УДК 004.9

Разработка алгоритмов прогнозирования дорожной ситуации с применением Big Data

Леонтьев С. С. (Санкт-Петербургский государственный университет),
Научный руководитель – доктор технических наук, профессор кафедры информационных систем в экономике, Стоянова О. В.
(Санкт-Петербургский государственный университет)

Аннотация

В данной работе представлена разработанная система, позволяющая пользователю на основании введенных параметров получить маршрут с указанием аварийно-опасных участков и коэффициентов опасности, рассчитанных с помощью созданных моделей для текущих условий. Для реализации алгоритмов анализа данных и прогнозирования опасности участка в работе использованы метод кластеризации DBSCAN, метод классификации случайный лес и определение значимости переменных методом Mean Decrease Accuracy. Для разработки использованы языки программирования Python и JavaScript, а также языки разметки HTML и CSS.

Введение

Обеспечение безопасности дорожного движения является ключевой целью любого государства. Ежегодно в мире на дорогах гибнет в среднем 1,35 миллионов человек. На Россию в 2020 году пришлось более 16,1 тысяч смертей и 183 тысяч ранений в 145 тысячах дорожно-транспортных происшествиях с пострадавшими (каждое девятое – со смертельным исходом). Несмотря на то, что ежегодно смертность на дорогах страны снижается, Россия находится в топ-20 стран мира по количеству смертей на дорогах. Любое ДТП с летальным исходом – это финансовые потери со стороны родственников погибшего, государства. Каждый год ДТП обходятся государству в среднем в 3% ВВП. Но это ничто по сравнению с человеческой жизнью. Существует множество способов повышения безопасности дорожного движения. Один из них – внедрение интеллектуальных информационных систем.

Основная часть

В работе использовались данные о дорожно-транспортных происшествиях, произошедших в г. Санкт-Петербург с 2015 по 2021 год, размещенные на сайте проекта «Карта ДТП» (URL: https://dtp-stat.ru). Первичный анализ данных о ДТП в городе включает в себя исключение аномальных и некачественных данных. Для начала из данных были удалены все записи, содержащие пустые значения, далее все записи, географические координаты которых не попадают в следующий квадрат: более 59.62 и менее 60.22 градусов северной широты и более 29.52 и менее 30.85 градусов восточной долготы. После первичной обработки данных осталось 36191 запись о ДТП в Санкт-Петербурге в период с 2015 по 2021 годы.

В соответствии с дополнением, внесенным в Федеральный закон "О безопасности дорожного движения", аварийно-опасный участок дороги - это «участок дороги, улицы, не превышающий 1000 метров вне населенного пункта или 200 метров в населенном пункте, либо пересечение дорог, улиц, где в течение отчетного года произошло три и более ДТП одного вида или пять и более ДТП независимо от их вида, в результате которых погибли или были ранены люди». Для определения таких участков был выбран метод кластеризации DBSCAN. Метод позволяет с большой скоростью определять кластеры произвольной формы. Он оперирует параметрами расстояния между двумя элементами и минимальным количеством элементов, образующих кластер. Стоит учесть, что у метода DBSCAN нет параметра, ограничивающего размер кластера, но на основании указанного выше Федерального закона было принято решение исключить из каждого получившегося кластера элементы, расстояние до которых от центра кластера более 200 метров.

Большая часть аварий в населенных пунктах происходят на перекрестках. Анализ размеров перекрестков Санкт-Петербурга показал, что в городе есть перекрестки в диаметре

50 и более метров. Так как расчет расстояния между авариями происходит на основании их долготы и широты, параметр расстояния необходимо брать в градусах. Для этого 50 метров необходимо разделить на радиус планеты в метрах, то есть на 6371210 метров. Параметр количества элементов, образующих кластер был установлен равным 1, чтобы каждая точка данных была назначена либо кластеру, либо образовала свой собственный кластер из одного элемента. Учитывая тот факт, что данные за 2015 и 2021 годы представлены не за весь период, будем считать, что продолжительность периода – 6 лет, поэтому в модели учитываются только те кластеры, в которых 30 или более элементов.

Для определения, опасен ли выбранный участок в текущих условиях, необходимо провести бинарную классификацию. В качестве переменных, по которым строится классификация, были выбраны: погодные и дорожные условия, освещение, время, месяц, день недели, и район. В исходных данных все записи являются элементами класса «авария». Для верного обучения классификатора необходимо иметь элементы другого класса: «не авария». Было принято решение для каждого элемента класса «авария» создать по одному элементу с отличной датой и временем (выбранными случайно в промежутке от наименьшей до наибольшей даты элементов класса «авария»). Выборка была разбита на тестовую и тренировочную в соотношении 30 к 70. Далее была проведена классификация. В качестве метрики, отражающей качество классификации выбрана F-мера. Начальным параметром «количества деревьев в лесу» был выбран 100. При указанных параметрах качество модели оказалось на уровне 73,5%. Дальнейшее изменение количества деревьев не привело к значительным изменениям качества классификации, поэтому было принято решение остановиться на 100 деревьях. Разработанная модель классификации используется для прогнозирования опасности участков в зависимости от введенных пользователем параметров маршрута.

Далее был проведен расчет значимости каждого параметра. На 44% аварийность участка определяет время поездки, на 14% - район города, остальные – по убыванию: состояние дорожного покрытия - 11,6%; месяц - 9,7%; погодные условия - 9,3%; день недели - 7,3%; освещение - 3,3%. Далее показатели значимости используются для расчета «коэффициента опасности участка».

Данные о кластерах, а также параметры классификации передаются в систему. Система представляет собой web-приложение, содержащее карту и окна ввода начальной и конечной точки маршрута, а также даты и времени поездки. После ввода, пользователю предлагается оптимальный маршрут, в котором отмечены аварийно-опасные места в зависимости от погодных и иных условий. После того, как пользователь вводит данные о маршруте, происходит обращение к форме, которая получает данные о координатах начальной и конечной точки, а также данные с сайта погоды на основании полученных от пользователя начальной и конечной точки и даты и времени. После этого на клиентскую часть приходит ответ, содержащий координаты начальной и конечной точки, погодные и дорожные условия в соответствующее время. На клиентской части происходит просчет маршрута, а затем - вывод аварийно-опасных участков.

Наведя курсор на маркер, пользователь может получить дополнительную информацию о каждом участке, включающую время суток и погодные условия, при которых на данном участке чаще всего происходят ДТП, наиболее частый тип аварий, количество ДТП, а также пострадавших и погибших на участке людей за период с 2015 по 2021 год. Также в дополнительной информации присутствует коэффициент опасности участка, рассчитываемый с помощью предложенной модели, учитывающей влияние факторов. После изучения маршрута пользователь может изменить параметры маршрута и рассчитать новый, либо очистить карту.

Заключение

В ходе работы автором была разработана система, позволяющая пользователю построить маршрут с отмеченными на нем опасными участками. Стоит отметить, что система

не является полной и может быть доработана (например, повышено качество прогнозов погодных и дорожных условий, или, возможно расширение данных за счет смежных баз), что позволит повысить качество моделей и точно определения опасности участков маршрута.