DOI 10.24182/2073-6258-2021-20-3-52-63

А. В. Щербинина

Специалист,
scherbinina.av@rea.ru

Кафедра маркетинга,
Российский экономический университет В.Г. Плеханова,
Москва, Российская Федерация

А. В. Алжеев

Магистр,
alzheev@gmail.com

Факультет прикладной математики и информационных технологий,
Финансовый университет при Правительстве Российской Федерации,
Москва, Российская Федерация

Сравнительный анализ качества прогнозирования классической статистической модели и модели машинного обучения на данных российского фондового рынка

Аннотация: *Основная задача данной работы — сравнение прогностической способности классической модели машинного обучения — ARIMA, как наиболее распространенной и хорошо изученной baseline модели, и ML модели на основе последовательной нейронной сети — в данном случае LSTM. Целью является максимизация точности и минимизация ошибки — подбор наиболее подходящей модели для прогнозирования временных рядов с наивысшей точностью. Для данных математических моделей приведено описание. Также предложен алгоритм для прогноза временных рядов в рамках использования данных моделей, основанный на подходе «Rolling window» («скользящее окно»). Практическая имплементация реализована с использованием среды программирования Python с библиотеками Pandas, Numpy, pmdarima, Keras, Statsmodels. Для обучения моделей использованы биржевые данные по цене закрытия за акцию ведущих российских компания: Яндекс, ВТБ, КамАЗ, Киви, Газпром, НЛМК, Роснефть, Алроса. Проведённые исследования демонстрируют прогностическое превосходство подхода, основанного на нейронных сетях, при этом среднеквадратическая ошибка RMSE на 71% меньше аналогичного показателя для модели ARIMA, что позволяет сделать вывод о предпочтительности использования модели LSTM для данного класса задач.*

Ключевые слова: *фондовые рынки; модель машинного обучения; эконометрика; ARIMA; LSTM; алгоритм.*

A. V. Shcherbinina

Specialist,
scherbinina.av@rea.ru

Department of Marketing,
Plekhanov Russian University of Economics,
Moscow Russian Federation

A. V. Alzheev

Masterstudent,
alzheev@gmail.com

Faculty of Applied Mathematics and Information Technology,
Financial University under the Government of the Russian Federation,
Moscow, Russian Federation

Comparative analysis of the forecasting quality of the classical statistical model and the machine learning model on the data of the Russian stock market

Abstracts: *The main objective of this work is to compare the predictive ability of the classical machine learning model — ARIMA, as the most common and well-studied baseline model, and the ML model based on a sequential neural network — in this case, LSTM. The goal is to maximize accuracy and minimize error — selecting the most appropriate model for predicting time series with the highest accuracy. A description is given for these mathematical models. An algorithm is also proposed for forecasting time series using these models, based on the «Rolling window» approach. Practical implementation is implemented using the Python programming environment with the Pandas, Numpy, pmdarima, Keras, Statsmodels libraries. To train the models, we used stock data at the closing price per share of the leading Russian companies: Yandex, VTB, KamAZ, Kiwi, Gazprom, NLMK, Rosneft, Alrosa for the period. The studies carried out demonstrate the predictive superiority of the approach based on neural networks, while the RMSE is 71% less than the same indicator for the ARIMA model, which allows us to conclude that the use of the LSTM model is preferable for this class of problems.*

Keywords*: stock markets; machine learning model; econometrics; ARIMA; LSTM; algorithm.*

Введение

Анализ динамики временного ряда — комплексный набор проблем, осложняющийся, помимо нетривиальности самой задачи прогнозирования, многими внешними факторами, такими как, например, шоки на экономических и финансовых рынках и несовершенством информации.

При этом до сих пор многие прогнозы осуществляются с использованием базовых регрессионных моделей, игнорируя при этом возможности получения более качественных исследований временных рядов.

Временным рядом является череда значений величины, измеренных через одинаковые временные промежутки. В качестве примеров можно привести, например, численность населения по годам, продажи и производство товаров и услуг, рождения и смерти. Анализ временных рядов — достаточно исследованная область, так как последовательность каких-либо значения часто встречается в различных сферах деятельности человека, например, экономике, инженерии, финансах и прочих, где значение величин необходимо измерять и анализировать через определённые временные интервалы. Для исследования временных рядов разработан ряд техник и математических алгоритмов, которые позволяют вычленить из ряда необходимую информацию и использовать её для различных целей, например прогнозирования значения в будущем, опираясь лишь на имеющихся значениях.

Временной ряд отличается от пространственной выборки данных упорядоченностью, то есть у каждого наблюдения имеется временная метка или порядковый номер. Так, например, анализ зависимости счастья граждан стран от ВВП соответствующей страны будет пространственным исследованиям, так как нет временной привязки данных. Так как значения, находящиеся на временном интервале близко к друг другу, имеются более сильную связь, чем разнесённые по временному интервалу, анализ временных рядов имеет свою специфику.

Так как исследовательский интерес к прогнозированию возник достаточно давно, был разработан и закрепился в использовании определённый ряд инструментов, одним из которых является модель ARIMA[1]. Она широко используется в научных и бизнес кругах как наиболее популярная baseline модель, равно как и её модификации SARIMA и XARIMA, однако в данном исследовании будет использован базовый вариант алгоритма. Среди алгоритмов машинного / глубокого обучения зарекомендовала себе модель RNN [2] и ее модификация LSTM, описанные впервые исследователями Хохрайтером и Шмидхубером [3]. Способность модели «запоминать» значения из прошлых периодов и соответствующим образом изменять веса сети позволяет отыскивать сложные взаимосвязи во временных рядах. Так, LSTM модель, например, была использована для прогнозирования волатильности индекса S&P 500 [4–6].

В работе проводится сравнение двух вышеуказанных математических моделей. Сравнение производится на основе минимума квадратного корня среднеквадратической ошибки. Как уже было указано, среди классических статистических моделей была выбрана ARIMA из-за возможности комфортной работы с нестационарными временными рядами. Baseline моделью машинного обучения же выбрана LSTM, так как возможность «запоминания» очень важна для работы с временными рядами большой протяжённости.

Использование алгоритмов на фондовом рынке

Автоматизация работы спекулятивных трейдеров на финансовых рынках за последнее время стала очень популярной сферой приложения усилия исследователей, так как одновременно имеет в себе и содержательную научную теоретическую задачу, и прикладной финансовый интерес. Из-за быстрого наступления технологий машинного обучения, разработки новых техник и подходов к сбору, обработке и хранению информации алгоритмическая торговля развивается стремительно.

Также, так как используемые участниками фондового рынка модели постоянно совершенствуются, все вынуждены постоянно искать новые подходы ввиду повышения уровня конкуренции. Это позволяет говорить об актуальности проведения исследований применимости математических алгоритмов в данной сфере. Причём в развитии таких разработок есть резон не только для компаний, использующих эти методы для алгоритмической торговли, таких как инвестиционные банки, но и для научного сообщества, так как применение математических моделей к торговле на финансовых рынках позволяет развивать не только данную сферу, но и множество смежных.

Теоретической значимостью работы является развитие исследований по применению методов математического моделирования к биржевой деятельности. Также подобные методы могут быть использован в других сферах деятельности человека, задачи в которых могут быть рассмотрены как задачи прогнозирования значения временного ряда в будущем периоде с максимальной точностью. Практической же ценностью является гипотетическая возможность монетизации результатов работы исследования на финансовых рынках путём создания автоматизированных систем принятия решений на основе рассмотренного в работе подхода.

Фондовый рынок — это общественная площадка для торговли акциями, облигациями и производными финансовыми инструментами различных компаний. Для многих компаний прогнозирование будущих цен на те или иные виды финансовых инструментов является одной наиболее приоритетных задач, так как является основной деятельностью, позволяющей извлекать прибыль. С появления на рынках финансовых инструментов компании искали оптимальный способ прогнозирования цен на рынке. Методы прогнозирования традиционного подразделяются на 3 вида: фундаментальный и технический анализ, а также технологические / инструментальные методы. При этом все они в разной степени продолжают использоваться на финансовых рынках. И традиционные методы статистики, и её развитие — классическое машинное обучение, равно как и, в свою очередь — его развитие — глубокое обучение, также используются для этого класса задач, разумеется, с различной долей охвата (например, обучение с подкреплением — реже, а случайные леса и LSTM модели — более часто).

Самой часто используемой разновидностью моделей машинного обучения является обучение с учителем (supervised). Для предсказания выбирается какая-либо целевая переменная (например, значение цены или её дисперсия). После успеха традиционных моделей машинного обучения (метод опорных векторов, метод главных компонент, ансамбли лесов) сообщество обратило своё внимание и на нейронные сети, которые также зарекомендовали себя и в других областях человеческой деятельности (например, Computer Vision — компьютерное зрение и NLP — обработка естественного языка). Следующим логичных шагом будет являться использование моделей обучения с подкреплением, которые на данный момент не пользуются самой широкой популярностью ввиду сложности обучения, высокими требованиями к аппаратной стороне оснащения и некоторой нестабильностью. Однако, обучение с подкреплением уже неоднократной демонстрировало свою силу на, например, компьютерных играх (Atari, Dota), в которых уровень дискретности среды и вариативности не уступает, а порой превосходит таковой на финансовых рынках.

Временные ряды

Ввиду большой содержательности сферы анализа временных рядов из-за её приложения к различным отраслям человеческой деятельности данными исследованиями занимаются и с научной, и с бизнес-точек зрения.

Классической отправной точкой являются статистические методы, в частности, эконометрические — модель ARIMA, которая является де-факто стандартом в данной области. Несмотря на это, у неё имеется ряд ограничений, например, линейность и строгость предпосылок. Дальнейшие модификации модели (SARIMA), равно как и новые модели (GARCH) позволяют избавиться от этих минусов. [7].

Классическое машинное обучение позволило применить и оценить подходы к прогнозированию таких математических моделей, как, например метод опорных векторов. Другим популярным методом является случайный лес. Они часто и повсеместно используется во множестве сфер деятельности человека для разных задач — начиная от фундаментальных наук, заканчивая прогнозированием спроса в розничной торговле.

Из плеяды моделей нейронных сетей для прогнозирования значений ряда в будущем чаще всего применяют особый тип свёрточной сети — модель LSTM. На данный момент подход пользуется успехом среди исследователей и практиков. Например, К. Краусс и его коллеги провели тестирование производительности и применимости математических алгоритмов (нейронные сети, леса) и пришли к выводу о затратности и сложности использования моделей глубокого обучения [8]. Качество различных подходов сравнивалось с использованием данных американской фондовой биржи [9, 10].

Математические основы алгоритмов

Ниже изложение математических основ моделей ARIMA и LSTM.

ARIMA — это обобщенная модель ARMA (Autoregressive moving average), которая соединяет в себе AR (Autoregressive) и MA (Moving average) процессы и интегрированность разностями. ARIMA(p, d, q) состоит из следующих частей [1, 11]:

1) AR — autoregressive (p).

2) I — integrated (d). Данные становятся стационарными взятием разностей (вычитание значений ряда друг из друга).

3) MA — moving average (q).

AR модель порядка p, т.е. AR(p) выглядит так:

$$x_t = c + \sum_{i=1}^{p} \varphi_i x_{t-1} + \epsilon_i, \quad (1)$$

где x_t — стационарная переменная, c — константа, φ_i — коэффициенты автокорреляции, а ϵ_i — остатки модели, белый шум с нулевым средним.

MA модель порядка q, т.е. MA(q) выглядит так:

$$x_t = \mu + \sum_{i=0}^{q} \theta_i \epsilon_{t-1}, \quad (2)$$

где μ — математическое ожидание процесса (обычно предполагается равным нулю), θ_i — веса, θ_0 предполагается равным 1. ϵ_t — white noise со средним, равным 0. Объединение AR модели и MA модели дает нам модель ARIMA порядка (p, q):

$$x_t = c + \sum_{i=1}^{p} \varphi_i x_{t-1} + \epsilon_i + \sum_{i=0}^{q} \theta_i \epsilon_{t-1}. \quad (3)$$

p и q — порядки модели. Введение взятия разностей позволяет использовать нестационарные временные ряды.

Показатели p и q называют порядками AR и MA. Использование интегрированности позволяет применять модель в условиях отсутствия стационарности временного ряда. Вычитание уровней ряда путём взятия разностей позволяет сгладить неравномерности во временном ряду.

Так как многие процессы в реальной жизни цикличны, необходимо также обращать внимание на сезонность. Для этого существует модификация модели — SARIMA. Если ряд нестационарен, т.е. его дисперсия растёт на протяжении временного ряда, необходимо использовать взятие разностей или другие техники для стационаризации временного ряда. Для исследования можно использовать ACF и PACF (автокорреляционную и частную автокорреляционную функции), что позволит увидеть необходимость в использовании взятия разностей и аналогичных методов.

LSTM — это одна из разновидностей рекуррентных нейронных сетей, в которых есть блоки «памяти», позволяющие идентифицировать, хранить и использовать взаимосвязи между различными наблюдениями временного ряда [2, 3].

Искусственная нейронная сеть состоит из как минимум трёх слоёв — входного, внутреннего и выходного. От количества переменных зависит число нейронов во входном слое. Слои соединены связями с соответствующими весами, определяющими возможность прохождения сигнала через них. Обучение производится путём изменения этих весов. Внутренний слой имеет функцию активации (ReLu, сигмоида, Leaky ReLu), которая преобразует данные для выходного слоя. Сам же выходной слой создаёт массив прогнозов, из которого путём минимизации ошибки выбирается один. Процесс обучения нейронной сети строится на весах предыдущей итера-

ции обучения сети, которые называют эпохами, которые продолжаются до нахождения оптимальных значений весов, минимизирующих ошибку.

Рекуррентная нейронная сеть — подвид нейронных сетей, в которые добавили слой запоминаний, которые позволяет извлекать зависимости из временного ряда и использовать их для совершения прогнозов будущих значений ряда. Для увеличения количества информации, которую способна запомнить рекуррентная нейронная сеть была разработана модель LSTM, в которой имеются, помимо обычных, ещё и обратные связи, что позволяет обрабатывать последовательности данных (не только, допустим, данные по динамике цен на финансовые инструменты, но и, например, изображения, тексты и видео), в которых зависимость между разнесёнными во времени наблюдениями может иметь большой временной лаг, который в обычных нейронных сетях приводит к проблеме затухания градиента, тогда как LSTM модель этому не подвержена.

Используемые данные и метрика качества

Был поставлен ряд экспериментов по анализу качества прогнозирования моделями LSTM и ARIMA на финансовых временных рядах.

Были собраны [1] временные ряды цен на акции компаний: Яндекс, ВТБ, КамАЗ, Киви, Газпром, НЛМК, Роснефть, Алроса. Для прогноза была выбрана цена закрытия за период с 15.10.2014 по 15.11.2019 с недельным интервалом. Данные были разбиты на обучающий (для построения моделей) и тестовый (для анализа качества построенной модели) датасеты (70:30). Ввиду природы данных (временные ряды) перемешивание не производилось. Данные по динамике значений приведены ниже на рисунке 1.

В качестве метрики оценки используется RMSE — квадратный корень среднеквадратичной ошибки (между реальным и прогнозными значениями).

Формула показателя [12]:

$$RMSE = \sqrt{\frac{1}{N}\sum_1^N (x_i - \widehat{x_i})^2}\,, \qquad (4)$$

где N — суммарное число значений, а $\sum_1^N (x_i - \widehat{x_i})$ — разность истинного и спрогнозированного значения, просуммированная по наблюдениям и возведённая в квадрат.

[1] URL: https://www.finam.ru/profile/moex-akcii/gazprom/export/ (дата обращения: 14.01.2021).

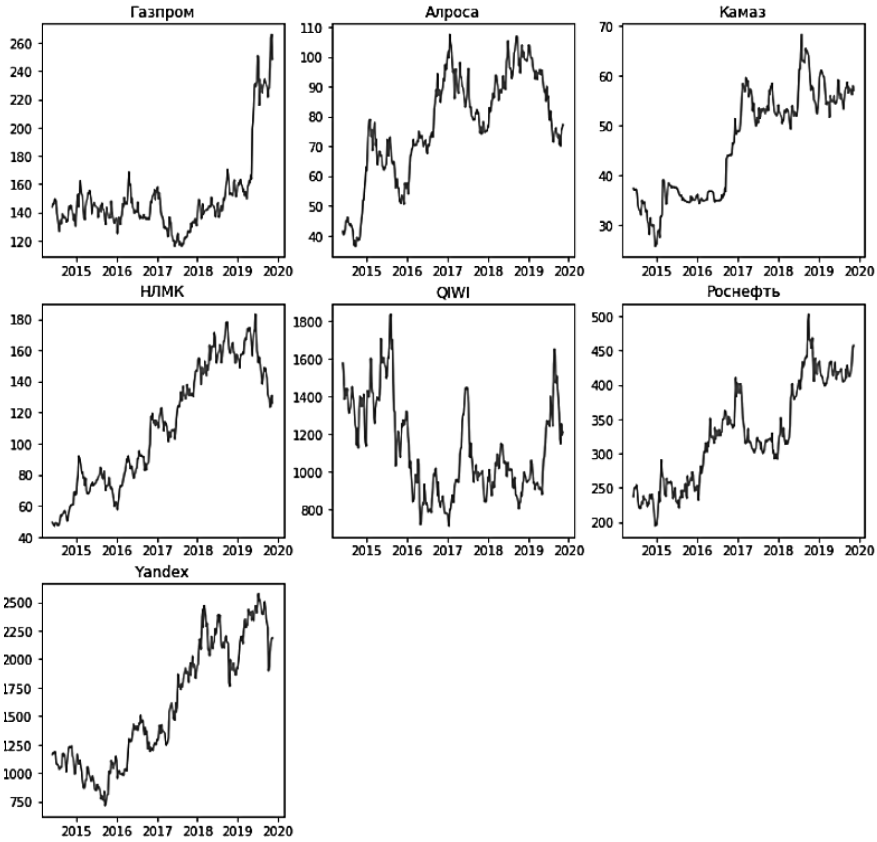

Рис. 1. Динамика изменения акций российских компаний

Источник: составлено авторами.

Имплементация алгоритмов ARIMA и LSTM

Для практической имплементации автором разработан и использован подход к прогнозированию временных рядов на основе методики «Rolling window» [13]. После каждой итерации прогноз — валидация последнее (N) значение временного ряда добавляется к пулу обучающего датасета для попытки обучения прогнозирования значения последующего члена временного ряда (N+1). Среди разных имплементация алгоритма есть, например, подобные:

1. Одношаговый без переоценки. Обучение модели на обособленном тренировочном наборе данных и производство прогноза для следующего значения ряда.

2. Многошаговый без переоценки. Обучение модели на обособленном тренировочном наборе данных и производство прогноза для следующих нескольких значений ряда.

3. Многошаговый с переоценкой. Обучение модели на динамическом тренировочном наборе данных (элементы ряда последовательно добавляются в тренировочный набор после произведения соответствующего прогноза на шаг вперед) и производство единичного или множественного прогноза для данной итерации обучения модели.

В данной работе используется третий подход. Подход реализован в среде программирования Python с библиотеками Pandas, Numpy, pmdarima, Keras, Statsmodels.

Алгоритм для моделей состоит из следующих шагов:

1. Разбиение данных на обучающий и тестовый наборы в соотношении 65:35.

2. Инициализация дополнительных переменных, в которых будут храниться исторические, будущие и прогнозные данные.

3. В цикле для каждого элемента из тестового набора (N наблюдений) обучается модель на тренировочном наборе (M наблюдений) и предпринимается попытка построения прогноза для (M+1)-го наблюдения. Прогноз сравнивается с истинным значением, записывается результат. Для прогноза используется подбор оптимальных гиперпараметров путём grid search.

4. Элемент тестового набора переходит в тренировочный. Тестовый набор становится размером N–1, а тренировочный — M+1.

5. Высчитывается на основе сделанных прогнозов ошибка.

Результаты

На рисунке 2 отображены результаты работы алгоритмов. Модель LSTM в среднем продемонстрировала более хорошую прогнозную силу, чем модель ARIMA. Среднеквадратичная ошибка моделей равна 10,75 и 26,67, соответственно, разница — 65.

Выводы

Сравнение математических моделей LSTM и ARIMA показал, что среднеквадратическая ошибка RMSE прогноза значений будущих периодов данных российской фондовой биржи ниже для модели LSTM в среднем на 64%, чем для модели ARIMA.

Тем не менее существуют различные задачи в научном и бизнес-сообществах, требующие как различных критериев оценки качества работы

	Имя компании	LSTM RMSE	ARIMA RMSE	Уменьшение RMSE, %
0	Алроса	1.120000	3.040000	63.157895
1	Газпром	3.810000	8.170000	53.365973
2	Камаз	0.380000	1.760000	78.409091
3	НЛМК	1.190000	5.340000	77.715356
4	QIWI	25.320000	63.130000	59.892286
5	Роснефть	4.610000	13.730000	66.423889
6	Yandex	39.190000	91.670000	57.248827
Среднее		10.802857	26.691429	65.173331

Рис. 2 Сравнительный анализ результатов алгоритмов ARIMA и LSTM

Источник: составлено авторами.

алгоритмов, так и различных подходов. В будущем видится релевантным сравнительное исследование других математических моделей и подходов для различных задач и используемых данных. Перспективным представляется оптимальная комбинация различных моделей, позволяющая путём «взвешивания» моделей минимизировать ошибку прогноза.

Список литературы

1. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. – М.: Дело; 2007. 504 с.
2. Schmidhuber J. Habilitation thesis: System modeling and optimization. – Munich: Technical University of Munich; 1993. 209 p. (In Germ.)
3. Hochreiter S., Schmidhuber J. Long short-term memory. *Neural Computation*. 1997; 9(8): 1735–1780. DOI:10.1162/neco.1997.9.8.1735. PMID 9377276.
4. Brownlee J. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. *Machine Learning Mastery*. 2016; https://machinelearning mastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras.
5. Gers F.A., Schmidhuber J., Cummins F. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*. 2000; 12(10): 2451–2471.
6. Кратович П.В. Нейронные сети и модели ARIMA для прогнозирования котировок. *Программные продукты и системы*. 2011; (1): 95–98.
7. Garcia1 F., Guijarro F., Moya I., Oliver J. Estimating returns and conditional volatility: a comparison between the ARMA-GARCH-M Models and the Backpropagation Neural Network. *Int. J. Complex Systems in Science*. 2012; 1(2): 21–26. ISSN 2174-6036.
8. Krauss C., Do X.A., Huck N. Deep neural networks, gradientboosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research*. 2016; 259(2). DOI:10.1016/j.ejor. 2016.10.031.
9. Chung J., Lee D., Seo Y., Yoo C.D. Deep attribute networks. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*. 2012; 3(2012). http://www.eng.uwaterloo.ca/~jbergstr/files/nips_dl_ 2012/Paper%2011.pdf.

10. Fischera T., Krauss C. Deep Learning with Long Short-term Memory Networks for Financial Market Predictions. *Eur. J. Oper. Res.* 2018; 270(2): 654–669.
11. Эконометрика. Под ред. Елисеевой И.И. – М.: Финансы и статистика; 2006. 576 с. ISBN 5-279-02786-3.
12. Armstrong J.S., Collopy F. Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons. *International Journal of Forecasting.* 1992; 8(1): 69–80. DOI:10.1016/0169-2070(92)90008-w.
13. Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. 2 edition. Melbourne, Australia: OTexts; 2018. 382 p. ISBN-13: 978-0987507112.

References

1. Magnus Ya.R., Katyshev P.K., Peresetskii A.A. Ekonometrika. Nachal'nyi kurs. – M.: Delo; 2007. 504 s.
2. Schmidhuber J. Habilitation thesis: System modeling and optimization. – Munich: Technical University of Munich; 1993. 209 p. (In Germ.).
3. Hochreiter S., Schmidhuber J. Long short-term memory. Neural Computation. 1997; 9(8): 1735–1780. DOI:10.1162/neco.1997.9.8.1735. PMID 9377276.
4. Brownlee J. Time Series Prediction with LSTM Recurrent Neural Networks in Python with Keras. Machine Learning Mastery. 2016; https://machinelearning mastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras.
5. Gers F.A., Schmidhuber J., Cummins F. Learning to Forget: Continual Prediction with LSTM. Neural Computation. 2000; 12(10): 2451–2471.
6. Kratovich P.V. Neironnye seti i modeli ARIMA dlya prognozirovaniya kotirovok. Programmnye produkty i sistemy. 2011; (1): 95–98.
7. Garcia1 F., Guijarro F., Moya I., Oliver J. Estimating returns and conditional volatility: a comparison between the ARMA-GARCH-M Models and the Backpropagation Neural Network. Int. J. Complex Systems in Science. 2012;1(2):21-26. ISSN 2174-6036.
8. Krauss C., Do X.A., Huck N. Deep neural networks, gradientboosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research. 2016; 259(2). DOI:10.1016/j.ejor. 2016.10.031.
9. Chung J., Lee D., Seo Y., Yoo C.D. Deep attribute networks. NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2012; 3(2012). http://www.eng.uwaterloo.ca/~jbergstr/files/nips_dl_ 2012/Paper%2011.pdf.
10. Fischera T., Krauss C. Deep Learning with Long Short-term Memory Networks for Financial Market Predictions. Eur. J. Oper. Res. 2018; 270(2): 654–669.
11. Ekonometrika. Pod red. Eliseevoi I.I. – M.: Finansy i statistika, 2006. 576 s. ISBN 5-279-02786-3.
12. Armstrong J.S., Collopy F. Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons. International Journal of Forecasting. 1992; 8(1): 69–80. DOI:10.1016/0169-2070(92)90008-w.
13. Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. 2 edition. Melbourne, Australia: OTexts; 2018. 382 p. ISBN-13: 978-0987507112.