Прогнозирование объемов грузоперевозок и оценка качества прогноза временных рядов

Газизуллина Римма

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва,
2015 г.

Цели исследования

Повышение качества прогнозирования объемов железнодорожных грузоперевозок

Используемые методы

Алгоритм, основанный на квантильной регрессии, модифицированный сверткой гистограммы

Что предлагается

Проводить прогноз путем учета топологии

Литература

Базовые публикации

- Вальков А.С.,Кожанов Е.М.,Медведникова М.М., Хусаинов Ф.И., Непараметрическое прогнозирование загруженности системы железнодорожных узлов по историческим данным // Машинное обучение и анализ данных, 2012. Т. 1, №4. С. 448-465.

Квантильная регрессия

- Kong L., Mizera I. Quantile tomography: using quantiles with multivariate data // Stat. Sinica, 2012. № 22. P. 1589-1610

Использование гистограмм

- Chen L., Dobra A. Histograms as statistical estimators for aggregate queries // Inf. Syst., Elsevier Science Ltd., 2013. №38. P. 213-230

Постановка задачи

Дано:

- $\mathbf{x} = \{x_i\}_{i=1}^{T}$ — временной ряд, который предполагается стационарным
- $\{y_1, y_2..y_k\}$ — реальные значения ряда x в некоторых точках

Вводится функция потерь $L(\hat{y}, y) = |\hat{y} - y|$.
Задача нахождения прогнозируемых значений $\hat{y}$:

$$\hat{y} = \underset{z \in \mathbb{N}}{\operatorname{argmin}} \sum_{k=1}^{K} L(z, y_k).$$

Пусть построен прогноз значений ряда $\{\hat{y_1}, \hat{y_2}..\hat{y_k}\}$ в некоторых точках, при этом реальные значения ряда x в этих же точках равны $\{y_1, y_2..y_k\}$.
Задача минимизации средней ошибки прогнозирования:

$$\text{MeanError} = \frac{1}{k} \sum_{i=1}^{k} L(\hat{y_i}, y_i) \to \min_{\hat{y_1}...\hat{y_k}}$$

Повышение качества прогноза путем учета топологии

Способы получения прогноза отправления вагонов:

- прогноз количества вагонов, проходящих через заданный узел,
- прогноз количества вагонов, проходящих с одного узла на другой.

Предполагается, что второй вариант прогноза более точен и более востребован, так как уточняет расписание сложившихся грузоперевозок.

Формат данных

Описание данных для блока вагонов содержит:

- код груза (нефть и нефтепродукты, сахар, продукты перемола и т.д.),
- род вагонов (полувагоны, крытые вагоны, цистерны, платформы, прочие).

Дата погрузки	Станция отправления	Станция назначения	Кол-во вагонов	Код груза	Род вагона	Суммарный вес груза, т.
2007-01-01	020108	932902	1	1	216	56
2007-01-01	032105	840109	1	19	040	63
2007-01-01	035508	843408	2	3	070	120

Алгоритм

1. По исходным данным для каждого типа груза G и каждой пары районов (L, A) строим временной ряд — объем перевозок груза G из района L в район A по дням
2. Строим временной ряд для суммарного количества каждого груза из каждого района
3. Применяем базовый алгоритм к временным рядам
4. Сравнение результатов

Базовый алгоритм

Для функции потерь $L(\hat{y}, y)$:

$$L(\hat{y}, y) = |\hat{y} - y|.$$

По временному ряду x построим гистограмму $\mathcal{H}$

$$\mathcal{H} = \{(y_k, g_k)\}_{k=1}^{K}, \tag{1}$$

K — число интервалов $[y_k^{\min}, y_k^{\max}]$ со средним значением y_k, на которые разбита ось значений ряда x,
g_k — высота столбца гистограммы на интервале y_k, равная взвешенной сумме количества точек ряда, попавших в этот интервал.

Тогда прогнозируемое значение ряда x_{T+h} находится как

$$\hat{y} = \underset{z \in \{y_1, \ldots, y_K\}}{\arg\min} \sum_{k=1}^{K} g_k L(z, y_k).$$

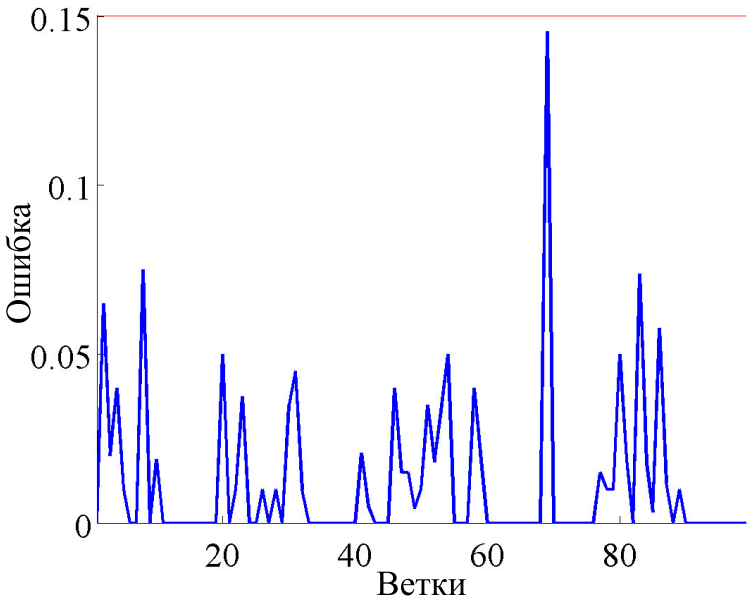
Цели вычислительного эксперимента

1. Применить алгоритм к данным после разбиения по парам районов для разных веток-грузов
2. Сравнить результаты с результатами для базового алгоритма

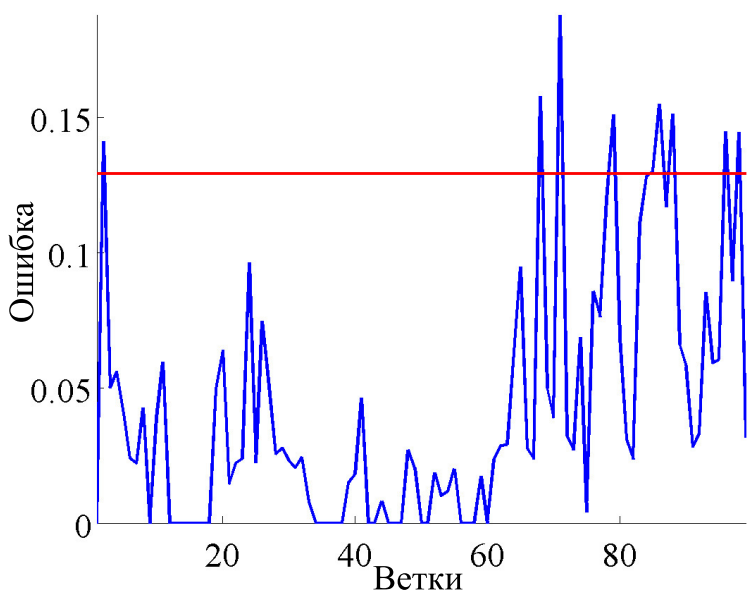
Для этого было проведено 2 эксперимента:

- Исследование детальных свойств временных рядов и прогнозов для нефти для отправленных вагонов из 83 района в другие
- Выполнение эксперимента для всех грузов для вагонов, отправленных во все районы, и сравнить показатели с результатами для базового эксперимента

Результаты вычислительного эксперимента



(a) Груз 2



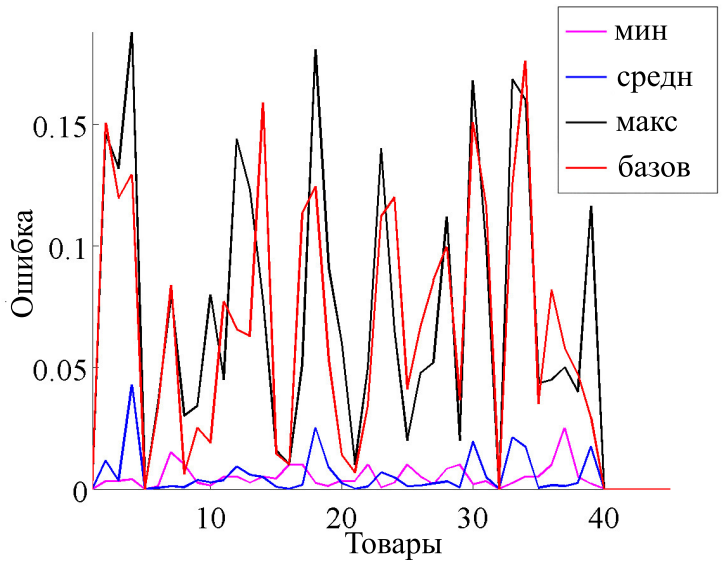
(b) Груз 4

Сравнение средней ошибки при прогнозе с узла на узел (синим) со средней ошибкой при прогнозе суммарного отправления вагонов с одного узла района на все узлы (красным).

Результаты вычислительного эксперимента

- Прогноз на отдельную ветку в большинстве случаев точнее
- Максимальная средняя ошибка для ветки близка к ошибке суммарного отправленного количества грузов
- Средняя ошибка по всем веткам значительно меньше, чем ошибка для суммы

Для каждого типа груза отложены: минимальная, средняя и максимальная ошибки при прогнозе с по парам веток и ошибка базового алгоритма, при котором прогноз выполняется для суммарного отправления на все ветки.



Заключение

Поставлен вычислительный эксперимент, который подтвердил предположение о том, что прогноз с учетом топологии точнее. Приведенные результаты подтверждают выдвинутое предположение.

Публикации по теме

Газизуллина Р.К., Стенина М.М., Стрижов В.В. Прогнозирование объемов железнодорожных грузоперевозок по парам веток // Системы и средства информатики, 2015. Т. 25, №1. С. 142-154.