Обнаружение конфликтов при эксплуатации поездовна высокоскоростных железных дорогах: простойподход Байеса

АННОТАЦИЯ

Точное обнаружение конфликтов движения поездов (TOC) имеет большое значение для улучшения способности диспетчеров справляться с аварийными ситуациями во время взаимодействия. В этом исследовании предлагается модель обнаружения конфликтов для движения высокоскоростных поездов с использованием данных о движении поездов от высокоскоростной железной дороги Сямынь до Шэньчжэня. Во-первых, определяется структура модели обнаружения TOC, учитывающая дисбаланс данных, на основе наивной байесовской модели Бернулли. Затем гиперпараметр предлагаемой модели настраивается с помощью набора данных для обучения и валидации. Затем результат производительности предлагаемой модели сравнивается с тремя другими широко используемыми наивными байесовскими моделями, а именно с гауссовой наивной байесовской, многочленной наивной байесовской и дополняющей наивный байесовский. Сравнительный анализ, основанный на обычно используемых оценочных индексах классификационных моделей, показывает, что точность обнаружения предлагаемой модели значительно выше, чем у других naiпяти байесовских моделей. Предлагаемая модель также обеспечивает высокую надежность и точность обнаружения в каждой категории.

о введении

Несмотря на современное оборудование связи и систему управления, высокоскоростные поезда неизбежно сталкиваются с помехами, создаваемыми инфраструктурой, людьми и окружающей средой, что приводит к конкуренции маршрутов и оборудования. Конфликт эксплуатации поезда (TOC) возникает, когда предыдущий поезд занимает маршруты или оборудование следующего поезда [1]. Расширенное обнаружение TOC необходимо для перепланирования поездов, управления задержками поездов и автоматического управления работой поездов [2]. Точное определение TOC имеет большое значение для улучшения планирования движения поездов и способности диспетчера реагировать на чрезвычайные ситуации в условиях помех, и департаменты железнодорожного транспорта должны повысить эффективность своей работы и предоставлять пассажирам высококачественные транспортные услуги [3,4]. Поскольку ТОС является основной причиной задержек поездов, исследования по обнаружению ТОС всегда были целью полевых исследователей и практиков [5,6]. Таким образом, компьютерное моделирование, математическая оптимизация и графовые и сетевые модели являются тремя наиболее часто используемыми методами для TOC. Компьютерное моделирование в основном включает имитационные модели и системы моделирования. Методы компьютерного моделирования в основном основаны на моделях или системах, которые были разработаны для моделирования работы поездов в реальном мире. Чтобы спрогнозировать TOC, Вен и др. смоделировали работу поездов в системе диспетчерского моделирования [7]. Они изучили разумные пропорции поездов с разными уровнями и выявили влияние помех на работу поездов с разными пропорциями поездов. Кроме того, было проведено множество исследований для улучшения управления работой поездов во время сбоев, таких как использование теории массового обслуживания для построения модели прогнозирования задержек [8], использование теории блочного времени для разработки модели обновления скорости [9], использование надлежащих карт перехода состояний для прогнозирования конфликтов [10], использование стохастической графовой модели для имитации аварийных задержек [11,12] и использование вероятностного символьного алгоритма для имитации железнодорожных расписаний. Кроме того, многие системы моделирования работы поездов, такие как LUKS [13], RailSys [14], OpenTrack [15] и системы прогнозирования задержек на высокоскоростной железной дороге Синкансэн [16], были разработаны и широко используются железнодорожными исследователями и менеджерами по всему миру. Однако параметры в компьютерных имитационных моделях и системах в основном зависят от гипотетических и теоретических моделей, например, от настройки задержек и возмущений в соответствии с конкретными моделями распределения.

Проблема разрешения TOC обычно рассматривается как проблема изменения расписания в условиях сбоев. Следовательно, наиболее распространенными подходами являются методы математической оптимизации [2,5]. В связи с этим инструменты математической оптимизации, такие как целочисленное программирование [17,18], смешанное целочисленное программирование [19-21] и нелинейное целочисленное программирование [22,23], обычно используются для достижения целей при изменении расписания, таких как максимизация удовлетворенности пассажиров, минимизация эксплуатационных расходов, задержек поездов и времени ожидания пассажиров, или комбинации двух или более целей [2,5]. Проблема разрешения конфликтов возникла в 1970-х годах [24]. Хиггинс и др. [25] предложили модель для оптимизации расписания поездов в однопутных железнодорожных коридорах. Модель была разработана с двумя основными приложениями, а именно (1) как инструмент поддержки принятия решений поездными диспетчерами при составлении расписания движения поездов и (2) как инструмент планирования для оценки влияния изменений расписания и инфраструктуры. Основываясь на проблеме конфликта между поездами,Ахин предложил эвристический алгоритм для изменения маршрута поезда путем изменения существующего расписания прибытия и отправления в случае однопутной железной дороги [26]. Была разработана двухцелевая модель оптимизации расписания для сокращения как времени ожидания пассажиров, так и выбросов углекислого газа в системе метро [27], и результаты показали, что предложенная структура модели может обеспечить хорошие показатели по обеим целям по сравнению с существующим расписанием. Дендар и др. [28] предложили метод разрешения конфликтов, основанный на генетическом алгоритме, и оценили его на основе операций диспетчеров и точных решений алгоритма. Интерактивная модель смешанного целочисленного программирования была разработана и внедрена в испанской национальной железнодорожной компании [29]. Авторы использовали эвристическую процедуру, основанную на обратном отслеживании, чтобы сократить время поиска решений. Чу и др. [30] отметили, что составление расписания является типичной проблемой NP-полноты, и они сравнили подходы генетического алгоритма (GA) и табуированного поиска (TS) для решения проблемы составления расписания. Результаты эксперимента показали, что подход TS показал лучшие результаты, чем подход GA, по срокам выполнения и эффективности. Все подходы к математической оптимизации были основаны на определенных допущениях, которые неизбежно вредят их практическому применению в реальных ситуациях.

Поскольку операции движения поездов состоят из событий движения поездов, таких как события прибытия, прохождения и отправления, графические и сетевые модели, которые могут описывать события движения поездов, широко используются для обнаружения TOC [2,3,31-33]. Капрара и др. [31] использовали направленные множественные графики для получения метода моделирования бесконфликтной работы поездов на однопутной железной дороге. Д'Ариано и др. [32] предложили эффективный метод обнаружения и разрешения конфликтов, основанный на альтернативной графовой модели, с учетом таких факторов, как остановка поезда, интервал отслеживания, безопасное расстояние и скорость поезда. Авторы использовали ту же графовую модель, то есть альтернативный граф, для поддержки диспетчеров в повседневной задаче управления помехами [33]; В этом исследовании использовался алгоритм с ответвлениями и границами для изменения порядка движения поездов, в то время как был разработан алгоритм локального поиска для изменения маршрута движения поездов с помощью системы управления движением в реальном времени. Беттинелли и др. [34] рассмотрели безопасность эксплуатации следующих транспортных средств, когда некоторые ресурсы были недоступны из-за помех и других факторов в диспетчеризации в режиме реального времени, а затем предложили пространственно-временную сеть, основанную на итеративном жадном алгоритме для разрешения TOC; Предлагаемый алгоритм определил исходные правила планирования поездов и итерационные методы с помощью множественной последовательности и еще больше улучшил производительность модели за счет принятия различных разреженных методов для пространственно-временной сети. Ван и др. [35] рассматривали

проблема перепланирования расписания как процесс обнаружения и разрешения конфликтов; Затем для моделирования железнодорожной системы были использованы временные цветные сети Петри, где места представляли ресурсы в железнодорожной системе, а жетоны представляли поезда. Кайми и др. [36] смоделировали операции поездов как конфликтный граф, в котором каждый рассматриваемый путь поезда соответствовал вершине, а ребра представляли попарные конфликты. Таким образом, бесконфликтное расписание соответствовало максимально независимому набору. Для каждого ресурса авторы могут эффективно определить максимальную конфликтную группу, сканируя время блокировки всех путей движения поездов и принимая его за жизненно важную разделительную плоскость целочисленного линейного программирования. Модель нечеткой сети Петри (FPN) для оценки задержек поездов, в которой знания экспертов использовались для определения нечетких множеств и правил преобразования экспертных знаний в модель для расчета задержек поездов, была предложена в [37]. Модель адаптивной сетевой системы нечеткого вывода (ANFIS), основанная на исторических данных о задержках поездов в системе железных дорог Сербии, была создана [38]. Модель состояния движения ВСМ, основанная на треугольных сетях рабочего процесса с нечеткими числами и нечетким временем движения поездов, была сгенерирована с использованием данных с пяти станций в системе ВСМ Пекин-Шанхай за период с 21 по 24 июня 2012 года [39]. Однако все графические и сетевые методы построены на предположении о свойстве Маркова, что означает, что операция поезда связана только с его последним состоянием. Хотя это упрощение может позволить проводить рассуждения и вычисления, которые в противном случае были бы трудноразрешимыми, оно чрезмерно упрощает работу поездов и пренебрегает многими критическими факторами, влияющими на работу поездов.

Анализ показывает, что существующие исследования по TOC в основном сосредоточены на разрешении конфликта. Однако существует нехватка исследований проблемы обнаружения TOC с точки зрения данных. Поэтому авторы представленного исследования используют модель машинного обучения, основанную на байесовской теории, для изучения обнаружения TOC по данным высокоскоростной железной дороги Сямынь-Шэньчжэнь (X-S HSR). Модель обучается, валидируется и тестируется на основе данных о работе поездов. Методы исследования могут помочь диспетчерам заранее прогнозировать оглавление, тем самым поддерживая принятие решений при диспетчеризации в режиме реального времени. Кроме того, выводы исследования могут дать диспетчерам лучшее понимание структуры оглавлений, что может повысить эффективность эксплуатации и диспетчеризации железнодорожных менеджеров.

о постановке задачи

Во время эксплуатации высокоскоростные поезда неизбежно будут испытывать помехи внутри или за пределами железнодорожной системы. Поскольку помехи сохраняются до определенной степени, поезд будет задержан. Когда продолжительность задержки поезда превышает буферное время между поездами, предыдущий поезд будет занимать маршрут следующего поезда, вызывая конфликты в работе поездов. На рис. 1(а) показана принципиальная схема конфликта движения высокоскоростного поезда на пространственно-временной диаграмме, где горизонтальная ось представляет время, а вертикальная ось - пространство (станцию). Пунктирная линия представляет запланированную линию движения поездов, а сплошная линия представляет фактическую линию движения поездов. На поезд i влияют помехи на станции Sn, и он задерживается на сутки;n минут. Когда di; n превышает запланированное буферное время между поездом i и поездом i 1, поезд i будет занимать маршрут поезда i 1 на станции Sn, вызывая конфликт в работе поезда. Другими словами, если задержки поезда i больше, чем разница во времени с запланированным интервалом Ii;i+1;n между поездом i и поездом i +1 и минимальным интервалом (Imin), указанным в техническом документе, поезд i вызывает конфликт между поездом i и поездом i + 1. Аналогично, при последующей эксплуатации поездов (на станции Sn+2, Sn+3 и SK), если di; n+1, di; n+2,di; K , больше, чем Ii; i+1; n+1 - Imin, Ii;i+1; n+2 - Imin и Ii;i+1; K - Imin, соответственно, поезд i также вызовет конфликты на станции Sn+1, Sn+2 и SK . Следовательно, мы определяем конфликт операций поезда как:

Определение 1. Конфликт в работе поезда: задержки поезда i в Sn di; n больше, чем разница во времени запланированного интервала Ii; i+1; n между поездом i и поездом i + 1 и минимальным интервалом (Imin), указанным в техническом документе, показанном на рисунке 1(a), или поезд i отправляется позже запланированного времени поезда i + 1, показанного на рисунке 1(b).

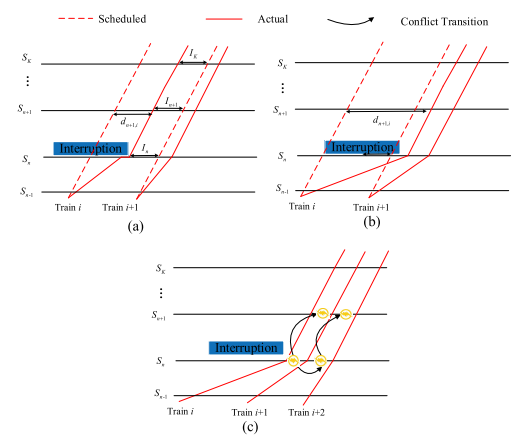


Рисунок 1 -указание на конфликт в работе поезда.

В этом исследовании используется известная информация от текущей и прошлых станций для определения TOC на ближайшей будущей станции. Обнаружение конфликта между поездом i и поездом i+1 на станции Sn взято в качестве примера, который показан на рисунке 1. В качестве входных данных используется известная информация о поезде i и поезде или бывших станциях.

Кроме того, чтобы исследовать распространение TOC между поездами, мы изучаем модель обнаружения, учитывающую зависимости / взаимодействия конфликтов между поездами и станциями. Возьмем в качестве примера рисунок 1(с). Сначала мы обучаем базовую модель (предположим, что она называется Model 1) обнаруживать конфликты между поездом i и поездом i 1 на станции Sn. Затем, чтобы продемонстрировать взаимодействие конфликтов между станциями, мы используем выходные данные базовой модели в качестве входных данных другой модели обнаружения (предположим, что она называется Model 2), и используем оба

Модели 1 и 2 для обнаружения конфликта поездов i и i 1 на станции Sn+1 . Кроме того, чтобы продемонстрировать взаимодействие конфликтов между соседними поездами, мы используем выходные данные базовой модели в качестве входных данных другой модели обнаружения (предположим, что она называется Model 3), и используем как модель 1, так и Модель 3 для обнаружения конфликта поезда i и поезда i 1 на станции Sn + 1. Здесь входные данные модели 1, Модели 2 и Модели 3 постоянны, но модели 2 и 3 принимают в качестве входных данных больше информации, т.е. Результатов обнаружения модели 1. Следовательно, эффекты распространения TOC могут быть количественно оценены путем сравнения модели 1 с моделью 2 и

Модель 3.

о способе

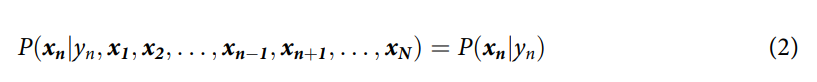
В существующих исследованиях в основном использовалось математическое программирование и графовые модели для изучения проблемы обнаружения конфликтов при эксплуатации поездов [2,5,32,33]. Однако эти методы неизбежно строятся на конкретных допущениях модели или упрощениях (например, модели математического программирования обычно предполагают, что задержки / распределения задержек известны, а методы, основанные на графах, оцениваются на основе марковского свойства для упрощения операций с поездами). Хотя эти методы могли бы обеспечить приемлемые решения для процесса эксплуатации поездов, они обычно чрезмерно упрощают проблему. Эксплуатация поездов - это динамический процесс в сложных системах. Состояния систем / поездов меняются во времени и пространстве, а факторы влияния (например, сбои инфраструктуры, плохая погода и человеческие ошибки) сложны и их трудно поддаются количественной оценке. Таким образом, особенности железнодорожных систем и недостатки математического программирования или моделей, основанных на графах, ограничивают применимость традиционных моделей обнаружения конфликтов на практике. Следовательно, необходимы передовые методы обучения обнаружению движения поездов, чтобы повысить практичность и удобство использования при диспетчеризации поездов в режиме реального времени. В этом отношении данные о работе поездов, содержащие исторические характеристики системы, могут лучше представлять функции и особенности системы. Кроме того, исторические данные являются следствием всех влияющих факторов, что позволяет модели, построенной на основе этих данных, давать более реалистичные результаты. Фактически, при наличии данных о работе поездов методы, основанные на данных, основанные на теореме Байеса, широко использовались для решения задач эксплуатации поездов, включая байесовские сети для прогнозирования задержек поездов, прогнозирования влияний и анализа надежности расписания [4,40,43]. Эти методы показали превосходство по сравнению с их аналогами. Поэтому мы инновационно применяем одну из широко используемых моделей, т.е. наивную модель Байеса, для обнаружения конфликтов при эксплуатации поездов. Ожидается, что предлагаемый метод преодолеет ограничения традиционных моделей обнаружения конфликтов при эксплуатации поездов и будет более практичным.

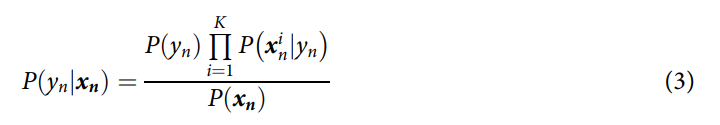
o Модель NB

Наивный Байес (NB) [44] - это модель контролируемого обучения, основанная на теореме Байеса. Принцип алгоритма относительно прост для понимания и устойчив к отсутствующим данным. В наивной модели Байеса каждая пара признаков условно независима в заданных категориях. T=  - это образец, сформированный из независимых и идентичных имения П Х; г . Предположим, что каждый образец имеет M характеристики, маркировку . Наивная модель Байеса может изучать совместное распределение вероятностей P X; Y данных из обучающих данных, то есть распределение вероятностей до данных Py cm и условное распределение вероятностей P X xn Y cm . Согласно теореме Байеса, учитывая характерные переменные и метки, их взаимосвязь может быть выражена следующим образом:

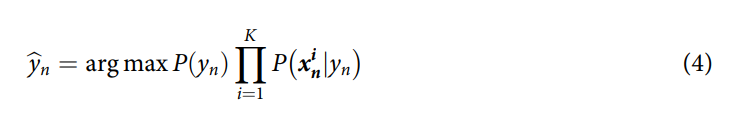


В наивном Байесе есть предположение, что наивный Байес является условно независимым при условии, что классификация определена. Учитывая предположение, уравнение (2) может быть получено:

Таким образом, приведенную выше формулу можно записать в виде уравнения (3):



Поскольку входные данные заданы, а P X является константой. Следовательно, наивный байесовский классификатор может быть выражен в виде уравнения (4):



Разница между различными типами наивных байесовских моделей заключается в их различных предварительных допущениях относительно данных.

o Методы понижающей выборки

Нисходящая выборка - это метод обработки данных. Понижающая выборка выполняется с использованием фильтра для уменьшения нежелательного шума в наблюдениях, например, для уменьшения частоты или разрешения сигнала / данных. Понижающая выборка уменьшает разрешение, сохраняя точное представление наблюдений. Метод понижающей выборки может привести к получению небольшого подмножества мажоритарного класса (es) для задач классификации, так что обучающие данные будут менее искаженными. Ожидается, что этот метод улучшит производительность классификационной модели для класса меньшинств (es).

Широко используемые методы понижающей выборки включают (1) случайную понижающую выборку (RD), (2) одностороннюю понижающую выборку (OSD), (3) понижающую выборку с близким пропуском (NMD) и (4) отредактированную понижающую выборку ближайшегососеда (ENND). Метод случайной выборки с уменьшением количества выборок недооценивает класс большинства (es) путем случайного выбора выборок с заменой / без замены. Односторонний метод понижающей выборки приводит к занижению выборки на основе метода односторонней выборки [45]. Метод понижающей выборки с близким промахом приводит к занижению выборки на основе метода близкого промаха [46]. Отредактированный метод понижающей выборки ближайшегососеда основан на отредактированном методе ближайшего соседа [47]; в этом методе данные очищаются путем удаления выборок, близких к границе принятия решения. В этом исследовании мы исследуем эти четыре метода понижающей выборки, чтобы улучшить производительность модели обнаружения конфликтов при работе поездов.

o Модель обнаружения конфликтов при эксплуатации поездов

o Описание и анализ данных

Данные о движении поездов, используемые в этом документе, получены от высокоскоростной железной дороги Сямынь-Шэньчжэнь (X-S HSR), как показано на рисунке 2. Протяженность X-S HSR составляет около 543 км, она оснащена системой управления поездом CTCS-2, которая позволяет управлять поездом с максимальной скоростью 250 км /ч. Данные, использованные в этом исследовании, включают поезда из сегмента Хуэйдун (HD) в сегмент Чаошань (CS), всего девять станций и восемь участков. Временной интервал используемых данных составляет около 20 месяцев с 24 марта 2015 года по 10 ноября 2016 года. В течение этого временного интервала из сегмента HD в CS ВСМ X-S курсирует около 41 000 поездов.

Случаи конфликтов при эксплуатации поездов извлекаются из данных о работе поездов. Возьмем в качестве примера случай, представленный в разделе 2. Если поезд i только прибывает в Sn, и мы намерены определить, вызовет ли поезд i конфликт на станции Sn + 1. В качестве входных данных модели обнаружения мы рассматриваем следующие факторы, которые могут оказать влияние на работу поездов:

o Особенности, связанные с инфраструктурой:

o X1: количество дорожек на целевой станции Sn+1;

o X2: расстояние между текущей станцией Sn и целевой станцией Sn+1 ;

o Особенности, связанные с погодой:



Рисунок 2. Принципиальная схема X-S HSR.

o X3: температурный диапазон i изменяется между текущей станцией Sn и целевой станцией Sn+1;

o X4: количество осадков, выпадающих на ближайшем часовом поезде i между текущей станцией Sn и целевой станцией Sn+1;

o X5: поезд скорости ветра i страдает между текущей станцией Sn и целевой станцией Sn+1;

o Функции, связанные с расписанием:

o X6: запланированное время движения поезда i от Sn до Sn+1;

o X7: запланированный интервал между поездом i и следующим за ним поездом i + 1 на целевой станции

Sn+1;

o X8: запланированное время прибытия поезда i на станцию Sn;

o X9: фактический интервал между поездом i и поездом i 1 на станции Sn;

o X10: фактическое время нахождения поезда i на станции Sn;

o X11: Задержка прибытия поезда i на станцию Sn

o X12: Задержка отправления поезда i на станции Sn°

В соответствии с определением TOC в разделе 2, в настоящем документе следующая ситуация обозначается как конфликты при эксплуатации поездов: (1) разница во времени между временем прибытия поезда на целевую станцию и запланированным временем прибытия следующего поезда на станцию меньше, чем Imin, и (2) время прибытия

поезд прибывает на целевую станцию позже запланированного времени прибытия следующего поезда на станцию. То есть отметка yi;n+1 для каждого образца показана в уравнении(5).

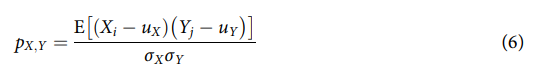


где, yi; n+1 = 1 означает, что есть ОГЛАВЛЕНИЕ, а в противном случае оглавления нет; di; n+1 - задержка прибытия поезда i на целевую станцию Sn+1, Ii; i+1; n+1 - запланированный интервал движения между поездом i и следующим поездом i 1 на целевую станцию Sn+1, а Imin - минимальный интервал прибытия поезда (в исследовании 5 минут).

После определения факторов влияния TOC и стандарта конфликтной маркировки из набора данных вышеупомянутых восьми станций было извлечено в общей сложности 324 563 случая. Характеристики распределения конфликтных поездов и поездов без конфликтов показаны на рисунке 3. На рисунке 3 показано, что данные обнаружения TOC представляют собой сильно несбалансированные данные, где бесконфликтных выборок больше, чем конфликтных выборок.

o Исследование зависимости переменных

Сначала мы исследовали их взаимозависимости с определенными факторами влияния (входными данными модели). Мы использовали корреляцию произведение-момент Пирсона, уравнение (6) для количественной оценки их взаимосвязей. Коэффициенты частичной корреляции для каждых двух переменных приведены в таблице 1. Следует отметить, что чем больше абсолютные значения, тем сильнее корреляции между переменными. Значения, близкие к нулю, указывают на отсутствие существенных зависимостей между переменными.



Где, pX; Y - корреляция между произведением Пирсона и моментом данных X и Y, u представляет средние значения, и ? представляет стандартное отклонение.

Результаты в таблице 1 показывают, что абсолютные коэффициенты корреляции переменных невелики, что указывает на тонкие зависимости между переменными. Это не относится к переменным X11 и X12. Это понятно, поскольку переменные X11 и X12 представляют задержки прибытия и отправления поездов. Результаты в таблице 1 показывают различия между переменными.

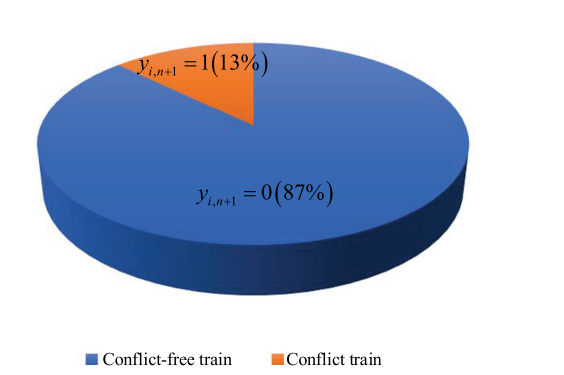
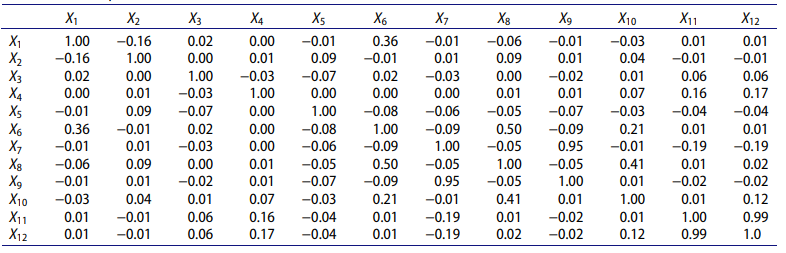


Рисунок 3. Доля конфликтных поездов и поездов без конфликтов в данных о движении поездов.

Таблица 1. Коэффициенты корреляции между произведением Пирсона и моментом для переменных.



o Структура модели

Результаты на рисунке 3 показывают , что данные о конфликте операций поездов сильно несбалансированы . Предположим, что классификатор используется непосредственно для классификации данных. В этом случае это приведет к низкой производительности классификатора на выборках с низкой частотой, поскольку модель легко предсказывает низкочастотные классы как высокочастотные классы. В этом случае высокий балл обычно может быть достигнут за счет использования общих показателей оценки модели, таких как точность. Однако модель прогнозирует только то, что все выборки будут бесконфликтными. Возьмем в качестве примера распределение меток на рисунке 3. Если классификатор предсказывает, что все выборки будут бесконфликтными, точность набора данных достигнет 87%. Чтобы решить упомянутую выше проблему, вызванную несбалансированностью данных, в данном исследовании предлагается структура фрейма модели, показанная на рисунке 4.

Шаг 1: Перед моделированием данных мы сначала используем метод понижающей выборки [48] для набора данных. Целью понижающей выборки является преобразование несбалансированных данных в сбалансированные. Наиболее сложной задачей является настройка частоты понижающей выборки для обеспечения надежности модели. В этом исследовании мы используем стандартный метод для установки частоты дискретизации с понижением [48], случайным образом выбирая 1 случай из бесконфликтных выборок (R - отношение бесконфликтных последовательностей к конфликтным последовательностям). Наконец, 81 707 примеров, содержащих 41232 конфликтных выборки и 40 475 бесконфликтных выборок, используются для обучения и тестирования модели.

Шаг 2: Стандартизируйте оставшиеся данные и преобразуйте каждый объект в стандартное нормальное распределение, чтобы исключить влияние измерений.

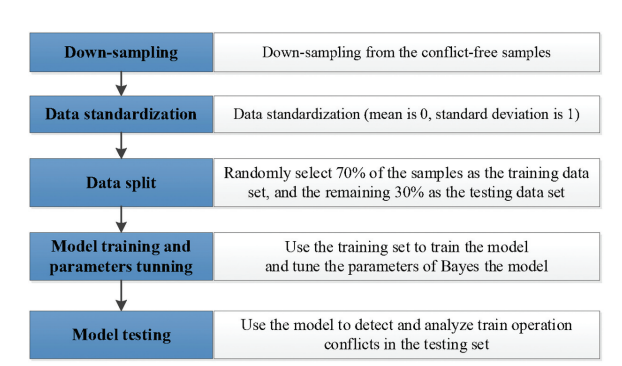


Рисунок 4. Структура модели обнаружения конфликтов.

Шаг 3: Случайным образом выбирается 70% выборок для обучения модели, включая 57 194 случая, в то время как остальные 30% выборок используются для тестирования модели, включая 24 513 случаев.

Шаг 4: Набор обучающих данных используется для обучения модели и настройки параметров наивной байесовской модели.

Шаг 5: Набор тестовых данных используется для проверки эффекта обнаружения конфликтов обученной модели

o Выбор метода понижающей выборки

Перед обучением моделей NB мы сначала исследуем влияние методов понижающей выборки на результаты обнаружения. При использовании различных методов понижающей выборки могут быть получены разные данные при обучении и тестировании модели. Здесь мы сравниваем результаты обнаружения четырьмя методами (т.Е. RD, OSD, NMD и ENND), представленными в разделе 3.2. Чтобы получить сбалансированные данные (т. Е. Каждый класс с одинаковым / схожим размером выборки), мы сначала выполняем методы RD, OSD, NMD и ENND соответственно. Затем для обучения модели BNB используются четыре группы данных с одинаковым объемом данных. Здесь мы используем наивную байесовскую модель Бернулли (BNB), чтобы продемонстрировать эффективность модели обнаружения на данных, полученных с помощью различных методов понижающей выборки, как показано на рисунке 5. Этот рисунок ясно показывает, что модель BNB на основе данных, полученных методом случайной понижающей выборки, имеет наивысшую точность. Поэтому мы используем метод случайной понижающей выборки в качестве метода понижающей выборки для балансировки данных о работе поездов. Это означает, что модели обнаружения конфликтов будут обучаться на данных, полученных методом случайной выборки с понижением.

o Настройка параметров модели

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 могут быть оптимизированы 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.

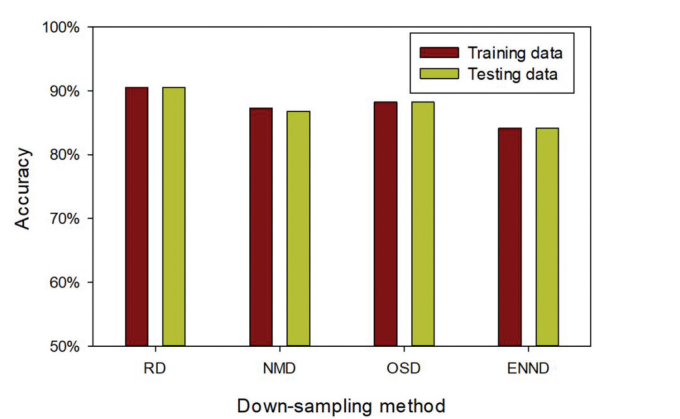


Рисунок 5- Производительность модели BNB при различных методах понижающей дискретизации.

Однако результаты показывают, что только второй параметр (т.е. порог для бинаризации выборок) существенно влияет на производительность модели BNB. Таким образом, мы в основном оптимизируем этот критический параметр, т.е. порог дуализации входных характеристик модели. Это значение определяет критическую точку бинаризации объектов входной выборки. Если значения выборочного признака превышают критическое значение, оно дискретизируется и помечается как 1; в противном случае оно помечается как 0. Следовательно, это значение оказывает существенное влияние на производительность модели. Этот параметр установлен в 0 по умолчанию в пакете Scikit-learn [49]. В этом исследовании мы выбрали более приемлемые значения от 0,2 до 0,1 с интервалом 0,01. Мы также экспериментируем с меньшими интервалами (например, 0,001). Однако меньшие интервалы редко влияют на производительность модели. Результаты с интервалом 0,01 показаны на рисунке 6. Из рисунка видно, что при критическом значении 0,14 модель достигла наивысшей точности обнаружения конфликтов как в обучающем, так и в тестовом наборе данных, т.е. 90,6% и 90,5% соответственно. Здесь, поскольку мы применяем метод понижающей выборки к несбалансированным данным, точность может быть использована в качестве оценочного показателя. Поэтому мы выбрали ?0.14 в качестве порога для бинаризации входных объектов, что означает, что когда модель обнаруживает TOC, она автоматически помечает значения входных объектов, которые больше порогового значения, как 1; в противном случае они помечаются как 0.

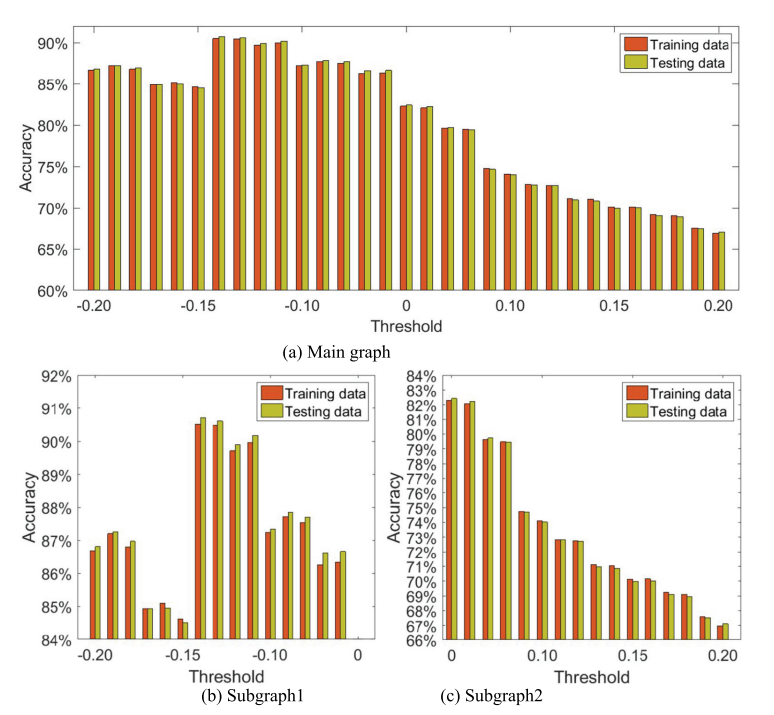


Рисунок 6. Результаты оптимизации критического значения (а) основного графика (б) подграфа1 (в) Подграфа2.

о Результатах

В этом разделе мы анализируем эффективность предлагаемой модели BNB при обнаружении TOC. Чтобы систематически оценивать производительность модели, мы также выбрали три другие широко используемые наивные байесовские модели в качестве эталона: гауссовскую наивную байесовскую (GNB), полиномиальную наивную байесовскую (MNB) и дополняющую наивную байесовскую (CNB). Поскольку входные данные моделей MNB и CNB должны быть неотрицательными, мы заменяем метод стандартизации данных, показанный в фреймворке (рисунок 4), операцией нормализации данных, чтобы масштабировать входные объекты до интервала (0, 1) и соответствовать ограничению входных данных модели. Кроме того, некоторые критические параметры этих трех наивных байесовских моделей также оптимизируются перед сравнительным анализом производительности модели при обнаружении конфликтов.

Обнаружение конфликтов при эксплуатации поездов в этом исследовании является проблемой классификации. Оценка модели классификации в основном основана на матрице путаницы, показанной в таблице 2. Сначала мы выбираем кривую точности и рабочих характеристик приемника (ROC), чтобы оценить эффективность модели на всех выборках. Точность отражает долю правильных выборок в общей выборке. Кривая ROC принимает частоту истинных положительных результатов (TPR) за горизонтальную ось, а частоту ложных положительных результатов (FPR) за вертикальную ось, которая используется для нахождения оптимальной критической точки для наибольшей частоты истинных положительных результатов и наименьшей частоты ложных положительных результатов. Расчет точности, TPR и FPR показан в (7), (8) и (9). Площадь под кривой ROC (AUC), которая является значением вероятности, может указывать на производительность модели. Для положительной пробы и отрицательной пробы вероятность ранжирования положительной пробы перед отрицательной может быть рассчитана алгоритмом классификации в соответствии с прогнозируемыми результатами, и эта вероятность представляет собой значение AUC. Чем больше значение AUC, тем лучше работает алгоритм. Как точность, так и кривая ROC оценивают общий классификационный эффект модели для всех категорий. Поскольку целью этого исследования является обнаружение конфликтов, то есть отличие конфликтов от обычных выборок, точность и отзывчивость также выбраны в качестве оценочных показателей для оценки производительности модели в различных категориях. Расчеты точности и отзыва показаны в уравнении 10 и уравнении 11 соответственно.

В общем, точность и отзыв - это пара взаимоисключающих показателей. Модели с высокой точностью, как правило, имеют более низкий отзыв, но выдающиеся показатели классификации позволяют получить высокий балл по обоим показателям. Кроме того, поскольку перед моделированием была проведена операция понижающей выборки, данные, используемые для моделирования, являются сбалансированными данными. Следовательно, необходимо проверить, точно ли модель обнаруживает конфликтные выборки и неправильно ли модель обнаруживает нормальные выборки. Другими словами, для систематической оценки эффективности предлагаемой модели обнаружения TOC в каждой категории должны быть рассчитаны показатели как точности, так и отзыва.

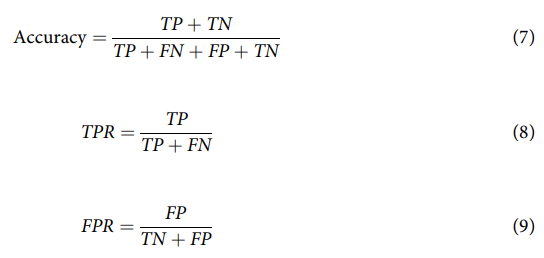
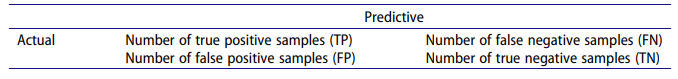
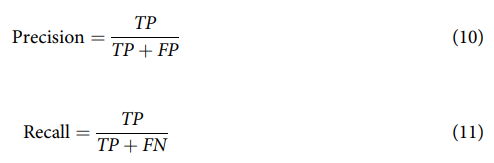


Таблица 2. Матрица путаницы.



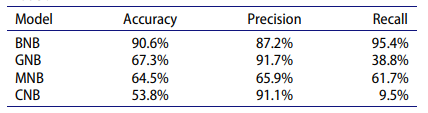


o Результаты работы модели NB

Во-первых, исследуется производительность каждой наивной байесовской модели на всех тестовых данных X-S HSR. Точность, прецизионность и отзыв показаны в таблице 3, а кривая ROC показана на рисунке 7. Результаты, приведенные в таблице 3, показывают, что предложенная модель BNB обладает наивысшей точностью прогнозирования по сравнению с другими байесовскими моделями на всех тестовых выборках. Кривые ROC на рис. 7 показывают, что кривая ROC модели BNB окружает кривые ROC других эталонных моделей. Значение AUC предлагаемой модели BNB составляет 0,948, что намного выше, чем у остальных трех наивных байесовских моделей (AUC GNB, MNB и CNB равны 0,848, 0,701 и 0,676 соответственно).

Кроме того, в последних двух столбцах таблицы 3 приведены значения точности и отзыва для каждой модели, что показывает, что, хотя значения точности моделей GNB и CNB немного выше, чем у модели BNB, значения отзыва моделей GNB и CNB составляют всего 38,8% и 9,5%, что намного ниже, чем у модели BNB. Вцеломthe BNB model BNB достигать 90,6%, 87,2%, и

Таблица 3. Точность и отзыв обнаружения конфликтов для каждой модели.



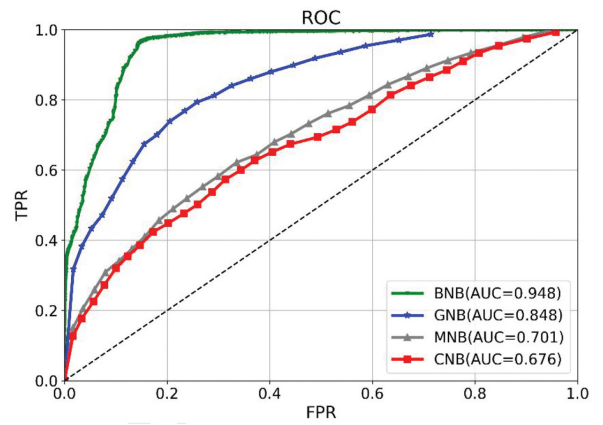
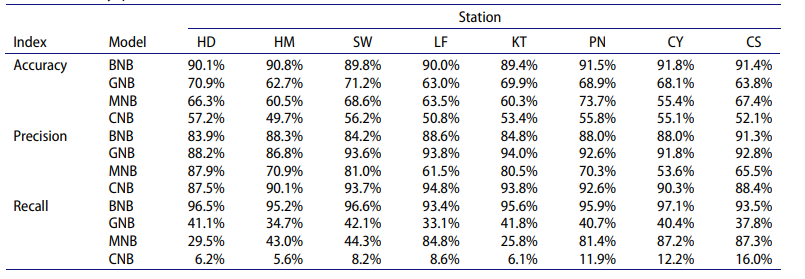


Рисунок 7. Кривые ROC и результаты AUC для каждой модели обнаружения конфликтов.

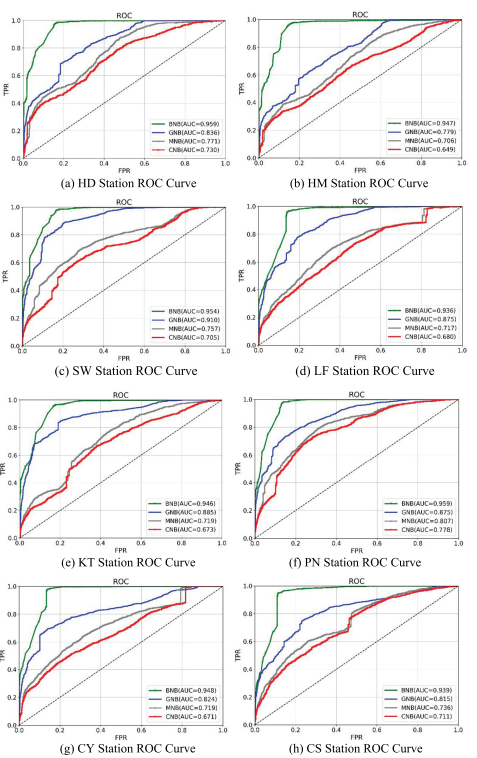
Таблица 4. Точность и отзыв обнаружения конфликтов на всех станциях каждой модели.



точность, прецизионность и отзываемость составляют 95,4%, но другие модели могут получить высокий балл только по одному конкретному показателю. Таким образом, учитывая все показатели, модель BNB значительно лучше других моделей в определении TOC.

Приведенный выше анализ показывает, что предлагаемая модель BNB превосходит другие наивные байесовские модели в обнаружении конфликтов при эксплуатации поездов. Чтобы систематически оценивать производительность модели, выбранные показатели (accuracy, ROC, прецизионность и отзыв) также используются для оценки производительности модели на каждой станции X-S HSR.

В таблице 4 приведены значения точности, отзыва и прецизионности каждой модели для обнаружения конфликтов в работе поездов на каждой станции. Модель BNB значительно лучше других моделей в отношении логичности, точности и показателей отзыва на каждой станции. Аналогично, значения точности моделей GNB и CNB немного выше, чем у BNB на некоторых станциях, но отзыв моделей GNB и CNB чрезвычайно низок. Следовательно, учитывая значения точности, прецизионности и отзыва, производительность модели BNB при обнаружении конфликтов значительно выше, чем у других моделей.



На рисунке 8 показана кривая ROC результатов обнаружения конфликтов каждой модели для работы поезда на каждой станции. Результаты показывают, что кривая ROC модели BNB окружает кривые ROC других наивных байесовских моделей на каждой станции. Все значения AUC модели BNB превышают 0,930 на каждой станции, в то время как максимальное значение AUC эталонных моделей составляет всего 0,910. Результаты показывают, что модель BNB неизменно превосходит другие широко используемые байесовские модели в обнаружении TOC.

o Результаты эффективности при распространении конфликтов/взаимодействиях

В разделе 5.1 мы показываем результаты работы предлагаемой модели обнаружения на ближайшей станции (т. е. Цель состоит в том, чтобы определить, есть ли конфликты между поездами на следующей станции). Однако управление поездами интерактивно благодаря взаимосвязанности инфраструктуры и сигнальных систем. Это означает, что конфликты зависят от соседних поездов и станций. Как только конфликт возникает между двумя соседними поездами на станции, он, скорее всего, распространится на будущие станции и следующие поезда. Поэтому в этом разделе мы исследуем эффективность предлагаемой структуры при распространении конфликта (на будущие станции и на следующие поезда). Возьмем в качестве примера рисунок 1. Если целью модели в разделе 5.1 является обнаружение конфликта между Train i и Train

i + 1 при Sn, модели в этом разделе предназначены для (1) обнаружения конфликтов между Train i и Train i + 1 при

Sn+1 (Задача 1) и (2) обнаруживать конфликты между Train формула и Train формула в Sn (Задача 2).

Чтобы продемонстрировать зависимости / взаимодействия между конфликтами, мы обучаем модели BNB с помощью

и без результатов обнаружения конфликтов, приведенных в разделе 5.1, в качестве входных данных для двух вышеупомянутых задач. Другими словами, для задач 1 и 2 мы обучаем модель BNB с результатом обнаружения конфликта (с именем BNB- with) в разделе 5.1 в качестве входных данных и модель BNB без результата обнаружения конфликта (с именем BNB-without) в разделе 5.1 в качестве входных данных соответственно. Следовательно, зависимости / взаимодействия между конфликтами могут быть количественно определены путем сравнения моделей BNB-with и BNB-without. Результаты сравнения BNB-с моделью и BNB-без модели в отношении точности, прецизионности и отзыва для задач 1 и 2 показаны в таблицах 5 и 6 соответственно. Сравнительные результаты BNB-с моделью и BNB-без модели с точки зрения ROC и AUC для задач 1 и 2 показаны на рисунке 9 и и рисунке 10 соответственно.

Результаты, приведенные в таблицах 5 и 6, показывают, что модель BNB-with неизменно превосходит модель BNB-without. Кроме того, AUCS моделей BNB-with также больше, чем у моделей BNB-without, показанных на рис. 9 и 10. Эти результаты означают, что использование обнаруженных результатов в разделе 5.1 в качестве входных данных может значительно улучшить производительность модели NB, также демонстрируя взаимодействия и зависимости конфликтов движения поездов между соседними станциями и поездами. Чтобы количественно оценить зависимости / взаимодействия конфликтов между станциями и поездами, мы рассчитали улучшение модели BNB-with по сравнению с моделью BNB without. Таблица 5 и рисунок 9 показывают, что модель BNB-with улучшена в среднем на, 3.38%, 4.39%, 3.37%, и 3,44% по сравнению с моделью BNB-without с точки зрения точности, прецизионности, отзыва и показателей AUC. Таблица 6 и рисунок 10 показывают, что модель BNB-with улучшена в среднем на, 6.55%, 4.23%, 15.96%, и 2,08% по сравнению с моделью BNB-without с точки зрения точности, прецизионности, отзыва и показателей AUC. Эти цифры количественно показывают зависимости между конфликтами и необходимостью учета взаимодействия между станциями и поездами.

о выводах

На основе данных о работе поездов X-S HSR мы проводим исследование по выявлению конфликтов в работе поездов. Предложена модель обнаружения конфликтов, основанная на наивном Байесе Бернулли. На основе стандартных показателей оценки моделей классификации, таких как точность, кривая ROC, прецизионность и отзыв, производительность предлагаемой модели обнаружения конфликтов сравнивается с другими широко используемыми наивными байесовскими моделями. Выводы таковы:

o Точность обнаружения конфликтов наивной байесовской модели Бернулли значительно выше, чем у других наивных байесовских моделей, для всех наборов данных и на каждой станции.

o Предлагаемая модель может получить высокий балл по точности и отзыву, но другие эталонные модели могут иметь удовлетворительный балл по точности.

o Кривая ROC предлагаемой наивной модели Байеса Бернулли окружает кривые ROC других моделей. Площадь под кривой ROC наивной модели Байеса Бернулли больше 0,930, в то время как площадь под кривой ROC других моделей меньше 0,910.

Это исследование является предварительным исследованием обнаружения конфликтов при эксплуатации поездов с точки зрения больших данных. Будущую работу можно было бы продолжить, рассмотрев влияние распространения задержек и взаимодействий. Влияние задержек поездов на будущее состояние движения поездов и взаимодействие в процессе движения поездов и между соседними поездами будет учитываться в методе обнаружения для улучшения характеристик модели