Jie Li, Zhongcan Li, Chao Wen, Qiyuan Peng & Ping Huang

**Train operation conflict detection for high-speed railways: a naïve Bayes approach**

International Journal of Rail Transportation, 2022

DOI: 10.1080/23248378.2022.2071346

Примитивный (исходный, упрощенный) метод Байеса

Abstract

Accurately detecting train operation conflicts (TOC) has great significance for improving the emergency handling ability of dispatchers during interference. In this study, a conflict detection model for high-speed train operation is proposed, with the train operation data from Xiamen to Shenzhen high-speed railway. Firstly, a TOC detection model framework considering data imbalance is determined, based on Bernoulli naïve Bayes model. Then, the hyper-parameter of the proposed model is tuned with the training and validation dataset. Next, the performance result of the proposed model is compared to other three commonly used naïve Bayes models, namely the Gaussian naïve Bayes, multinomial naïve Bayes and complement naïve Bayes. Comparison analyses based on the commonly used classification model evaluation indexes show that the detection accuracy of the proposed model is significantly higher than other naïve Bayes models. The proposed model also achieves high robustness and detection accuracy in each category.

Точное обнаружение конфликтов движения поездов (TOC) имеет большое значение для улучшения способности диспетчеров реагировать на чрезвычайные ситуации во время появления препятствий. В этом исследовании предлагается модель обнаружения конфликтов при движении высокоскоростных поездов с использованием данных о поездной работе от Сямыня до Шэньчжэня на высокоскоростной железной дороге. Во-первых, определяется структура модели обнаружения TOC, учитывающая дисбаланс данных, основанная на исходной байесовской модели Бернулли. Затем настраивается гиперпараметр предлагаемой модели с использованием совокупности данных для обучения и проверки. Затем результат работы предложенной модели сравнивается с тремя другими широко используемыми наивными байесовскими моделями, а именно, с гауссовской наивной байесовской моделью, полиномиальной наивной байесовской моделью и дополнительной наивной байесовской моделью. Сравнительный анализ, основанный на обычно используемых индексах оценки классификационных моделей, показывает, что точность обнаружения предлагаемой модели значительно выше, чем у других наивных байесовских моделей. Предлагаемая модель также обеспечивает высокую надежность и точность обнаружения в каждой категории.

1. Introduction

Train operation conflict (TOC) occurs when the preceding train occupies the routes or equipment of its following train [1]. Advanced TOC detection is essential for train rescheduling, train delay management, and automatic train operation control [2]. Accurate detection of TOC is of great significance to improve train operation rescheduling and dispatcher’s emergency handling ability under interference conditions, and railway transport departments must improve their operational efficiency and provide high-quality transport services to passengers [3,4].

Конфликт (TOC) возникает, когда предыдущий поезд занимает маршруты или оборудование, предназначенное для следующего за ним поезда [1]. Усовершенствованное обнаружение TOC необходимо для изменения расписания поездов, управления задержками поездов и автоматического управления движением поездов [2]. Точное обнаружение ТОС имеет большое значение для улучшения планирования движения поездов и способности диспетчера действовать в аварийных ситуациях в условиях препятствий, а подразделения железнодорожного транспорта должны повышать эффективность своей работы и обеспечивать качественное транспортное обслуживание пассажиров [3,4].

The NB model

In the naïve Bayes model, each pair of features are conditional independent in given categories. T={(x1,y1), (x2,y2)…(xn,yn)} is a sample generated from independent and identical distribution P(X ,Y). Suppose that each sample has M features, labelled as yE{c1, c2,…cM}

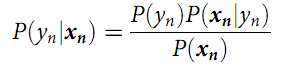
The naïve Bayes model can learn the joint probability distribution P(X,Y) of the data from the

training data, that is, the data prior probability distribution P(y=cm) and the conditional

probability distribution P(X=xn|Y=cm). According to Bayes’ theorem, given characteristic

variables and labels, their relationship can be expressed as follow:

В наивной байесовской модели каждая пара признаков условно независима в заданных категориях. T={(x1,y1), (x2,y2)…(xn,yn)} — выборка, сгенерированная из независимого и идентичного распределения P(X ,Y). Предположим, что каждая выборка имеет M признаков, помеченных как yE{c1, c2,…cM}. Наивная байесовская модель может изучать совместное распределение вероятностей P(X,Y) данных из обучающих данных, то есть, априорное распределение вероятностей P(y=cm) и условное распределение вероятностей P(X=xn|Y =см). Согласно теореме Байеса, при заданных характеристических переменных и признаках их взаимосвязь может быть выражена следующим образом:



There is an assumption in the naïve Bayes, that is the naïve Bayes is conditionally independent

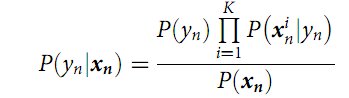
under the condition that the classification is determined. Considering the assumption, Equation (2)

can be got:

В наивном Байесе есть допущение, что наивный Байес условно независим при условии, что определено выделение (классификация). С учетом допущения можно получить уравнение (2):



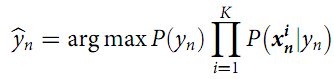
Thus, the above formula can be written as Equation (3):



Since the input data is given and P(xn) is a constant. Therefore, the naïve Bayes classifier can be

expressed as Equation (4):

Поскольку входные данные заданы и P(xn) является константой. Следовательно, наивный байесовский классификатор может быть выражен уравнением (4):



The difference between different types of naïve Bayes models lies in their different prior assumptions of the data.

Разница между различными типами наивных байесовских моделей заключается в их различных априорных предположениях о данных.

**3.2. Down-sampling methods Методы уменьшения выборок**

Down-sampling is a technique for data processing. Down-sampling is performed by using a filter to

reduce undesired noise in the observations, e.g., to reduce the frequency or resolution of signal/data.

The down-sampling reduces the resolution while keeping the exact representation of the observa-

tions. The down-sampling method can result in a small subset of the majority class (es) for

classification problems so that the training data is less skewed. This technique is expected to

improve the performance of the classification model on the minority class (es).

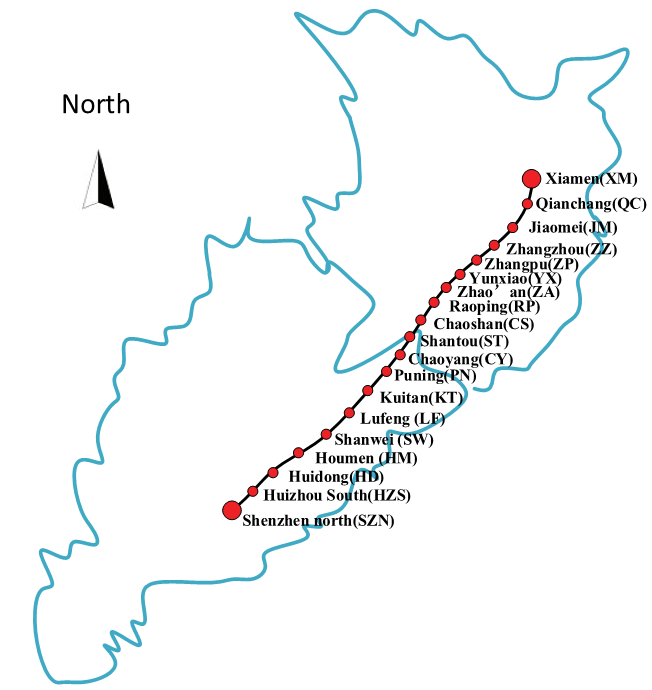
Уменьшение выборок — это метод обработки данных. Понижающая дискретизация выполняется с использованием фильтра для уменьшения нежелательного шума в наблюдениях, например, для уменьшения частоты или разрешения сигнал/данные. Понижающая дискретизация снижает разрешение, сохраняя при этом точное представление наблюдений. Метод уменьшенной выборки может привести к небольшому подмножеству большинства классов для задач классификации, чтобы обучающие данные были менее искаженными. Ожидается, что этот метод улучшит производительность модели классификации для класса (ов) меньшинства.

**4. Train operation conflict detection model**

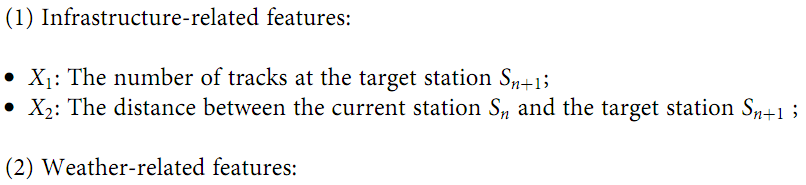
**4.1. Data description and analysis**

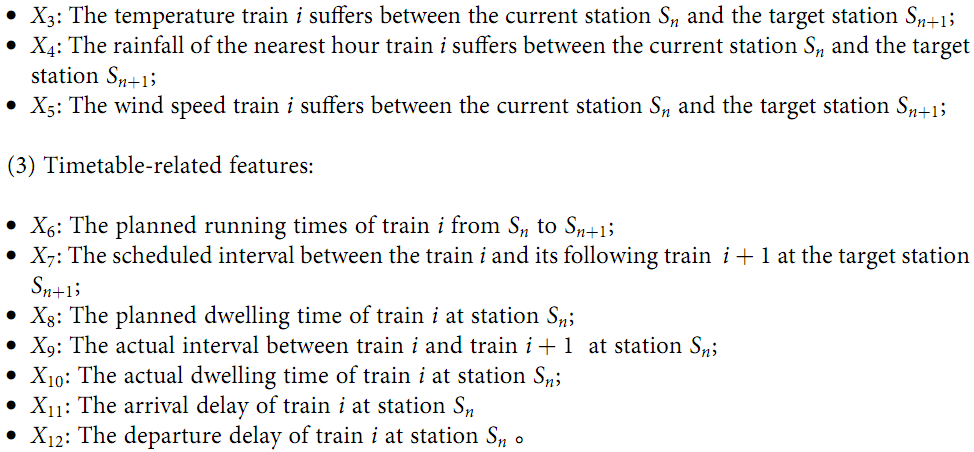
Линия 543 км, 8 участков, 9 станций. Данные с марта 2015 по ноябрь 2016. Около 41 тыс. поездов.

Задача: зная момент прибытия на станцию Sn, определить, вызовет ли поезд конфликт на станции Sn+1.



Учитывались следующие факторы:





According to the definition of TOC in Section 2, this paper marks the following situation as train

operation conflicts: (1) the time difference between the train’s arrival time at the target station and

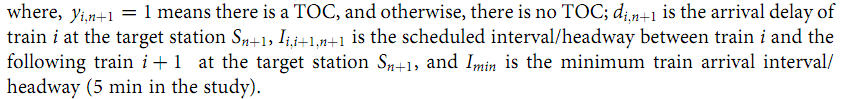
the scheduled arrival time of the next train at the station is less than I min, and (2) the arrival time of

the train at the target station is later than the scheduled arrival time of the next train at the station.

That is, the mark y1,n+1 of each sample is shown in Equation(5).

В соответствии с определением ТОС в разделе 2, в данной работе в качестве конфликтов движения поездов отмечена следующая ситуация: (1) разница во времени между временем прибытия поезда на целевую станцию и плановым временем прибытия следующего поезда на станцию меньше, чем 1 мин, и (2) время прибытия поезда на целевую станцию позже планового времени прибытия следующего поезда на станцию. То есть метка y1,n+1 каждой выборки показана в уравнении (5)





After determining the influence factors of TOC and the conflict marking standard, a total of

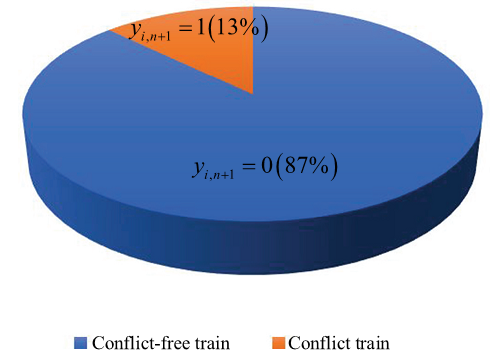
324,563 cases are extracted from the dataset of the above eight stations. The distribution character-

istics of the conflict trains and conflict-free trains are shown in Figure 3. Figure 3 shows that the

TOC detection data is severely unbalanced data, where the conflict-free samples are more than the

conflict samples.

После определения факторов влияния на TOC и стандарта конфликтной маркировки, из набора данных восьми вышеупомянутых станций было извлечено в общей сложности 324 563 случая. Характеристики распределения конфликтных и бесконфликтных поездов показаны на рисунке 3.



Из рисунка 3 следует, что данные обнаружения TOC являются сильно несбалансированными данными, где бесконфликтных выборок больше, чем конфликтных.

**4.2. Variable dependency investigation**

**5.2. Performance results on conflict propagations/interactions**

In Section 5.1, we shows the performance results of the proposed detection model at the nearest

station (i.e., the target is to detect whether there are conflicts between trains at the next station).

However, train operations are interactive due to infrastructure and signal systems’ interlocking.

This means that conflicts are dependent between adjacent trains and stations. Once a conflict occurs

between two adjacent trains at a station, it will likely propagate to the future stations and the

following trains. Therefore, in this section, we investigate the performance of the proposed frame-

work on conflict propagation (to the future stations and to the following trains). Take Figure 1 as an

example. If the goal of the model in Section 5.1 is to detect the conflict between Train i and Train

i þ1 at Sn, the models in this section are to (1) detect the conflicts between Train i and Train i þ1 at

Snþ1 (Task 1), and (2) detect the conflicts between Train i þ1 and Train i þ2 at Sn (Task 2).

В разделе 5.1 мы показываем результаты работы предложенной модели обнаружения на ближайшей станции (т. е. цель состоит в том, чтобы обнаружить, есть ли конфликты между поездами на следующей станции). Однако движение поездов является интерактивным из-за взаимосвязи инфраструктуры и сигнальных систем. Это означает, что конфликты распространяются между соседними поездами и станциями. Как только конфликт возникает между двумя соседними поездами на станции, он, вероятно, распространится на будущие станции и последующие поезда. Поэтому в этом разделе мы исследуем эффективность предложенного метода в задаче распространения конфликтов (на будущие станции и следующие поезда). Возьмите рисунок 1 в качестве примера. Если целью модели в разделе 5.1 является обнаружение конфликта между поездом i и поездом i +1 на Sn, модели в этом разделе должны (1) обнаруживать конфликты между поездом i и поездом i +1 на Sn+1 (задача 1) и (2) обнаруживать конфликты между поездом i +1 и поездом i +2 на Sn (задача 2).

4.5. Parameter tuning of the model

The critical parameters of the model are optimized before applying the model for detection. The Bernoulli naïve Bayes (BNB) model is chosed as the conflict detection model. The BNB model assumes that the conditional distribution of training data follows the Bernoulli distribution. Generally, three hyper-parameters can be optimized for the BNB model to improve the performance, including the smoothing parameter, the threshold for binarizing the samples, and the prior probability for the classes. We experiment and show the influences of these parameters on the performance of the BNB model. However, the results show that only the second parameter (i.e., the threshold for binarizing the samples) substantially influences the BNB model’s performance.

Критические параметры модели оптимизируются перед применением модели для обнаружения. Наивная байесовская модель Бернулли (BNB) выбрана в качестве модели обнаружения конфликтов. Модель BNB предполагает, что условное распределение обучающих данных соответствует распределению Бернулли. Как правило, для повышения производительности модели BNB можно оптимизировать три гиперпараметра, включая параметр сглаживания, порог для бинаризации выборок и априорную вероятность для классов. Мы экспериментируем и показываем влияние этих параметров на производительность модели BNB. Однако результаты показывают, что только второй параметр (то есть, порог для бинаризации выборок) существенно влияет на производительность модели BNB.

We thus mainly optimize this critical

parameter, i.e., the threshold of the model’s dualization of input features. This value

determines the critical point of the binarization of the input sample features. If the sample

feature values are greater than the critical value, it is discretized and marked as 1; otherwise,

it is marked as 0. Therefore, this value has a significant impact on the performance of the

model. This parameter is set to 0 by default in the Scikit-learn package [49]. In this study,

we chose more reasonable values from −0.2 to 0.1 with an interval of 0.01. We also

experiment on the smaller intervals (e.g., 0.001). However, the smaller intervals rarely

influence the performance of the model. The results with an interval of 0.01 are shown in

Figure 6. It can be seen from the figure that when the critical value is −0.14, the model

achieved the highest conflict detection accuracy on both the training dataset and the testing

dataset, i.e., 90.6% and 90.5%, respectively. Here, because we apply the down-sampling

technique on the imbalanced data, the accuracy could be used as an evaluation metric.

Therefore, we chose −0.14 as the threshold for binarizing the input features, which means

when the model detects the TOC, it automatically marks the input feature values which are

greater than the threshold value as 1; otherwise, they are marked as 0.