

Регрессионная модель прогнозирования скорости движения транспорта

И. Чучуева
МГТУ им Н.Э.Баумана
chuchueva@yandex.ru

АННОТАЦИЯ

В настоящей работе представлено описание алгоритма решения задачи прогнозирования загруженности автомобильных дорог по историческим данным, предоставленным компанией «Яндекс». В статье представлен общий подход к решению задачи прогнозирования, а также приведены основные математические зависимости для вычисления прогнозных значений. В результате реализации предложенной модели были получены результаты, оценка точности которых составила 65.118.

Ключевые слова

прогнозирование, регрессионная модель, пробки

1. ВВЕДЕНИЕ

С 1 марта по 16 мая 2010 года в рамках проекта «Интернет-Математика» компания «Яндекс» проводила конкурс. В качестве конкурсной была предложена задача прогнозирования загруженности автомобильных дорог внутри одного дня на основе исторических данных. С условиями конкурса, формами регистрации, исходными данными и способами оценки результатов можно ознакомиться по ссылке [1].

В рамках данного конкурса автором был реализован ряд моделей прогнозирования скорости движения транспорта (СДТ). В настоящей статье приведено описание регрессионной модели прогнозирования СДТ, показавшей наиболее точный результат согласно *публичной оценке* (ссылка [1]).

2. ИСХОДНЫЕ ДАННЫЕ

По условиям конкурса исторические данные о СДТ содержатся в файле jams.txt. Наблюдения охватывают 31 день: для первых 30 дней в файле содержится информация о скорости движения потока автотранспорта с 16:00 до 22:00, для последнего дня — с 16:00 до 18:00. Здесь и далее мы будем говорить о времени, имея в виду формат записи «ЧЧ:ММ». Данные файла jams.txt представлены в таблице 1.

Таблица 1. Исходные данные и обозначения переменных

п/п	Идентификатор дороги, Y	Номер дня, d	Время, t		СДТ, км/ч, $Z^Y(t_i^d)$
			Часы	Минуты	
1	317744	11	16	26	62
2	317744	11	16	30	62
3	317744	11	16	34	62
...	317744	11	16	40	63

Во второй колонке содержится идентификатор дороги; в третьей – номер дня, четвертая и пятая колонки представляют собой время; шестая колонка содержит информацию о СДТ в км/ч.

Исходный файл jams.txt содержит около 30 млн. строк. Данные по каждой дороге даны за дни в диапазоне $d \in [11;41]$. Время t лежит в диапазоне с 16:00 до 22:00 для дней с 11 по 40, для последнего дня время t лежит в диапазоне с 16:00 до 18:00. Согласно файлу task.txt необходимо спрогнозировать около 700 тыс. значений СДТ для 29335 дорог за период с 18:00 по 22:00 для последнего дня (т.е. $d = 41$).

3. АЛГОРИТМ РЕШЕНИЯ ЗАДАЧИ

Алгоритм решения задачи прогнозирования представлен на рис. 1, ниже приведено описание действий на каждом шаге.

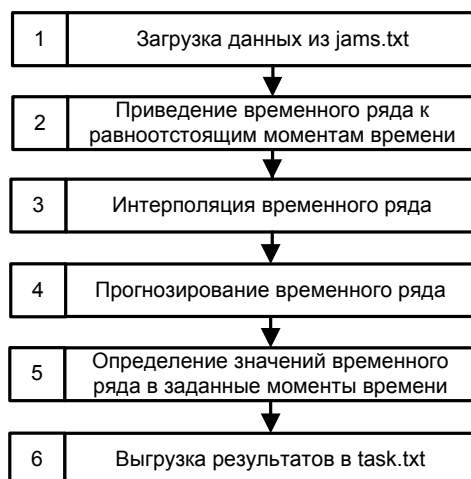


Рис 1. Алгоритм решения задачи прогнозирования СДТ

3.1 Загрузка данных

На первом шаге исходные данные выгружались из файла jams.txt с целью их последующей обработки. Исторические данные по каждой дороге представляют собой временной ряд $Z^Y(t_i^d) = Z^Y(t_1^1), Z^Y(t_2^1), Z^Y(t_3^1), \dots, Z^Y(t_{Q1}^1), Z^Y(t_1^2), Z^Y(t_2^2), \dots$, содержащий значения СДТ в неравноотстоящие моменты времени t_i^d , где верхний индекс отражает номер дня, а нижний индекс – время внутри этого дня. Необходимо отметить, что для различных дорог длина временного ряда, содержащего исторические значения, различна и колеблется от нескольких значений до нескольких тысяч значений.

3.2 Приведение временного ряда к равноотстоящим моментам времени

Неравноотстоящие во времени значения СДТ временных рядов $Z^Y(t_i^d)$ приводились к равноотстоящим моментам времени $Z^Y(t_p^d)$, где p назовем периодом. Длина периода рассматриваемой модели составляет 30 минут, так например, период $p = 1$ соответствует интервалу времени $t \in [16:00; 16:30)$.

Таблица 2. Соответствие номера периода и времени

Период, р	1	2	3	4	5	6	7	8	9	10	11	12
Время, t	16:00	16:30	17:00	17:30	18:00	18:30	19:00	19:30	20:00	20:30	21:00	21:30

Значение СДТ за период p определяются как

$$Z^Y(t_p^d) = \sum_{i=1}^M Z^Y(t_i^d), \quad (1)$$

где M – количество значений СДТ, попавших внутрь периода t_p^d , т.е. определялось среднее значение СДТ внутри каждого периода каждого дня.

3.3 Интерполяция временного ряда

В результаты вычислений, описанных на предыдущем шаге, для каждой дороги был получен временно ряд $Z^Y(t_p^d)$ с равноотстоящими моментами времени. Внутри полученного ряда содержались периоды с пропущенными значениями, которые необходимо интерполировать для последующего решения задачи прогнозирования. Для интерполяции значений временных рядов $Z^Y(t_p^d)$ прежде определялись среднее значение СДТ за период

$$Z^Y(t_p) = \sum_{d=1}^D Z^Y(t_i^d). \quad (2)$$

Здесь D – есть количество дней, содержащих исторические данные по временному ряду за указанный период p . Интерполяция выполнялась замещением отсутствующих значений $Z^Y(t_p^d)$ средними значениями за данный период. Пояснения к интерполяции представлены в таблице 3.

Таблица 3. Пояснения к интерполяции временных рядов $Z^Y(t_p^d)$

Период, p	1	2	3	...	8	9	10	11	12
$Z^Y(t_p^d)$	50	30	NULL	...	60	50	NULL	55	50
			↑ $Z^Y(t_3)$				↑ $Z^Y(t_{10})$		

Таким образом, пропущенные значения временного ряда $Z^Y(t_p^d)$ замещались средним значением СДТ $Z^Y(t_p)$ для данного периода. Полученный временной ряд обозначим $\bar{Z}^Y(t_p^d)$.

3.4 Прогнозирование временного ряда

Каждый временной ряд $\bar{Z}^Y(t_p^d)$ прогнозировался внутри периодов $p \in [5;12]$ для $d = 41$ на основании следующего прогнозного уравнения

$$\bar{Z}^Y(t_p^d) = \alpha_2 \cdot Z^Y(t_p) + \alpha_1 \cdot \bar{Z}^Y(t_{p-1}^{41}) + \alpha_0. \quad (3)$$

Здесь α_2 , α_1 и α_0 численные константы, $Z^Y(t_p)$ – среднее значение СДТ в этом периоде, определенное по формуле (2), $\bar{Z}^Y(t_{p-1}^{41})$ – СДТ в предыдущем периоде. Таким образом, будущее значение СДТ зависит от среднего значения СДТ в данный период по итогам предыдущих дней, а также от СДТ в предыдущем периоде. Значения констант α_2 , α_1 и α_0 определялись при помощи метода наименьших квадратов, описанного в литературе [2].

Алгоритм получения будущих значений СДТ имеет вид, представленный на рис. 2.

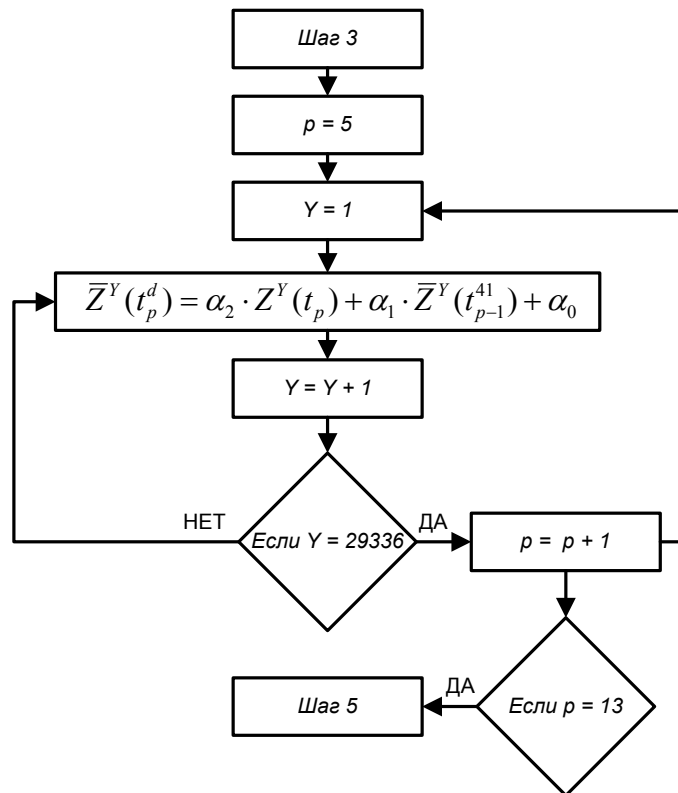


Рис 2. Алгоритм вычисления прогнозных значений

3.5 Определение значений временного ряда в заданные моменты времени

Полученные на предыдущем шаге прогнозные значения временного ряда $\bar{Z}^Y(t_p^{41})$ внутри периодов $p \in [5;12]$ были преобразованы во временной ряд, с требуемыми в задании неравноотстоящими моментами времени

$$Z^Y(t_i^{41}) = \bar{Z}^Y(t_p^{41}), \forall t_i \in t_p. \quad (4)$$

Таким образом, значения СДТ в требуемые моменты времени t_i приравнивались к значению СДТ внутри соответствующего периода t_p .

3.6 Выгрузка результатов

Множество временных рядов, содержащих будущие значения СДТ для заданных дорог, выгружались в файл task.txt согласно заданному формату данных.

4. РЕЗУЛЬТАТЫ

При использовании данной модели прогнозирования были получены результаты, публичная оценка которых составила 65.118. Полные результаты прогнозирования СДТ будут объявлены 31 мая 2010 года.

ЛИТЕРАТУРА

- [1] «Интернет-математика - 2010» [Электронный ресурс] // «Яндекс». DOI= <http://imat2010.yandex.ru/>.
- [2] Draper N. R., Smith H. Applied regression analysis. – New York: Wiley, In press, 1981. – P. 693.