УДК 004.942:519.216.3
DOI: 10.20535/SRIT.2308-8893.2021.1.06

ЗАГАЛЬНА МЕТОДИКА ПРОГНОЗУВАННЯ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ НА ОСНОВІ МАТЕМАТИЧНИХ МОДЕЛЕЙ З ВИКОРИСТАННЯМ СТАТИСТИЧНИХ ДАНИХ

О.М. БЕЛАС, А.О. БЕЛАС

Анотація. Розглянуто проблематику прогнозування нелінійних нестаціонарних процесів, поданих у вигляді часових рядів, що можуть собою описувати динаміку процесів як у технічних, так і в економічних системах. Детально описано загальну методику аналізу таких даних і побудови відповідних математичних моделей на базі авторегресійних моделей та рекурентних нейронних мереж. Методику застосовано на практичних прикладах — виконано порівняльний аналіз моделей прогнозування кількості каналів обслуговування абонентів стільникового зв'язку для конкретної базової станції, виявлено переваги та недоліки кожного з методів. Сформульовано необхідність удосконалення існуючої методики та розробленні нового підходу.

Ключові слова: математичне моделювання, оброблення сигналів, нестаціонарні процеси, авторегресійні моделі, нейронні мережі, рекурентні нейронні мережі.

ВСТУП

Прогнозування на основі моделей, побудованих за експериментальними (статистичними) даними — один з найпопулярніших підходів до прогнозування динаміки процесів у технічних, соціально-економічних та фінансових системах. Завдання прогнозування нелінійних нестаціонарних процесів натепер є дуже актуальним.

Розвиток інформаційних технологій і розширення обсягів інформаційних послуг значною мірою ґрунтується на науково-технологічних розробках у галузі телекомунікаційних мереж. Дослідження показують, що сучасні мережеві технології своїм зростанням випереджають теоретичне та аналітичне розуміння мережевих взаємодій. Методи розрахунку характеристик телекомунікаційної мережі (пропускної здатності каналів, ємності буферів та ін.), що засновані на класичних моделях, не відповідають необхідним вимогам і не дозволяють адекватно оцінювати навантаження в мережі.

У праці [1] встановлено, що завдання прогнозування процесів у технічних системах ґрунтовно проаналізовано із застосуванням класичних регресійних підходів, які доволі просто використовуються як у теоретичному аспекті, так і в обчислювальному. Проте такий підхід має певні недоліки.

Тому у праці [1] запропоновано розглянути нейронні мережі типу LSTM, які також вирішують завдання моделювання послідовностей, до того ж ураховують нелінійний або комбінований вплив зовнішніх факторів, але мають деякі недоліки отриманих моделей (їх складність).

У цьому дослідженні сформулювано й описано загальну методику прогнозування нелінійних нестаціонарних процесів на основі математичних моделей, застосовано методологію на практичних прикладах — виконано порівняльний аналіз моделей прогнозування кількості каналів обслуговування абонентів стільникового зв'язку для конкретної базової станції.

ОПИС МЕТОДИКИ

Основи першої методики побудови моделей часових рядів запропонували Бокс і Дженкінс у праці [2]. Модифікована авторами методика побудови математичної моделі процесу, що складається з декількох кроків, поклала основу процесу моделювання та прогнозування часових рядів. Із розвитком математичного моделювання, появою нових завдань і методів їх розв'язання методика модифікувалась і поліпшувалась.

Існує кілька методик побудови таких математичних моделей: KDD, SEMMA, CRISP-DM [3].

У цій роботі пропонується використовувати сучасну аналітичну методологію SEMMA (Sampling, Exploring, Modifying, Modeling, Assessing) [4]. Методологія добре орієнтована на побудову моделей процесів, для яких, як правило, набагато складніше поставити експеримент та отримати інформативні експериментальні дані в достатньому обсязі. Методологія більш орієнтована саме на процес дослідження, більше фокусуючись на побудові моделей, а не на бізнес-розумінні або інтеграції з технічними системами.

1. Sampling — відбір і завантаження даних у проект, попереднє оброблення експериментальних даних На цьому етапі виконуються ретельний відбір даних, фільтрація, видалення пропусків та пошкоджених або некоректних даних, зведення до найбільш зручного вигляду, підготовка остаточної «вітрини даних» до роботи. Формально можна виокремити такі операції на цьому етапі:

– коригування даних — заповнення пропусків та зменшення викидів (екстремальних імпульсних значень), що виходять за основний діапазон значень змінних. Некоректні виміри замінюються інтерпольованими або усередненими даними;

– ортогональні перетворення та цифрова фільтрація даних з метою вилучення шумових складових (за необхідності);

– нормування даних: їх логарифмування або зведення до зручного діапазону їх зміни, наприклад, від 0 до 1; від -1 до +1; від +10 до -10 і т.ін.

2. Exploring — розуміння суті даних, пошук трендів, аномалій та взаємозв'язків. Для цього використовуються візуалізація, кластеризація та асоціація даних. Виконується аналіз даних на можливу наявність нелінійностей за допомогою множини статистичних критеріїв. Використовуються тести на тренд та на гетероскедастичність для розуміння належності досліджуваного процесу до певного класу нестаціонарності [5].

3. Modifying — модифікування даних шляхом створення, обрання та перетворення змінних; остаточне перетворення вихідних даних, розбиття

вибірки на навчальну і тестову. Правильне обрання тестової вибірки є важливим завданням та запорукою коректного навчання моделі.

4. Modeling — складається з аналітичних інструментів, призначених для побудови обраних математичних моделей, що дають бажаний результат. У праці [1] виконано огляд сучасних підходів до прогнозування процесів, подано статистичні дані у вигляді часових рядів. Розглянуто модель авторегресії з інтегрованим ковзним середнім (ARIMA):

$$y'^{(k)} = a_0 + \sum_{i=1}^{p} a_i y'(k-i) + \sum_{j=1}^{q} b_j \varepsilon(k-j) + \varepsilon(k).$$

Методику автоматичної побудови такої моделі для обраного ряду досліджено та описано Хайндманом у праці [6].

Розглянуто нейронні мережі типу LSTM. Мережі з довгокороткостроковою пам'яттю (Long Short Term Memory) — спрощено LSTM — особливий вид рекурентних нейронних мереж, здатних до навчання довгострокових залежностей. Їх запропонували Хохрейтер і Шмідхубер [7] і доопрацювали та популяризували інші дослідники [8, 9]. Вони дають змогу отримати високоякісні результати на великій різноманітності проблем і нині широко застосовуються. Однак поки не існує загальної методики побудови такого типу нейронних мереж, зокрема методики вибору початкових ваг, алгоритму оптимізації, функції активації, архітектури мережі тощо. Через складність даних моделей створення такої методики є надважливим завданням.

5. Assessing — побудова графіків та використання критеріїв для оцінювання якості прогнозів, отриманих у процесі моделювання та дослідження; Обрання кращої з оцінених моделей-кандидатів. Застосування моделі до розв'язання основного завдання — прогнозування, керування, поглиблене дослідження процесу та остаточне встановлення її придатності.

МЕТРИКИ ДЛЯ ОЦІНЮВАННЯ ЯКОСТІ ПРОГНОЗІВ

Щоб визначити найбільш придатну модель для розв'язання певної задачі, використовують різні метрики для оцінювання якості прогнозів.

Для оцінювання якості прогнозів зазвичай використовують множину взаємно доповнювальних статистичних критеріїв. Наприклад, значення середньоквадратичної похибки залежить від масштабу даних, а тому недостатньо використовувати тільки цей статистичний параметр для аналізу якості прогнозу. Поглиблене оцінювання якості прогнозів досягається з використанням критеріїв, які дають відносні оцінки у відсотках. Найпоширеніші статистичні критерії якості прогнозу та їх призначення [10]:

— середньоквадратична похибка моделі

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{k=1}^{N} [\hat{y}(k) - y(k)]^2}{N}};$$

— середня абсолютна похибка моделі

$$\mathrm{MAE} = \frac{\sum_{k=1}^{N} |\hat{y}(k) - y(k)|}{N};$$

— середня абсолютна похибка моделі у відсотках

$$\text{MAPE} = \frac{100}{N}\sum_{k=1}^{N}\left|\frac{\hat{y}(k)-y(k)}{y(k)}\right|,$$

де $\hat{y}(k)$ — прогнозоване за моделлю значення; *y(k)* — реальне вимірювання; N — довжина вибірки.

Із можливих кандидатів необхідно обирати ту модель, для якої RMSE, MAE і MAPE набувають мінімального значення.

Однак трапляються випадки, коли деяка модель є пріоритетною за одними критеріями, а інша модель — за іншими. У таких випадках потрібно обрати пріоритетний критерій або побудувати інтегральний критерій оцінювання якості прогнозів, що наразі є дуже актуальним і досі невирішеним завданням.

ПРИКЛАДИ ЗАСТОСУВАННЯ МЕТОДИКИ НА СТАТИСТИЧНИХ ДАНИХ

Побудовані моделі для прогнозування необхідної кількості каналів обслуговування абонентів стільникового зв'язку деякого українського оператора у формі регресії та LSTM нейронних мереж.

Реалізація алгоритмів спирається на праці [11, 12].

Основні показники точності даних моделей та якості оцінювання RMSE, MAE, MAPE подано у табл. 1–3.

Побудовані графіки порівняння реальних та прогнозованих значень різними моделями показано на рис. 1–5.

Таблиця 1. Порівняльна характеристика моделей прогнозування кількості каналів обслуговування абонентів стільникового зв'язку (базова станція 1)

Тип моделі	**Оцінки якості прогнозу**		
	RMSE	MAE	MAPE
АРІКС(2,1,2)	0,5392346	0,404728	9,289125
Нейронна мережа типу LSTM	0,3718906	0,288897	1,988698

Таблиця 2. Порівняльна характеристика моделей прогнозування кількості каналів обслуговування абонентів стільникового зв'язку (базова станція 2)

Тип моделі	**Оцінки якості прогнозу**		
	RMSE	MAE	MAPE
АРІКС(2,1,2)	0,6349195	0,474474	1,587532
Нейронна мережа типу LSTM	0,4224225	0,315366	0,961578

Таблиця 3. Порівняльна характеристика моделей прогнозування кількості каналів обслуговування абонентів стільникового зв'язку (базова станція 3)

Тип моделі	**Оцінки якості прогнозу**		
	RMSE	MAE	MAPE
АРІКС(3,1,2)	1,19619	1,019926	2,341559
Нейронна мережа типу LSTM	0,7199281	0,613519	1,314227

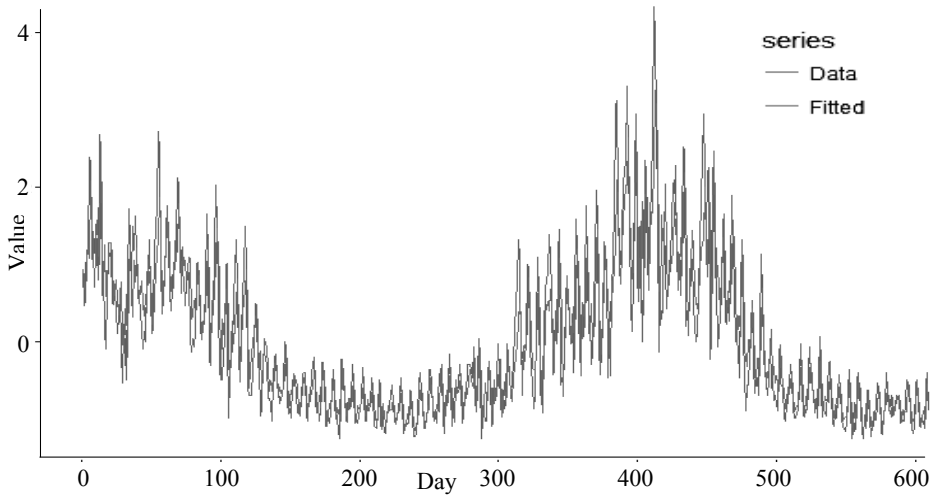

Рис. 1. Порівняльний графік реальних та прогнозованих значень регресійної моделі для базової станції 1 стільникового зв'язку

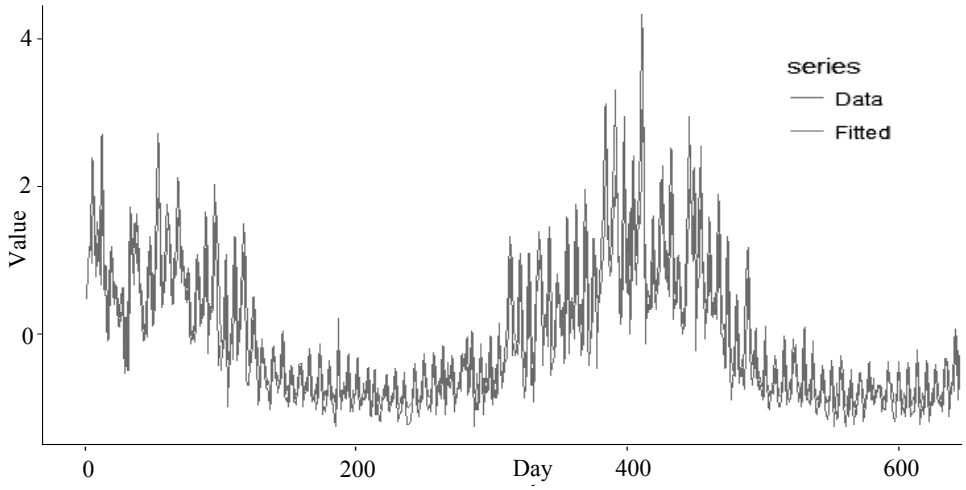

Рис. 2. Порівняльний графік реальних та прогнозованих значень нейронної мережі типу LSTM для базової станції 1 стільникового зв'язку

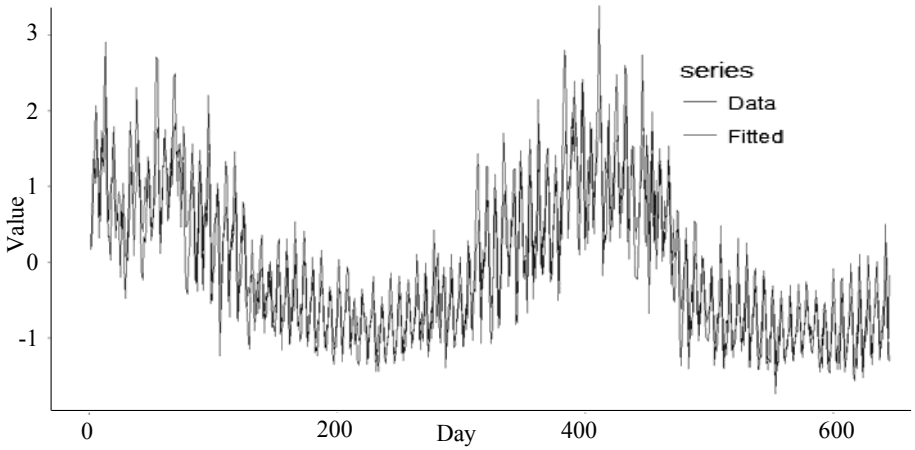

Рис. 3. Порівняльний графік реальних та прогнозованих значень нейронної мережі типу LSTM для базової станції 2 стільникового зв'язку

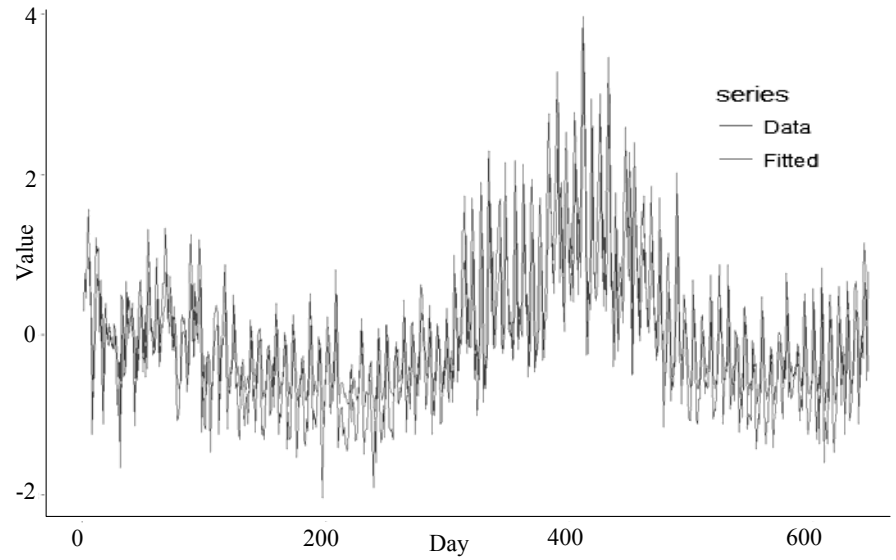

Рис. 4. Порівняльний графік реальних та прогнозованих значень регресійної моделі для базової станції 3 стільникового зв'язку

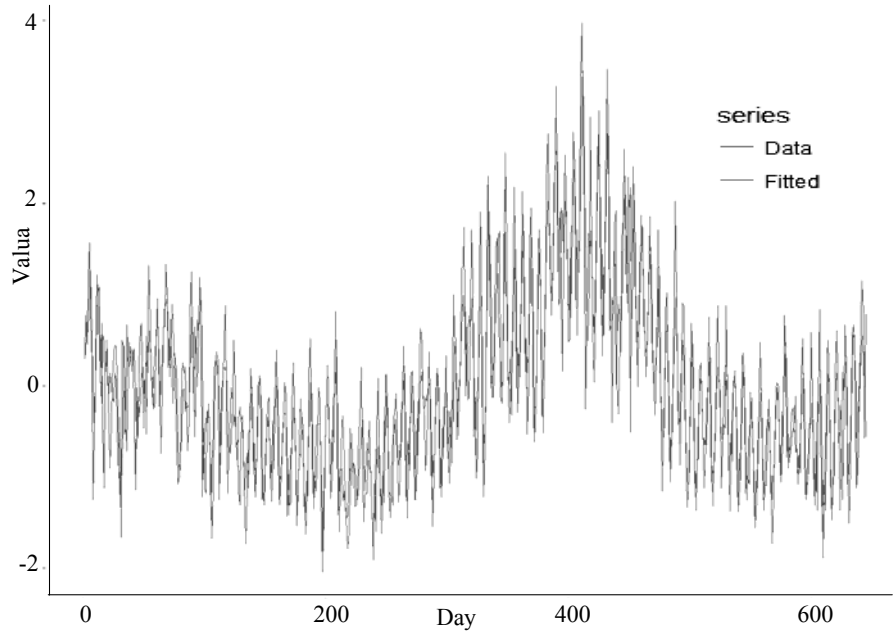

Рис. 5. Порівняльний графік реальних та прогнозованих значень нейронної мережі типу LSTM для базової станції 3 стільникового зв'язку

АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ

У ході дослідження кращою виявилася модель з використанням нейронних мереж (MAPE=1,98). Регресійні методи виявились менш точними (MAPE=9,29), однак швидшими для побудови і використання. Серед регресійних моделей найкращою виявилась модель АРІКС(2,1,2). Обрання відповідної структури моделі для автоматизованої роботи в системі на час дослідження спирається саме на важливість швидкості або точності обчислень, що підтверджує проблематику, описану у праці [1].

ВИСНОВКИ

Сформульовано й описано основні етапи загальної методики прогнозування нелінійних нестаціонарних процесів на основі сучасної аналітичної методології SEMMA. Показано досліджену методику для автоматичної побудови регресійних моделей, сформульовано проблеми існуючої методики для побудови моделей на основі нейронних мереж. Сформульовано проблему єдиного критерію якості прогнозів. Побудовано нові моделі для прогнозування вибраних процесів і у формі регресії та нейронних мереж типу LSTM, показано їх переваги та недоліки.

ЛІТЕРАТУРА

1. O.M. Belas, P.I. Bidyuk, and A.O. Belas, "Comparative analysis of autoregressive approaches and recurrent neural networks for modeling and forecasting of nonlinear nonstationary processes", *Information Technology and Security*, vol. 7, no. 1, pp. 91–99, 2019.
2. J. Box and G. Jenkins, *Time Series Analysis*. Moscow, USSR: Mir, 1974.
3. A. Azevedo and M. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview", in *Proc. of the IADIS European Conference on Data Mining*, pp. 182–185, 2008.
4. *SEMMA in SAS Enterprise Miner* [Online]. Available: https://web.archive.org/web/20120308165638/http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html/. Accessed on: Sep. 12, 2020.
5. R. Shumway and D. Stoffer, *Time Series Analysis and Its Applications*. New York, USA: Springer, 2011. doi: 10.1007/978-1-4757-3261-0.
6. R. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for R", *Journal of Statistical Software,* vol. 27, no. 3, 2008.
7. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
8. F.A. Gers, D. Eck, and J. Schmidhuber, "Applying LSTM to Time Series Predictable Through Time-Window Approaches", in *Proc. of International Conference on Artificial Neural Networks*, pp. 669–676, 2001.
9. S. Hochreiter, Y. Bengio, and J. Schmidhuber, *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies* [Online]. Available: http://www.bioinf.jku.at/publications/ older/ch7.pdf . Accessed on: Dec. 12, 2018.
10. P.I. Bidyuk, V.D. Romanenko, and O.L. Timoschuk, *Analysis of time series.* Kyiv, Ukraine: Polytechnic, 2010.
11. R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice.* Melbourne, Australia: OTexts, 2013.
12. F. Chollet, *Deep Learning with R*. New York, USA: Manning, 2017.

Надійшла 28.11.2020

INFORMATION ON THE ARTICLE

Oleg M. Belas, ORCID: 0000-0002-1595-3029, Institute of Special Communication and Information Security of National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: belas@ukr.net

Andrii O. Belas, ORCID: 0000-0001-7883-2489, Educational and Scientific Complex "Institute for Applied System Analysis" of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, e-mail: andrii.belas@gmail.com

GENERAL METHODS OF FORECASTING NONLINEAR NONSTATIONARY PROCESSES BASED ON MATHEMATICAL MODELS USING STATISTICAL DATA / O.M. Belas, A.O. Belas

Abstract. The article considers the problem of forecasting nonlinear nonstationary processes, presented in the form of time series, which can describe the dynamics of processes in both technical and economic systems. The general technique of analysis of such data and construction of corresponding mathematical models based on autoregressive models and recurrent neural networks is described in detail. The technique is applied on practical examples while performing the comparative analysis of models of forecasting of quantity of channels of service of cellular subscribers for a given station and revealing advantages and disadvantages of each method. The need to improve the existing methodology and develop a new approach is formulated.

Keywords: mathematical modeling, signal processing, nonstationary processes, autoregressive models, neural networks, recurrent neural networks.

ОБЩАЯ МЕТОДИКА ПРОГНОЗИРОВАНИЯ НЕЛИНЕЙНЫХ НЕСТАЦИОНАРНЫХ ПРОЦЕССОВ НА ОСНОВЕ МАТЕМАТИЧЕСКИХ МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ СТАТИСТИЧЕСКИХ ДАННЫХ / О.Н. Белас, А.О. Белас

Аннотация. Рассмотрена проблематика прогнозирования нелинейных нестационарных процессов, представленных в виде временных рядов, которые могут собой описывать динамику процессов как в технических, так и в экономических системах. Подробно описана общая методика анализа таких данных и построения соответствующих математических моделей на базе авторегрессионных моделей и рекуррентных нейронных сетей. Методика применена на практических примерах — выполнен сравнительный анализ моделей прогнозирования количества каналов обслуживания абонентов сотовой связи для конкретной базовой станции, выявлены преимущества и недостатки каждого из методов. Сформулирована необходимость усовершенствования существующей методики и разработки нового подхода.

Ключевые слова: математическое моделирование, обработка сигналов, нестационарные процессы, авторегрессионные модели, нейронные сети, рекуррентные нейронные сети.