УДК 519.688

Создание статистической модели определения времени прохождения перегона поездами по реальным данным

Е. Ю. Бобарико[1]

[1]Московский физико-технический институт (государственный университет)

В настоящее время во всем мире и в России в частности актуальна проблема оптимизации железнодорожных перевозок. Представляемая работа является частью работ по организации оптимального пропуска поездов по железнодорожной сети (ЖС). Цель работы – создание модели определения времени прохождения перегона поездом по статистическим данным. Алгоритмы оптимального пропуска поездов оперируют временами прохода поездов по перегонам. В практике железнодорожных перевозок эти времена определяются из тяговых расчётов, которым посвящено большое количество литературы (в том числе учебной).

Проанализировав реальные статистические данные по временам прохода поездов разного веса по перегону 25793 (Рабак) - 25800 (Куеда) длиной 22.9 км с локомотивом серии 240, мы пришли к заключению, что реальные времена хода поездов значительно отличаются от теоретических, которые получаются расчётным путём. Из графика времён прохождения практически одинаковых поездов по перегону следует, что время прохождения поезда имеет вероятностный характер.

Так, например, статистические данные дают среднее время прохода по рассматриваемому перегону $T_{ср}$ = 25,6 мин, а в основной диспетчерской информационной системе графика исполненного движения ГИД-Урал ВНИИЖТ для этого перегона заложено время $T_{гид}$ = 23,0 мин. Ошибка составляет более 11%.

Не имея достоверных времён хода поездов (различной массы, с различными начальными скоростями выхода на перегон, с различными ограничениями максимальной скорости на перегоне, с различными локомотивами и локомотивными бригадами) сделать программу оптимального управления движения поездами невозможно.

Таким образом, необходимо, или привлечь статистический аппарат обработки данных, или внести соответствующие поправки в тяговые расчёты на основе статистических данных.

Проанализируем фактические данные.

1. Время прохода по перегону для рассматриваемого локомотива практически не зависит от массы состава.

2. Наши данные подчиняются нормальному распределению.

Если по фактическим данным построить гистограмму и наложить на неё график функции плотности распределения вероятностей нормального распределения со средним и среднеквадратичным отклонением, посчитанными по фактическим данным, то наблюдается хорошее совпадение.

Согласно детерминированным тяговым расчётам для поездов рассматриваемой массы, ПК «Искра» выдаёт результат, который почти на 14% отличается от среднего значения фактических данных. Скорее всего, в основной диспетчерской системе ОАО «РЖД» ГИД-Урал данные поправляли, исходя из опыта эксплуатации.

В любом случае, для практических целей результатами тяговых расчётов пользоваться нужно аккуратно, они не совпадают с практикой.

Следовательно, возникает следующая задача. Получить достоверную прогнозную модель для определения времён хода различных поездов по перегонам сети ОАО «РЖД».

В распоряжении имеется выборка данных по временам движения поездов по различным перегонам Горьковской железной дороги.

Из детерминированных тяговых расчетов следует, что время движения поезда по перегону зависят от следующих параметров:

- масса состава;
- тип локомотива (несколько параметров);
- параметры конкретного перегона: длина, текущие ограничение максимальной скорости на различных участках перегона, профиль путей (нужно параметризовать);

• номер локомотивной бригады (считаем, что все бригады пронумерованы и каждая из бригад имеет некоторый стиль вождения).

Из-за актуальности проблемы ученые предлагали достаточно много вариантов решения данного вопроса.

1.Теоретический метод.

Теоретический метод составляют тяговые расчёты, состоящие из решения дифференциальных уравнений, в основе которых лежат законы механики.

В современной постановке тяговые расчёты расширяются до оптимального управления, например, с точки зрения экономии электроэнергии.

Оптимальное управление с экономией энергии состоит в определении режимов управления безостановочным движением поезда по участку между начальным и конечным пунктами за заданное время, обеспечивающих минимальный расход энергии на тягу с учётом плана и профиля пути, длины состава, типа и загруженности вагонов, тяговых и тормозных характеристик локомотива, ограничений скорости движения. Началом решения является поиск исходного допустимого режима управления движением, наиболее близко удовлетворяющего всем условиям и ограничениям задачи. При этом используется алгоритм, основанный на регулировании времени хода, а затем—собственно, оптимизационный алгоритм последовательного улучшения исходного приближения по расходу энергии. То есть, на каждом шаге итерационного процесса строится удовлетворяющая всем условиям и ограничениям задачи траектория движения, при реализации которой расход энергии меньше, чем при реализации траектории, рассчитанной на предыдущем шаге.

2.Статистический метод.

Мы имеем статистический массив времён по данным прохождения каждого перегона. Однако статистика достаточная бедная, а нам нужен результат при любой комбинации входных параметров.

Но при изменении любого из параметров время может поменяться. Статистический массив не содержит всех возможных комбинаций по перечисленным параметрам. В отдельных случаях, возможно, при прохождении перегона может случиться нечто, что сильно задержит поезд, данные факты должны выявляться и не влиять на общую статистику. Необходимо построить модель, которая при задании параметров, влияющих на время прохождения перегона, выдаст прогнозируемую скорость его прохождения, а также сможет выдавать прогноз при неизвестных значениях того или иного параметра.

Размерности модели:

- Перегонов порядка 10000
- Масса поезда – непрерывная величина
- Серии локомотивов порядка 100
- Бригад порядка 10000
- Экземпляров локомотива порядка 10000
- Время прохождения – непрерывная величина
- Абсолютное время входа и выхода с перегона (дата, время суток) – непрерывная величина
- Длина поезда
- Тип поезда

Мы должны определить, каким инструментом наиболее эффективно пользоваться, чтобы построить модель.

Для исследования мы решили обратиться к MATLAB R2014b (сокращение от англ. *«Matrix Laboratory»*)— пакет прикладных программ для решения задач технических вычислений и одноимённый язык программирования, используемый в этом пакете. Наш выбор пал именно на этот инструмент по нескольким главным причинам:

1. MATLAB позволяет сгенерить C++ текст программы (то есть обученной модели).
2. Есть функции, которые не подлежат кодогенерации (производитель не обеспечил эту возможность). Тогда MATLAB позволяет сделать отдельную библиотеку dll, exe файл, Java приложение, WEB приложение, .NET библиотеку для C# и т.д.

Без этих возможностей MATLAB для нас не представлял бы особого интереса.

С помощью него мы обратимся к машинному обучению, а конкретно к одному из его разделов - регрессионному анализу.

Регрессионный анализ - метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений зависимой переменной (переменной отклика) и независимой переменной (объясняющей переменной). Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные. Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента. Предполагается, что зависимая переменная есть сумма значений некоторой модели и случайной величины. Относительно характера распределения этой величины делаются предположения, называемые гипотезой порождения данных. Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом остатков. При этом предполагается, что независимая переменная не содержит ошибок.

Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

Цель использования регрессионного анализа:

1. Определение степени детерминированности вариации критериальной (зависимой) переменной предикторами (независимыми переменными)
2. Предсказание значения зависимой переменной с помощью независимой(-ых)
3. Определение вклада отдельных независимых переменных в вариацию зависимой

Регрессионный анализ нельзя применять для определения наличия связи между переменными, поскольку наличие такой связи и есть предпосылка для использования анализа.

Для построения регрессионного анализа мы использовали одиннадцать методов, которые включают линейные подходы, построение деревьев решений и нейронные сети.

Так как у нас имеется общая статистика по 6 месяцам, а каждый день случаются разные непредвиденные ситуации, которые могут привести к задержке поездов, то получаются «грязные данные». Поэтому, чтобы получить наиболее правильную модель и определить, какие же переменные дают основной вклад во время движения поезда, нам необходимо провести «очистку» данных.

Для этого было сделано следующее: исключили строки, для которых время движения поезда по перегону отличается в два, и более раз, от среднестатистического времени движения (скорее всего, была остановка на перегоне, т.е. не регулярное движение). Для этого была посчитана для каждого перегона медиана (median) времени прохождения, затем умножена она на 2. Потом были удалены все времена для данного перегона, превышающие данное время.

В итоге новая статистика содержала в себе 3613754 строк. Было выброшено 2,3% строк, для которых время превышало среднестатистическое в 2 и более раз.

В качестве обучающейся выборки мы взяли 40% данных.

Для учета всех ситуаций при построении модели было проведено несколько серий экспериментов с различными критериями отбора входных данных. Построена регрессия:

1. с учетом ограничений скоростей на перегонах;
2. для длинных перегонов;
3. для коротких перегонов;
4. для определенного перегона (25793 (Рабак) - 25800 (Куеда)).

В первом эксперименте мы ввели для каждого ограничения скорости 3 параметра (три дополнительных переменных к искомой функции). Первый параметр - расстояние от начала перегона по ходу движения поезда, когда начинает действовать ограничение, второй параметр – величина ограничения скорости, третий параметр – расстояние, когда ограничение заканчивает действовать. Если ограничения отсутствует, то первый или третий параметр соответственно принимает значение 0, второй (максимальная скорость поезда) для грузового поезда равна 80 км/ч, а пассажирский – 120 км/ч. Данные ограничения позволяют более реальную ситуацию отразить в модели.

Следующими исследованиями нами было установлено, что независимо от длины перегона наименьшее среднеквадратичное отклонение между выходными данными и теоретическими наблюдалось у метода BAGTREE. Именно его необходимо использовать для создания модели определения времени прохождения поездом перегона.

Стоит отметить, что в последнем исследовании мы показали, что ошибка в предсказанном времени отличается от фактического всего на 0,2%. Это очень хороший показатель, особенно, если сравнивать с результатами модели, которую используют на данный момент РЖД, где ошибка составляет 14%.

Результаты данной работы являются необходимой частью более крупного исследования – создания модели определения времени прохождения перегона поездом по реальным данным. Нам удалось определить наиболее эффективный алгоритм для построения модели. Полученный результат дает меньше ошибку, чем используемая в настоящее время на практике модель.

Литература

1. *Баранов Л.А., Ерофеев Е.В., Мелёшин И.С., Чинь Л.М.* Оптимизация управления движением поездов: учебное пособие / под ред. д.т.н., проф. Л.А. Баранова. -М.: МИИТ, 2011. – 164 с.
2. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных/ пер. с англ. А.А.Слинкина.-М.: ДМК Пресс, 2015. - 400с.:ил.