

Math-Net.Ru

All Russian mathematical portal

O. E. Sorokoletova, T. V. Zakharova, Random sampling method for cryptocurrency market time series forecasting, *Sistemy i Sredstva Inform.*, 2019, Volume 29, Issue 4, 65–72

DOI: <https://doi.org/10.14357/08696527190406>

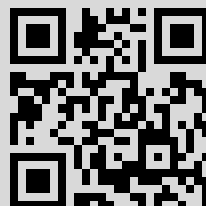
Use of the all-Russian mathematical portal Math-Net.Ru implies that you have read and agreed to these terms of use

<http://www.mathnet.ru/eng/agreement>

Download details:

IP: 2.197.88.168

April 5, 2022, 14:51:36



МЕТОД СЛУЧАЙНОГО ОТБОРА ПРИ ПРОГНОЗИРОВАНИИ ВРЕМЕННЫХ РЯДОВ РЫНКА КРИПТОВАЛЮТ

О. Е. Сороколетова¹, Т. В. Захарова²

Аннотация: Работа посвящена рассмотрению применения метода случайного отбора, или RSM-метода, в задаче классификации (прогнозировании динамики) нестационарных временных рядов рынка криптовалют. Название метода RSM — аббревиатура его полного названия Random Sampling Method. RSM представляет собой метод глубокого обучения. Одними из основных методов глубокого обучения среди использовавшихся ранее для решения данной задачи и показавших свою эффективность остаются рекуррентные LSTM (long short term memory) нейросети. В настоящей работе представлена более гибкая архитектура, построенная на базе LSTM-ячеек и имеющая таким образом все преимущества традиционного алгоритма, однако более устойчивая к проблеме дисбаланса классов. Основным отличительным признаком RSM является использование метрического обучения.

Ключевые слова: криптовалюта; временные ряды; задача классификации; прогнозирование динамики; метрическое обучение; LSTM; нейросети; глубокое обучение

DOI: 10.14357/08696527190406

1 Введение

Взгляды на вопрос применимости методов машинного (в том числе глубокого) обучения в качестве инструмента для алгоритмической торговли полярны, и в этом кроется причина большого числа дискуссий и появления множества научных, а также экономических и финансовых публикаций, посвященных доказательству эффективности того или иного инструмента. Криптовалютные временные ряды — это феномен недавнего времени, поэтому он мало изучен и представляет интерес для исследований. В данной работе рассматривается алгоритм машинного обучения, эффективный при анализе временных рядов рынка криптовалют.

Методы глубокого обучения требуют большого объема данных для тренировки. В случае криптовалютных временных рядов это означает, что данные должны быть высокочастотными, т. е., как показывает практика, интервал временной метки должен составлять минуту и не более. Однако на большом объеме

¹Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, olgas020697@gmail.com

²Московский государственный университет имени М. В. Ломоносова, факультет вычислительной математики и кибернетики; Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, tvzaharova@mail.ru

высокочастотных данных, особенно таких нестабильных, как криптовалюта, возникает новая проблема — проблема нестационарности, с которой и призван бороться метод случайного отбора.

Метод случайного отбора представляет собой метод метрического обучения, который учится оценивать схожесть между целевым и эталонным экземпляром перед выставлением метки, где количество эталонных экземпляров должно быть равно числу классов в задаче классификации. Важной особенностью метода является алгоритм выбора эталонных последовательностей, который будет определен ниже как схема отбора.

Метрическое обучение хорошо изучено на задаче классификации изображений. В случае значительно ограниченного числа входных данных и, главное, недостаточных знаний об их природе, что, в частности, не позволяет обнаружить подходящее признаковое пространство (как качественно, так и количественно), установить метрику бывает затруднительно. Этот случай полностью соответствует характеристикам рынка криптовалют. При решении задачи классификации изображений в работе [1] схожие затруднения были разрешены при помощи вовлечения нейросетевой архитектуры. В данной статье также при помощи нейронной сети рассмотрен алгоритм решения задачи прогнозирования временных рядов.

Идеи реализации метрического обучения в задаче классификации описаны в источниках [2, 3].

2 Постановка задачи классификации

Формальная постановка задачи предсказания движения цены акции на рынке может быть представлена следующим образом.

Имеются три возможных варианта движения цены после временной отметки t : вверх (up), вниз (down) или без изменений (static). Будем рассматривать распределение логарифмической доходности R_t :

$$R_t = \log \left(\frac{P_t}{P_{t-1}} \right),$$

где P_t — значение цены в момент времени t , взятое из временного ряда. Обозначим через $p_t^{(\text{up})}$, $p_t^{(\text{down})}$ и $p_t^{(\text{static})}$ вероятности наступления соответствующих верхним индексам событий. Эти вероятности будут получены в результате обучения модели. Таким образом, возникает задача трехклассовой классификации:

$$X_t \in \{\text{up}, \text{down}, \text{static}\}.$$

3 Предобработка данных и схема отбора

На вход модели подаются OHLC-временные ряды (Open, High, Low, Close — цена открытия, наиболее высокая цена, наиболее низкая цена и цена

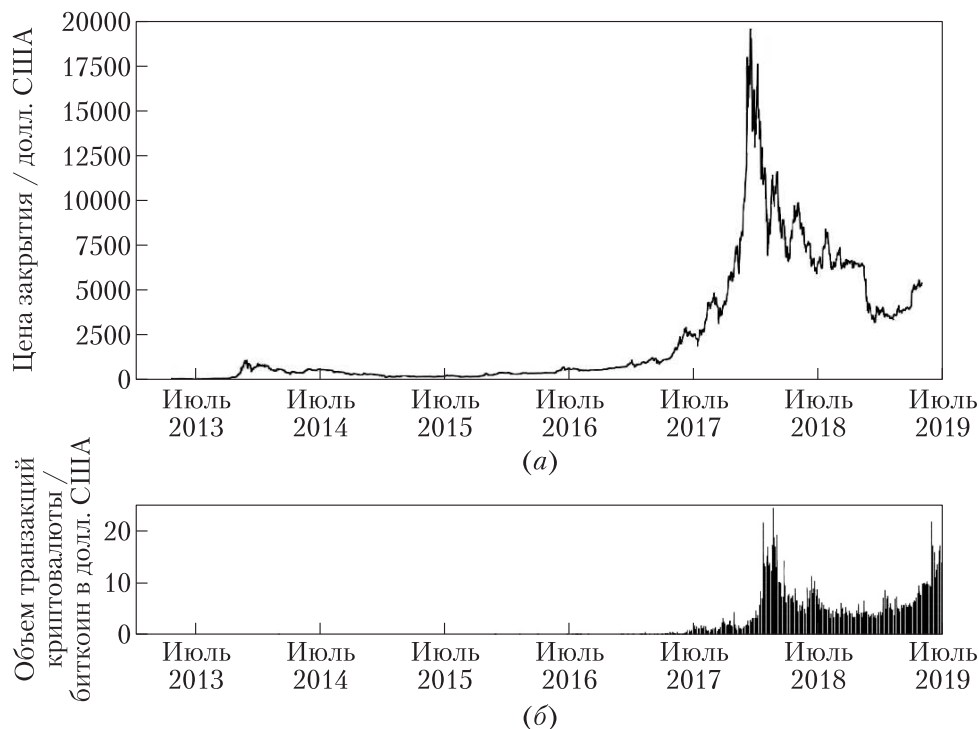


Рис. 1 Цены закрытия и динамика объема транзакций криптовалюты биткоин в долларах США

закрытия соответственно). Как известно, дисбаланс между значениями признаков может вызвать неустойчивость работы модели, ухудшить результаты обучения и замедлить его процесс, поэтому необходимо производить нормализацию данных. Так как изучаемые ряды нестационарны, будем сначала извлекать последовательность из набора, а затем применять к ней нормализацию. Для того чтобы определиться со стратегией нормализации, обратимся к рис. 1.

Большинство криптовалютных временных рядов, как и тот, что представлен на рис. 1, имеют пик в конце 2017 г., а следовательно, известен максимум ряда. Из этого можно заключить, что оптимальным выбором будет применение минимаксной нормализации:

$$x_{ti}^{\text{norm}} = \frac{x_{ti} - \min(x_t)}{\max(x_t) - \min(x_t)},$$

где индекс i соответствует номеру признака (Open, High, Low или Close), индекс t — временной отметке во входной последовательности, а операции взятия

минимума и максимума применяются к компонентам x_{ti} для всех $i = 1, \dots, 4$ на отрезке $[t - j + 1, t]$, где j — длина входной последовательности. В результате получим значения от 0 до 1.

Следующим шагом избавимся от очевидных выбросов, например, по порогу α : удалим имеющие метку «up» или «down» последовательности, для которых $|R_t| > \alpha$. На этом вся необходимая предобработка данных будет выполнена, и можно приступать к применению схемы отбора.

Схема отбора — это алгоритм, по которому осуществляется выбор тройки эталонных последовательностей, подающей затем на вход модели. Она построена, основываясь на предположении, что финансовые данные нестационарны, что обуславливает необходимость выбора последовательностей исключительно среди недавнего прошлого. Более формально, для заданной входной последовательности x_t схема подразумевает выбор последовательностей из интервала $[t - k - l, t - k]$, где k — размер окна для простого скользящего среднего, а l — гиперпараметр, определяющий размер этого интервала.

4 Основная идея метода случайного отбора

В основе метода лежит предположение о том, что схожесть пары последовательностей может быть охарактеризована при помощи классов, к которым они принадлежат: последовательность с меткой «up» более близка последовательности с меткой «up», чем последовательности с меткой «down» или «static».

На вход модели подаются пары элементов, один из которых — это последовательность, подлежащая классификации, а другой — набор последовательностей (три последовательности, помеченные как «up», «down» и «static» соответственно), выбранных из недавнего прошлого. Затем осуществляется получение внутренних представлений путем независимого кодирования каждой последовательности и сравнение внутреннего представления входной последовательности и внутренних представлений выбранных последовательностей. Выходным значением является класс, к которому предположительно принадлежит целевая последовательность. Работа метода представлена на рис. 2.

Оптимизация параметров модели производится таким образом, чтобы достигались внутренние представления, для которых верно: среди внутренних представлений последовательностей из тройки более близко внутреннему представлению целевой последовательности то, которое имеет метку класса, совпадающую с истинной в данном случае меткой. Близость внутренних представлений оценивается при помощи косинусной меры сходства.

5 Работа кодировщика

Рассмотрим процесс получения внутренних представлений подробнее. Кодировщик, осуществляющий эту работу, состоит из двух частей: LSTM и дву-

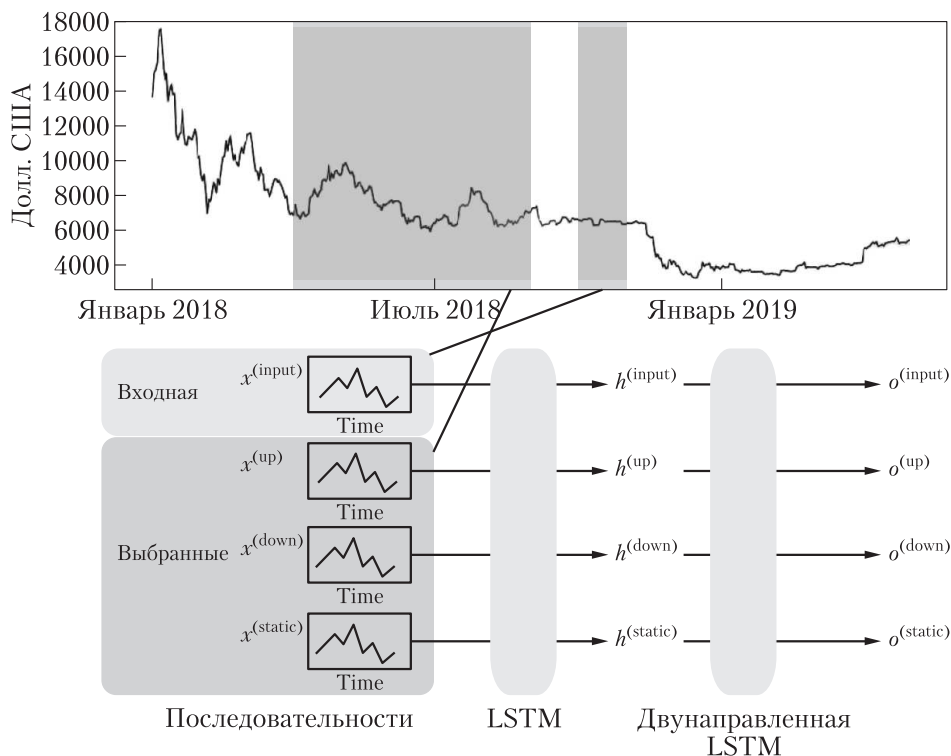


Рис. 2 Полная схема метода. Дана входная последовательность $x_t^{(input)}$, последовательности $x_t^{(up)}$, $x_t^{(down)}$ и $x_t^{(static)}$ случайно выбираются из темного окна $[t - k - l, t - k]$. Затем входная и выбранные последовательности независимо проходят через кодировщик с целью получения и обработки внутренних представлений: LSTM и двунаправленную LSTM

направленной LSTM. Долгая краткосрочная память (LSTM) — это особая разновидность архитектуры рекуррентных нейронных сетей, способная к обучению долговременным зависимостям. Не будем останавливаться на ней подробно, но заметим, что в ее блоке используются четыре взаимодействующих слоя, функционирование которых описывает следующая система уравнений:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{mi}m_{t-1} + W_{ci}c_{t-1} + b_i); \\ f_t &= \sigma(W_{xf}x_t + W_{mf}m_{t-1} + W_{cf}c_{t-1} + b_f); \\ c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{mc}m_{t-1} + b_c); \\ m_t &= \text{out}_t \tanh(c_t), \quad \text{out}_t = \sigma(W_{xo}x_t + W_{mo}m_{t-1} + W_{co}c_t + b_o), \end{aligned}$$

где σ — сигмоидальная функция; W_{ij} и b_j — параметры модели; x_t несет смысл вновь приобретенной информации, а m_t — накопленной памяти.

Для того чтобы описать первую часть кодировщика, зафиксируем входную последовательность в момент времени t : $x_t^{(\text{input})}$, а также соответствующие ей выбранные $x_t^{(\text{up})}$, $x_t^{(\text{down})}$ и $x_t^{(\text{static})}$. Затем преобразуем их независимо во внутренние представления $h_t^{(\text{input})}$, $h_t^{(\text{up})}$, $h_t^{(\text{down})}$ и $h_t^{(\text{static})}$ с помощью LSTM:

$$h_t = \text{LSTM}(x_t).$$

Здесь верхние индексы не указаны для краткости.

Следующим шагом необходимо установить отношения между $h_t^{(\text{input})}$, $h_t^{(\text{up})}$, $h_t^{(\text{down})}$ и $h_t^{(\text{static})}$, за что отвечает вторая часть кодировщика — двунаправленная LSTM-нейросеть. Она пропускает последовательности в прямом ($h_t^{(\text{input})}$, $h_t^{(\text{up})}$, $h_t^{(\text{down})}$, $h_t^{(\text{static})}$) и обратном ($h_t^{(\text{static})}$, $h_t^{(\text{down})}$, $h_t^{(\text{up})}$, $h_t^{(\text{input})}$) порядке, генерируя выходное значение как результат сложения:

$$o_{ti} = h_{ti}^{\rightarrow} + h_{ti}^{\leftarrow} + h_{ti},$$

где h_{ti}^{\rightarrow} (h_{ti}^{\leftarrow}) — i -й выход из двунаправленной LSTM-нейросети при прямом (обратном) прохождении; o_{ti} — i -й выход кодировщика.

Стандартно при решении задачи классификации во время обучения модели производится минимизация кросс-энтропийной функции потерь, не составляет исключения и данный случай. Выход представлен one-hot-векторами, в которых на месте верного класса установлена 1, а все остальные разряды обращены в 0.

Обозначим через e_t среднее цен скользящего окна размера T . Тогда целевые метки представимы следующим образом:

$$y_t = \begin{cases} -1, & \text{если } e_t > e_{t+T} + \epsilon; \\ 1, & \text{если } e_t < e_{t+T} - \epsilon; \\ 0 & \text{иначе.} \end{cases}$$

Здесь ϵ — пороговый параметр, введенный с целью контроля баланса классов.

6 Заключение

Криптовалютный рынок — получившая стремительное развитие и нестабильная область финансовой системы, требующая комплексного изучения. В частности, актуальна задача прогнозирования временных рядов рынка криптовалют, решение которой — трудоемкий процесс, имеющий множество нюансов. Применение глубоких технологий открывает новые возможности во многих прикладных областях, так как нейросети, благодаря своему алгоритмическому устройству, способны выучивать сложные нелинейные зависимости и выявлять скрытые

структурные признаки. Это делает поиски универсального инструмента для алгоритмической торговли на рынке криптовалют среди методов глубокого обучения осмысленными.

Основной целью данной работы ставилось исследование нового метода обучения — метода случайного отбора (RSM) — на задаче классификации (предсказания динамики цен) временных рядов, специфика которого заключается в том, что он имеет высокую эффективность в областях применения, в которых верно предположение о нестационарном характере ряда. Заметим, что данное условие может быть выполнено и в случае финансовых временных рядов общего вида. На практике метод случайного отбора имеет превосходство над ранее использовавшимися для решения рассматриваемой задачи алгоритмами глубокого обучения, в частности над теми из них, которые считаются наиболее производительными: рекуррентными LSTM-нейросетями.

Литература

1. Hilliard N., Phillips L., Howland S., Yankov A., Corley C. D., Hodas N. O. Few-shot learning with metric-agnostic conditional embeddings. arXiv:1802.04376v1, 2018.
2. Fei-Fei L., Fergus R., Perona P. One-shot learning of object categories // IEEE T. Pattern Anal., 2006. Vol. 28. Iss. 4. P. 594–611. doi: 10.1109/TPAMI.2006.79.
3. Lake B. M., Salakhutdinov R., Tenenbaum J. B. Human-level concept learning through probabilistic program induction // Science, 2015. Vol. 350. P. 1332–1338. doi: 10.1126/science.aab3050.

Поступила в редакцию 08.05.19

RANDOM SAMPLING METHOD FOR CRYPTOCURRENCY MARKET TIME SERIES FORECASTING

O. E. Sorokoletova¹ and T. V. Zakharova^{1,2}

¹Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation

²Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: This paper applies Random Sampling Method (RSM) to classification task for cryptocurrencies time series, which are not-stationary Long Short Term Memory (LSTM) networks have been demonstrated to be particularly useful for learning sequences containing longer term patterns of unknown length, such as at this task. But RSM represents another deep learning algorithm with more

flexible architecture, built on the basis of LSTM cells and thus having all the advantages of the traditional algorithm, but more resistant to the class imbalance problem. The main distinguishing feature of RSM is the use of metric learning and special sampling scheme.

Keywords: cryptocurrency; time series; forecasting; classification; metric learning; LSTM; random sampling; neural networks; deep learning

DOI: 10.14357/08696527190406

References

1. Hilliard, N., L. Phillips, S. Howland, A. Yankov, C. D. Corley, and N. O. Hodas. 2018. Few-shot learning with metric-agnostic conditional embeddings. *arXiv*. Available at: <https://arxiv.org/abs/1802.04376v1> (accessed February 12, 2018).
2. Fei-Fei, L., R. Fergus, and P. Perona. 2006. One-shot learning of object categories. *IEEE T. Pattern Anal.* 28(4):594–611. doi: 10.1109/TPAMI.2006.79.
3. Lake, B. M., R. Salakhutdinov, and J. B. Tenenbaum. 2015. Human-level concept learning through probabilistic program induction. *Science* 350:1332–1338. doi: 10.1126/science.aab3050.

Received May 8, 2019

Contributors

Sorokoletova Olga E. (b. 1996) — student, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; olgas020697@gmail.com

Zakharova Tatiana V. (b. 1962) — Candidate of Science (PhD) in physics and mathematics, associate professor, Department of Mathematical Statistics, Faculty of Computational Mathematics and Cybernetics, M. V. Lomonosov Moscow State University, 1-52 Leninskiye Gory, GSP-1, Moscow 119991, Russian Federation; senior scientist, Institute of Informatics Problems, Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; tvzaharova@mail.ru