КРАТКОСРОЧНОЕ ПРОГНОЗИРОВАНИЕ УРОВНЯ ЗАБОЛЕВАЕМОСТИ COVID-19 В РОССИИ НА ОСНОВЕ ШТРАФНОГО СПЛАЙНА В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

Я.А. Михайлова, студент гр. 8И8Б
Е.А. Кочегурова, к.т.н., доц. ОИТ ИШИТР
Томский политехнический университет
E-mail: kocheg@tpu.ru

Введение

Прогнозирование временных рядов является актуальной задачей во множестве прикладных областей. В настоящее время самой острой проблемой, охватившей весь мир, является пандемия COVID-19. И прогнозирование в этом вопросе является одним из шагов в борьбе с пандемией. Задача прогнозирования заключается в предсказании поведения временного ряда в будущие моменты времени.

Данная работа посвящена созданию программно-математических средств прогнозирования временных рядов для предсказания заболеваемости COVID-19 в России.

Существующие алгоритмы

На данный момент во всем мире активно обсуждаются результаты исследований по анализу данных о COVID-19, включая прогнозирование распространения заболевания, смертности и выздоровления. В основе прогнозирования лежат математические модели на основе искусственных нейронных сетей и машинного обучения, авторегрессионные модели, марковские процессы, на основе имитационного моделирования и робастного статистического прогнозирования и ряд других подходов [1].

Различны и результаты прогнозирования заболеваемости COVID-19. Так, в работе [2] рассматривается модель прогнозирования на основе нейронных сетей с точностью прогноза 24,7% и 37,9% для Бразилии и Португалии соответственно. Точность прогноза оценивалась по формуле относительной погрешности. Прогнозирование распространения COVID-19 в Малайзии, Марокко и Саудовской Аравии также проведено с использованием нейронных сетей [3]. Варьируя параметры нейронной сети, авторы получили разные результаты, например, для некоторых случаев ошибка RMSE составляет более 8,5%.

Описание алгоритма

В данной работе для решения задачи прогнозирования был разработан алгоритм на языке программирования Python в облачной платформе Colaboratory. Алгоритм построен на основе штрафного P-сплайна для данных, поступающих в реальном масштабе времени группами.

Сплайны – базисные, регрессионные, сглаживающие – довольно популярная модель для решения задачи прогнозирования в апостериорном режиме. В данной работе используется вариационный подход к получению штрафного сплайна (P-сплайна). Штрафные сплайны аналогично регрессионным, имеют малое число узлов, а со сглаживающими их объединяют штрафы за негладкость. Применение штрафного сплайна позволяет получить искомое гладкое значение.

Для прогнозирования в реальном времени получены рекуррентные формулы расчета штрафного сплайна на основе модификации экстремального функционала для данных, объединенных в группы [4].

$$J(S) = (1-\rho)(h\square \int_{t_0^i}^{t_h^i} t)\Big]^2\, dt + \rho\sum_{j=0}^{h}\Big[S(t_j^i) - y(t_j^i)\Big]^2 \qquad (1)$$

Расчётные формулы для прогноза, полученные на основе (1), содержат настраиваемые параметры алгоритма, позволяющие регулировать точность и другие параметры эффективности. Параметрами настройки являются: весовой коэффициент ρ, устанавливающий баланс между сглаживающими и интерполяционными свойствами сплайна, h – количество измерений внутри i-го звена сплайна, интервал дискретизации процесса Δt. Для вычисления прогноза в режиме реального времени используется текущий режим функционирования сплайна, т.е. звено сплайна вычисляется при поступлении нового значения с использованием h-1 предыдущих значений.

Результаты прогноза

В данной работе исследования проведены с использованием наборов данных ежедневно регистрируемых новых заболеваний COVID-19 в России. Данные взяты с сайта некоммерческого электронного проекта Our World in Data, публикующего данные о глобальных проблемах человечества [5]. Анализируемый временной ряд содержит значения за период с 31.01.2020 по 25.12.2021.

В качестве основного показателя точности прогноза использована оценка RMSE, дополненная процентным нормированием – RMSPE.

$$RMSPE = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \cdot \frac{100}{y_{\max} - y_{\min}}, [\%], \qquad (2)$$

где ($\hat{y}_i$ - y_i) – разность между прогнозируемым и реальным значением.

Для исследования глубина h предистории выбрана равной h = 10. Сглаживающий параметр ρ нормирован в диапазоне [0,1]. На рисунке 1 представлен график фрагмента данных, отражающий реальные данные, 3 варианта прогноза для значений ρ: 0,2, 0,5 и 0,9. Также на рисунке приведены значения соответствующих RMSPE-ошибок. Наилучшая оценка RMSPE = 1,82 соответствует прогнозу на основе сплайна с ρ = 0,5. Если сравнивать поведение это кривой с кривой реальных данных, она довольно сглаженная, не имеет резких скачков. Но визуально красная кривая (ρ = 0,9) больше соответствует кривой реальных данных, повторяя скачки.

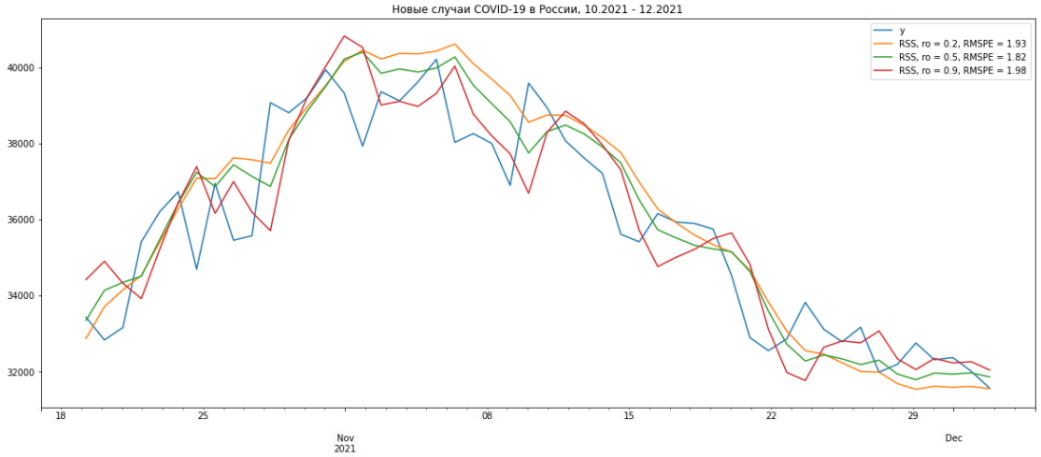

Рис. 1. Новые случаи COVID-19 в России, 10.2021 – 12.2021.

Заключение

В результате выполнения исследований был реализован алгоритм прогнозирования на основе штрафного P-сплайна на языке программирования Python. Полученный подход показал вполне допустимый результат: оценка точности RMSPE не превосходит 2%. При этом алгоритм показал лучший результат по сравнению с результатами, полученными другими авторами с использованием модели нейросетей в области прогнозирования пандемийных данных.

Список использованных источников

1. Харин Ю.С., Волошко В.А., Дернакова О.В., Малюгин В.И., Харин А.Ю. Статистическое прогнозирование динамики эпидемиологических показателей заболеваемости COVID-19 в Республике Беларусь. – Журнал Белорусского государственного университета. Математика. Информатика. 2020, №3, С. 36–50.
2. De Carvalho, K.C.M., Vicente, J.P., Teixeira, J.P., COVID-19 Time Series Forecasting – Twenty Days Ahead. Procedia Computer Science – 2022. – vol. 196. – P. 1021-1027.
3. Alassafi, M.O., Jarrah, M., Alotaibi, R. Time series predicting of COVID-19 based on deep learning. Neurocomputing. – 2022. – vol. 468. – P.335-344.
4. Кочегурова, Е.А. Гибридный подход для краткосрочного прогнозирования временных рядов на основе штрафного P-сплайна и эволюционной оптимизации /Е.А. Кочегурова, Е.Ю. Репина, О.Б. Цехан // Компьютерная оптика. – 2020. – Т. 44, № 5. – С. 821- 829. – DOI: 10.18287/2412-6179-CO-667.
5. Russia: Coronavirus Pandemic Country Profile. [Электронный ресурс]. – URL: https://ourworldindata.org/coronavirus/country/russia#how-many-tests-are-performed-each-day (дата обращения: 18.01.2022).