |

3. Data Preprocessing
It is necessary to preprocess the data before using a neural network to extract the features. The data preprocessing includes data cleaning, data fusion and data encoding. After encoding, the data put into the RNN (Recurrent Neural Network) will be serialized, and the data put into the CNN (Convolutional Neural Networks) will be matrixed. By preprocessing, all categorical variables are converted to numerical variables since deep learning algorithms exhibit a better performance with numerical variables. The whole data preprocessing is shown in Figure 1.
3.1. **Data Cleaning**

Data cleaning is an essential step in data preprocessing, which mainly deals with the null values and outliers in the data. For large data sets, when the eigenvalues are missing, the method adopted in this paper is to delete the flight or the meteorological data, when the data has an abnormal value, mostly appears in the meteorological data, the method is filling the data with the average value of the features.

3.2. **Data Fusion**

Since making features of data putting into the model completely and diversely is beneficial for the neural network to extract the information more fully between features, the flight and meteorological data is fused in this paper. Before fusing it, we need to know the associated information of the data and take the information as the key value. Hongqiao airport is taken as an example below.

The flight data and corresponding meteorological data of the airport are extracted firstly. because the meteorological data is updated every hour, but the time of flight data is not strictly consistent with the observation time of meteorological data, so the second step is extracting the hourly information of the departure or landing time of the flight, and then extracting the time information of observation in the meteorological data. Finally, take the region and time as the key value to fuse the flight data and meteorological data.

3.3. **Data Encoding**

In order to obtain the corresponding matrix or vector and put them into the model for learning, it is always necessary to encode the original data in deep learning algorithms. It can be seen from the examples in Table 1 and Table 2 that data features include both continuous and discrete features. For example, features such as departure time and date are all given as continuous values, while flight number, airport ID, and weather conditions are all given as discrete values. The different methods will be introduced in this section to encode the different type of data. Figure 2 shows the flow chart of data encoding.
3.3.1. Continuous features.

Since continuous features have different ranges of dimension, it will have influences on data analysis and process. The influences mainly include the following two points. One is that the features with large dimensional range will be in dominant, which will increase the weight of the feature artificially. The other one is that the iterative convergence rate will drop slowly, which means the performance of the model is not good enough. In order to eliminate the influence of different feature dimensions and increase the contrast between data, the continuous features are normalized through Min-Max encoding, and all the eigenvalues are mapped to the range of 0 to 1. The function of data normalization is shown in equation (1).

\[
\hat{x} = \frac{x - \text{min}}{\text{max} - \text{min}}
\]  

Where $x$ is the original data, $\hat{x}$ is the normalized data, min is the minimum value in the sample data, and max is the maximum value in the sample data.

3.3.2. Discrete features.

For discrete data, its characteristics are not sequential, so it can be divided into low based category data and high based category data. Low base category data can be encoded using One-Hot encoding [8], which can make the category under each feature attribute have independent registers, thus ensuring that the encoding is uniquely valid. Although this coding expands the features, it makes the data sparse.

The feature attributes of the fused data set are very complex, especially the flight data with high cardinal number, such as departure city ID and airline ID, etc. The feature attributes contain hundreds of categories. If processed by one-hot coding, the size explosion will occur, and the data will be very sparse, which is not conducive to the learning of subsequent algorithms. In the dataset used in this paper, discrete data are almost all high base class data, so the following methods are proposed to deal with high base class data.

(a) Count encoding. Count encoding uses the number of times that a certain characteristic category appears to replace the category, which is characterized by frequency statistics. This coding method will not produce dimension explosion, but different categories will appear the same frequency, coding the two categories into the same number will lead to low accuracy of model learning.

(b) Label encoding. Label encoding uses the sequence of categories under a certain feature attribute, that is, the category is replaced by 1 to n-1 sequential numbers. This encoding method does not cause a dimensionality explosion, but adds order to high-based data that is not sized, which will also lead to low accuracy of the algorithm model.

(c) CatBoost encoding. CatBoost encoding [7] is a supervised coding method, which converts high base class data into numerical data based on statistics, it can not only avoid dimension explosion, but also not make the encoded data sequential. At the same time, the encoded data is in the same dimensional range. CatBoost encoding works as follows:

![Figure 2. Presentation of serialization process.](image-url)
Assume that the data set is \( D = (X_i, Y_i)_{i=1,...,n} \), where \( X_i = (x_{i,1},...,x_{i,m}) \) is a vector containing \( m \) features, \( Y_i \in R \) is the label value. The coding will shuffle the data set randomly first, assuming that the random sort sequence is \( \sigma = (\sigma_1, \sigma_2, ..., \sigma_n) \), then the following category can be expressed as

\[
x_{\sigma, j, k} = \frac{\sum_{i=1}^{n-1} [x_{\sigma_i, k} = x_{\sigma_j, k}] * Y_{\sigma_j} + \hat{\theta} * P}{\sum_{j=1}^{n-1} [x_{\sigma_j, k} = x_{\sigma_j, k}]} + \hat{\theta}
\]  

(2)

Where \( P \) is the added prior term, \( \hat{\theta} \) is generally a weight coefficient greater than 0. By adding these two terms, the influence of noise and low frequency category data on the data distribution can be reduced.

In regression problems, the prior term \( P \) is expressed as the mean of the data label value. In classification problems, such as dichotomous, the prior term \( P \) is expressed as the probability of positive examples.

3.3.3. Hybrid encoding.
Because the feature attributes of the fused data set are complex, the sample size is large, and it contains continuous features and discrete features, a hybrid coding method is proposed. Min-Max coding is used for continuous features and CatBoost encoding for discrete features.

3.4. Serialization
Considering the time correlation of the flight data, the flight data input to the Recurrent Neural Network is serialized to construct a time series. A fixed time series length is used in this paper, the serialization process is shown in Figure 3.

First, the data set \( E \) is sorted according to the flight operation time to obtain the data set \( E_t \), and then the window with the step of length \( L \) is used for sliding segmentation, and a number of sequence data with a length of \( L \) are obtained as the input of the network.

![Figure 3. Presentation of serialization process.](image)

3.5. Matrix
For the data input into the Convolutional Neural Network, it is necessary to do matrix, one-dimensional data is converted by size, so that it will be convenient to input into the network. The matrix process is shown in figure 4, where \( x_1, x_2, ..., x_n \) are represented as the sample data of each input into the model,
$f_1, f_2, ..., f_m$ are represented as different feature attributes, $x'_1, x'_2, ..., x'_n$ represent the matrixized data after size conversion respectively.

![Matrix Process Diagram](image)

**Figure 4. Presentation of matrix process.**

## 4. Analysis of results

The experimental environment and parameter configuration are introduced in this section, then the effects of different encoding methods will be discussed through three indicators.

### 4.1. Experimental environment and parameter configuration

The configuration of the computer used in the experiment is as follows: the processor is the Intel® Xeon® E5-2630, and the CPU frequency is 3.6GHz with the memory of 16.00 GB. The operating system is the Ubuntu16.04.3 operating system, while the GPU accelerated graphics card is the GTX TITAN Xp. The deep learning platform and data preprocessing library are Tensorflow 2.3.0, Numpy 1.18.5, Pandas 1.0.5, and scikit-learn 0.23.1 respectively.

| Parameter name                        | Parameter value |
|----------------------------------------|-----------------|
| The maximum number of iterations       | 100             |
| Loss function                          | Cross entropy   |
| Optimizer                              | Adam            |
| Learning rate                          | 0.0001          |
| Number of batches during training      | 128             |

### 4.2. Comparison of data set results with different encoding methods

According to the three encoding methods described in this paper, the dimensions of the data after Label encoding and Count encoding are pretty large, which is not conducive for the algorithm model to learn the correlation between features. In deep learning algorithms, since larger value of the dimension will increase the learning weight and cause the wrong direction of the network, normalization is required to reduce the dimension range after these two encoding methods.
There are 2 data samples and 7 features are selected from the data set, and the results of the three encoding methods are shown in Table 4.

Table 4. The result of different coding methods.

| Feature       | Raw data | Min-Max + Label encoding | Min-Max + Count encoding | Min-Max + CatBoost encoding |
|---------------|----------|---------------------------|---------------------------|-----------------------------|
| PlanDestAirport | ZGSZ     | 0.229885057               | 0.269451132               | 0.413828149                 |
|               | ZSYT     | 0.781609159               | 0.05982385                | 0.497508658                 |
| PlanAircraft  | A321     | 0.083333                  | 0.50469558                | 0.373616692                 |
|               | B738     | 0.305556                  | 1                         | 0.383826702                 |
| TailNumber    | B8540    | 0.916222                  | 0.600175554               | 0.345605457                 |
|               | B5785    | 0.538462                  | 0.111915734               | 0.525628716                 |
| TaskType      | S        | 0.5                       | 1                         | 0.365917744                 |
| GuaranteeType | S        | 0.5                       | 1                         | 0.365917744                 |
| GuaranteeLevel| NML      | 0.5                       | 0.378189722               | 0.378189722                 |

It can be seen from the Table 4 that for different samples of the same feature, both “Min-Max + Label encoding” and “Min-Max + Count encoding” could have an encoding result of 0 in some discrete features, which makes the original feature meaningless and weakens the correlation between the them, and also have no advantage to the analysis of subsequent data and effect of algorithm models. The “Min-Max + CatBoost encoding” method encodes discrete feature supervisory with a statistically-based way, which can encode data effectively and provide a favorable support for the subsequent learning of algorithm models.

4.3. The impact of coding methods on prediction accuracy

In order to verify the impact of different encoding methods on the flight delay prediction model, different encoding methods are used preprocess the data which is put into the ResNet neural network [9] under the same variable controlled. The trends of loss and accuracy is shown in Figure 5-7, and the results are shown in Table 5.

Figure 5. The trends of loss and accuracy with Label encoding. (a) shows that the trend of loss with Label encoding, (b) shows that the trend of accuracy with Label encoding.

The trends in Figure 5 show that the model has no phenomenon of over-fitting or under-fitting, but there is a large fluctuation between 10th to 20th epoch. What’s more, the final accuracy rate of the training and test set is not high enough.
Figure 6. The trends of loss and accuracy with Count encoding. (a) shows that the trend of loss with Count encoding, (b) shows that the trend of accuracy with Count encoding.

It can be seen from Figure 6 that the model has no over-fitting or under-fitting, but fluctuations occurred during the period from 10th to 40th epoch, the convergence of the train and test set is not very satisfactory either.

Figure 7. The trends of loss and accuracy with CatBoost encoding. (a) shows that the trend of loss with CatBoost encoding, (b) shows that the trend of accuracy with CatBoost encoding.

The trends of the loss and accuracy value of the training and test set in Figure 7 show that the model has no over-fitting or under-fitting. Compared with the other two methods of encoding, this method has a better performance in training process. The loss value dropped extraordinary smoothly, the accuracy rose gently, and the final accuracy is relatively high.

Table 5 clearly shows the accuracy value of flight delay prediction with the three different encoding methods.

| Encoding            | Accuracy /% |
|---------------------|-------------|
| Min-Max + Label encoding | 81.53       |
| Min-Max + Count encoding  | 83.99       |
| Min-Max + CatBoost encoding | 94.02      |
It can be seen from the experimental results, as for a flight delay prediction model using ResNet neural network with the above three methods of encoding, CatBoost encoding has better results than the other two encoding methods. What’s more, for structured data, especially the data that fused with flight and meteorological information, the hybrid encoding method composed of Min-Max and CatBoost encoding is more applicable.

5. Conclusion
In dealing with the structured data set that fused flight and meteorological data, a preprocessing method is proposed based on hybrid encoding in this paper. By processing the data set of Shanghai Hongqiao Airport, this method shows a better result. The main conclusions are as follows:

1) Since the dimension of the data between features is totally different after Label encoding and Count encoding, the neural network will increase the weight for the larger value in small dimension, which can cause a wrong direction of learning.

2) CatBoost encoding does not encode all the same data into 0, which can preserve the meaning of the feature. After encoding, the data between features have the same dimension, so there is no need for normalization.

3) The hybrid encoding method adopts targeted coding for different types of data, and unifies the coding results into the same dimension, the performance of ResNet has been improved.

When processing structured data, the data preprocessing method based on hybrid coding is more suitable, which can reduce the workload of the model and improve the accuracy of the flight delay algorithm model.

Acknowledgments
This work was supported by the Natural Science Foundation of China (U1833105), Tianjin Natural Science Foundation (19JCYBJC15900) and the Fundamental Research Funds for the Central Universities (3122019185).

References
[1] Choi, S., Kim, Y. J., Briceno, S., Mavris, D., (2016) Prediction of weather-induced airline delays based on machine learning algorithms, In: Digital Avionics Systems Conference, Sacramento. CA. USA, pp. 1-6.
[2] Sina, K., Salih, T., Yunus, K., (2016) A new multilevel input layer artificial neural network for predicting flight delays at JFK airport, Procedia Computer Science, 95: 237-244.
[3] Luo, Y. Q., Chen, Z. J., Tang, J. H., Zhu, Y. W., (2015) Flight delay prediction using support vector machine regression, Journal of Transportation Systems Engineering and Information Technology, 15: 143-149.
[4] Cheng, H., Li, Y. M., Luo, Q., Li, C., (2014) Study on flight delay with C4.5 decision tree based prediction method, Systems Engineering Theory & Practice, 34: 239–247.
[5] Candes, E. J., Li, X. D., Ma, Y., Wright, J., (2011) Robust principal component analysis, Journal of the ACM, 58: 11-11.
[6] Ioffe, S., Szegedy, C., (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift, In: International Conference on Machine Learning, Lille. France, pp. 448–456.
[7] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., Gulin, A., (2018) CatBoost: unbiased boosting with categorical features, In: Conference on Neural Information Processing Systems, Montréal. Canada, pp.6639-6649.
[8] Wu, R. B., Li, J. Y., Qu, J. Y., (2018) Flight delay prediction based on dual-channel convolutional neyral networks, Journal of Computer Applications, 38:2100- 2112.
[9] He, K. M., Zhang, X. Y., Ren, S. Q., Sun, J., (2016) Deep residual learning for image recognition, IEEE, In: Conference on Computer Vision & Pattern Recognition, Las Vegas. Nevada. USA, pp. 770-778.