Предсказание метеорологических величин с помощью гибридного метода обработки временных рядов

Е.А. Черных
Московский государственный университет имени М.В.Ломоносова
Москва, Россия
chernykh.ea18@physics.msu.ru

Н.Е. Шапкина
Московский государственный университет имени М.В.Ломоносова
ИТПЭ РАН
Москва, Россия
neshapkina@mail.ru

П.В. Голубцов
Московский государственный университет имени М.В.Ломоносова
Москва, Россия
golubtsov@physics.msu.ru

***Аннотация* — В данной работе представлен гибридный метод анализа временных рядов, основанный на авторегрессионной интегрированной модели скользящего среднего (ARIMA) и модели линейной регрессии, адаптированной к эффективной обработке больших данных, накапливаемых в режиме реального времени. Его практическое применение продемонстрировано на примере обработки реальных данных для предсказания временного ряда метеорологических величин.**

Ключевые слова— временные ряды, большие данные, ARIMA, линейная регрессия

1. Введение

В силу большого объёма данных и их постоянного пополнения, анализ динамики метеорологических показателей в реальном времени является крайне ресурсоёмкой задачей.

В данной работе предложен гибридный метод обработки временных рядов, сочетающий в себе возможность параллельной обработки накапливаемой в течение длительного промежутка времени информации и построения достаточно точного прогноза в кратковременной перспективе без привлечения больших вычислительных мощностей.

2. Эффективная обработка данных в модели линейной регрессии

Рассматриваются n пар наблюдений вида $(t_i, y_i),\ i = 1, \dots, N$, t_i – временная метка, y_i – измерение прибора. В данной работе строится математическая модель линейной регрессии вида:

$$y = f(t) + \varepsilon(t), \tag{1}$$

где $f(t)$ – функция регрессии, $\varepsilon(t)$ – случайная величина. Функция регрессии представима в виде

$$f(t_i) = \beta_1 f_1(t_i) + \cdots + \beta_m f_m(t_i) = \vec{F}(t_i) \cdot \vec{\beta} \tag{2}$$

где $\vec{F}(t) = (f_1(t), \dots, f_m(t))$ – вектор строка из m функций, $\vec{\beta} = (\beta_1, \dots, \beta_m)$ – вектор-столбец неизвестных коэффициентов. Коэффициенты оцениваются с помощью метода наименьших квадратов в сочетании с методом накопления «канонической информации» (КИ) [1] (T, v, V, n), где

$$T = \sum_{i=1}^{n} \vec{F}^T(x_i)\vec{F}(x_i),\ \ v = \sum_{i=1}^{n} \vec{F}^T(x_i) y_i, \tag{3}$$

$$V = \sum_{i=1}^{n} y_i^2. \tag{4}$$

На основании этих данных получается оценка коэффициентов регрессии $\vec{\beta} = T^{-1}v$, функции $\overline{f(x)} = \vec{F}(x)T^{-1}v$, которая аппроксимирует исходный ряд, и коридора погрешности этой функции $D\,\overline{f(x)} = \frac{V - v^T T^{-1} v}{n-m}\,\vec{F}(x)T^{-1}\vec{F}^T(x)$.

Такой подход позволяет разделить обработку данных на две фазы: выделение промежуточной информации и её последующая обработка [1]. Если рассмотреть два набора статистических данных размера n_1 и n_2, то в силу аддитивности КИ $(T, v, V, n) = (T_1 + T_2,\ v_1 + v_2,\ V_1 + V_2,\ n_1 + n_2)$. Это позволяет извлекать КИ из предварительно разделённого ряда одновременно на нескольких устройствах и оперативно обновлять ее при поступлении новых измерений.

3. Интегрированная модель авторегрессии скользящего среднего

Другой распространённый способ анализа временных рядов – это интегрированная модель авторегрессии скользящего среднего (ARIMA). Описывающая стационарный стохастический процесса модель состоит из авторегрессионной модели порядка p и модели скользящего среднего порядка q:

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \tag{5}$$

Модель ARIMA способна работать с нестационарными рядами. Она включает в себя операцию дифференцирования $\Delta y_t = y_t = y_{t-1}$, которая при её d -кратном повторении позволяет сделать ряд стационарным [2].

4. Преимущества и недостатки моделей

Каждая из описанных выше моделей обладает своими особенностями. Модель линейной регрессии с накоплением КИ предоставляет возможность параллельной обработки данных, позволяет добавлять новые измерения в режиме реального времени, нечувствительна к пропускам измерений во временных рядах, используется для выделения систематической составляющей ряда, учитывающей как дневную, так и годовую сезонность, по большому временному промежутку и отслеживания отклонений от общей тенденции. Однако она не позволяет делать прогноз для локальных отклонений от систематического поведения.

Модель ARIMA, напротив, способна делать точные предсказания на короткие временные промежутки, учитывая с наибольшими весами именно последние измерения во временных рядах. Однако, при добавлении новых измерений необходимо перестраивать модель или

вовсе подбирать новую с совершенно другими параметрами (p, d, q) . Кроме того, эта модель требовательна к данным, не допускает наличие пропущенных значений, а для обработки слишком большого количества измерений, например, чтобы отразить дневную и годовую сезонность, необходимы значительные вычислительные мощности.

Предлагается одновременно использовать преимущества каждой модели для эффективной обработки данных. Идея заключается в следующем: выполняется построение модели линейной регрессии с накоплением КИ для всех имеющихся значений временного ряда. Таким образом выделяется систематическая компонента ряда, которая включает в себя тренд, годовые изменения и суточный профиль. К ней добавляется прогноз модели ARIMA для фрагмента последних измерений и так формируется уточнённое предсказание.

5. Моделирование и прогнозирование рядов

Исследовался временной ряд показателей атмосферного давления на территории заповедника во Вьетнаме (метеорологическая станция "AsiaFlux" [3]) в период с 2013 по 2021 годы с интервалом в 30 минут.

Рис. 1 и 2 демонстрируют предсказания, полученные тремя способами: моделью линейной регрессии,

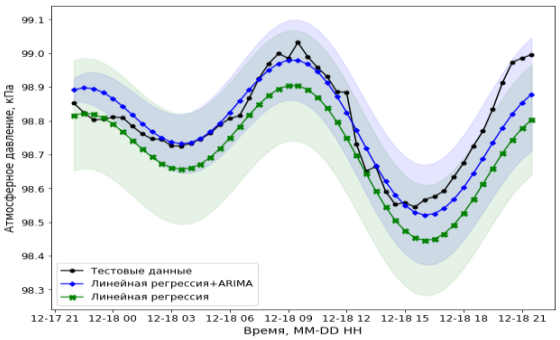

Рис.1. Сравнение предсказания гибридной модели и линейной регрессии на сутки вперёд для атмосферного давления.

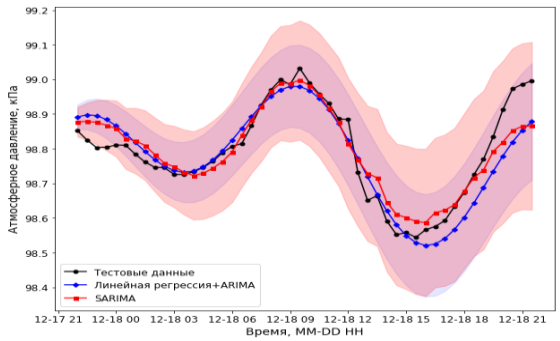

Рис.2. Сравнение предсказания гибридной модели и сезонной ARIMA на сутки вперёд для атмосферного давления.

сезонной ARIMA и предложенным гибридным методом.

В модели линейной регрессии для временного ряда использованы периодические функции синуса и косинуса вплоть до третьей гармоники, учитывающие сезонность, как годовую, так и суточную. Линия тренда аппроксимирована полиномом второй степени. Для предсказания используются данные, накопленные за длительный промежуток времени в несколько лет. Параметры сезонной модели ARIMA подобраны исходя из анализа корреляционной и автокорреляционной функций исходного ряда. Предсказание строилось по данным за предшествующую неделю.

Для сравнения полученных предсказаний используются следующие метрики: средняя абсолютная ошибка (MAE), средняя квадратическая ошибка (MSE) и средняя абсолютная ошибка в процентах (MAPE) [4]. Случайным образом выбраны 10 недельных отрезков из всего ряда и построены предсказания на сутки вперёд для каждой модели (таблица I).

Таблица I. Среднее значение метрик для разных моделей временных рядов

	MSE	***MAE***	***MAPE***
Линейная регрессия	0,0315	0,1473	0,1981
Сезонная ARIMA	0,0093	0,0749	0,0758
Гибридная модель	0,0074	0,0436	0,0441

Минимальное значение среди всех метрик достигается гибридной моделью, что говорит о более высокой точности построения предсказания, чем у каждой из исходных моделей. Кроме того, гибридная модель не требует больших вычислительных и временных ресурсов.

6. Заключение

В работе было проведено сравнение трёх моделей прогнозирования метеорологических временных рядов – линейной регрессии с возможностью накопления канонической информации, сезонной модели ARIMA и гибридной. Для каждой из моделей помимо предсказания был получен его коридор погрешности. Предложенная смешанная модель представила достаточно точный прогноз с наименьшими затратами времени и ресурсов, поэтому её можно считать оптимальной для быстрой обработки метеорологических временных рядов.

Благодарности

Авторы благодарят сотрудников ИПЭЭ РАН им. А.Н. Северцова Ю.А. Курбатову и В.К. Авилова за предоставленные данные метеорологических величин.

Литература

[1] Golubtsov, P. V. The concept of information in big data processing / P. V. Golubtsov // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52. – P. 38-43.

[2] Бокс, Дж. Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1994. – 407с.

[3] Сайт метеорологической станции AsiaFlux [Электронный ресурс]. – Режим доступа: http://asiaflux.net/index.php?page_id=86 (25.05.2022).

[4] Chicco, D. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation / D. Chicco, M. J. Warrens, G. Jurman // PeerJ Computer Science. – 2021. – Vol. 7. – P. e623.