УДК 004.942 **DOI** 10.15622/ia.21.4.1

А.Ю. Попков , Ю.А. Дубнов , Ю.С. Попков

РАНДОМИЗИРОВАННОЕ МАШИННОЕ ОБУЧЕНИЕ И ПРОГНОЗИРОВАНИЕ НЕЛИНЕЙНЫХ ДИНАМИЧЕСКИХ МОДЕЛЕЙ С ПРИМЕНЕНИЕМ К ЭПИДЕМИОЛОГИЧЕСКОЙ МОДЕЛИ SIR

Попков А.Ю., Дубнов Ю.А., Попков Ю.С. **Рандомизированное машинное обучение и прогнозирование нелинейных динамических моделей с применением к эпидемиологической модели SIR.**

Аннотация. В работе предлагается подход к оцениванию параметров нелинейных динамических моделей с помощью концепции Рандомизированного машинного обучения (РМО), основанной на переходе от детерминированных моделей к случайным (со случайными параметрами) с последующим оцениванием вероятностных распределений параметров и шумов по реальным данным. Главной особенностью данного метода является его эффективность в условиях малого количества реальных данных. В работе рассматриваются модели, сформулированные в терминах обыкновенных дифференциальных уравнений, которые преобразуются к дискретному виду для постановки и решения задачи энтропийной оптимизации. Применение предлагаемого подхода демонстрируется на задаче прогнозирования общего количества инфицированных COVID-19 с помощью динамической эпидемиологической модели SIR. Для этого в работе строится рандомизированная модель SIR (R-SIR) с одним параметром, энтропийно-оптимальная оценка которого реализуется его функцией плотности распределения вероятностей, а также функциями плотности распределения вероятностей измерительных шумов в точках, в которых производится обучения. Далее применяется техника рандомизированного прогнозирования с фильтрацией шумов, основанная на генерации соответствующих распределений и построении ансамбля прогнозных траекторий с вычислением средней по ансамблю траектории. В работе реализуется вычислительный эксперимент с использованием реальных оперативных данных о заболеваемости в виде сравнительного исследования с известным методом оценивания параметров модели, основанным на методе наименьших квадратов. Полученные в эксперименте результаты демонстрируют существенное снижение средне-абсолютной процентной ошибки (MAPE) при по отношению к реальным наблюдениям на интервале прогноза, что показывают работоспособность предложенного метода и его эффективность в задачах рассматриваемого в работе типа.

Ключевые слова: рандомизированное машинное обучение, энтропия, энтропийное оценивание, прогнозирование, рандомизированное прогнозирование.

1. Введение. Традиционно, все нелинейные случаи представляют собой трудоемкие задачи, основным подходом к решению которых до сих пор остается подход, основанный на переходе к линейным зависимостям с помощью переформулировки задач, ввода некоторых допущений, которые делают возможной линеаризацию используемых функциональных объектов и т.п.

Тем не менее, многие практические задачи, связанные с различными процессами, которые требуется моделировать, очевидно содержат в себе

нелинейные процессы, эффекты и связи, работа с которыми напрямую, без линеаризации, была бы предпочтительна. Разумеется, при условии эффективной возможности получения полезных результатов.

Одной из таких задач является задача прогнозирования количества инфицированных при распространении какого либо заболевания. Существенную актуальность данная задача получила при возникновении пандемии новой коронавирусной инфекции и вызываемой ей болезни COVID-19, которая к настоящему времени далека не только от завершения, но и от надежного контролирования. До сих пор не разработан единый эффективный подход к моделированию процесса распространения этой инфекции, но наиболее распространенная группа подходов к решению этой проблемы основана на использовании «золотого стандарта» эпидемиологии, а именно компартментных эпидемиологических моделей [2]. Эта группа моделей математически реализована в виде системы нелинейных дифференциальных уравнений, параметры которых необходимо настроить по реальным данным.

Стандартный путь решения этой формальной задачи основан на использовании метода наименьших квадратов, после чего, модель с оптимальными значениями параметров используется для прогнозирования [3–5]. При всей эффективности и теоретической обоснованности данного подхода, он не лишен недостатков, особенно в контексте нелинейных моделей, главным из которых является доказательность свойств получаемых оценок параметров и соответствующих им прогнозов. Эта проблема в той или иной степени успешно решается в рамках методов машинного обучения, также основанных на статистической теории, тем не менее, предполагающих определенные свойства используемых данных типа нормальности [6–8]. В рамках практических применений методов машинного обучения обычно эта проблема решается путем нормализации данных, что становится возможным в условиях, когда элементов данных много. В условиях же малого количества данных и при неизвестном механизме их генерации (а точнее, механизме генерации их стохастической составляющей) актуализируется проблема разработки новых подходов к решения задач в таких условиях.

Теория рандомизированного машинного обучения [9] ориентирована на решение задач с малым количеством данным без учета их вероятностных свойств. В [10] этот подход применялся к решению задач с линейными динамическими моделями, в [1] к задаче прогнозирования развития эпидемии COVID-19 с использованием статических моделей. В данной работе предлагается развитие метода энтропийного оценивания параметров нелинейных моделей [11] в направлении оценивания параметров

нелинейных динамических моделей, основанных на дифференциальных уравнениях.

2. Постановка задачи. Рассмотрим объект, описываемый моделью в виде «вход-выход»:

$$\mathbf{y} = F(\mathbf{x}, \mathbf{a}), \tag{1}$$

где векторы $\mathbf{y}$, $\mathbf{x}$ и $\mathbf{a}$ определяют выход, вход и параметры модели соответственно, вектор-функция F реализует связь между входом и выходом. Предполагается, что выход объекта в процессе своего функционирования «наблюдается» и «измеряется», в результате образуется массив его измерений.

Функциональная связь, реализуемая в модели вектор-функцией F, может быть статической, когда следующее значение выхода предполагается зависящим только от текущего значения входа, и динамической, когда следующее значение выхода зависит от нескольких значений входа. Статические и динамические связи входа и выхода модели следует понимать в общем смысле без привязки к временной шкале и т.п. Объектом исследования настоящей работы являются динамические модели.

Традиционно, динамические объекты принято реализовывать с помощью дифференциальных уравнений. В связи с этим, рассмотрим обыкновенное дифференциальное уравнение первого порядка в виде:

$$\frac{dx}{dt} = \Phi(t,x,a), \tag{2}$$

где x реализует состояние исследуемого объекта, a — параметр. В общем случае все величины кроме времени t могут быть векторными, а Φ в том числе и нелинейная.

Для перехода от представления (2) в виде дифференциального уравнения к виду (1) используем схему Эйлера с шагом h:

$$y[n+1] = y[n] + h\Phi(t[n], y[n], a) = F(\mathbf{x}, \mathbf{a}), \tag{3}$$

где n является индексом текущего узла равномерной сетки с шагом h и:

$$F(\mathbf{x},\mathbf{a}) = y[n] + h\Phi(t[n], y[n], \mathbf{a}), \quad \mathbf{x} = (t[n], x[n], y[n]).$$

Здесь и далее будем использовать систему обозначений, которая предполагает обозначение соответствующих значений величин в узлах реальной или виртуальной дискретной сетки в виде имени величины с индексом узла в квадратных скобках. Эти обозначения не подразумевают,

что все используемые величины дискретны, а подразумевают обозначение значений этих величин в конкретных точках.

3. Энтропийное оценивание параметров. Энтропийное оценивание параметров модели, подробно описанное в работах [9, 11, 12], базируется на переходе от модели с детерминированными параметрами к модели со случайными параметрами. Этот переход называется *рандомизацией* и состоит в придании неслучайным объектам случайных свойств.

Для реализации этого перехода с дальнейшей постановкой и решением оптимизационной задачи требуется использование модели в виде «вход-выход». В случае рассматриваемой здесь реализации динамических моделей в виде обыкновенных дифференциальных уравнений (ОДУ), будем использовать представление (3).

Рандомизация приводит к тому, что параметры модели рассматриваются случайными величинами со значениями из заданных интервалов и соответствующими распределениями вероятностей. То же касается и шума, который добавляется к выходу модели в каждой точке, в которой происходит оценивание. В непрерывном случае распределения будут определяться соответствующими функциями плотности распределения вероятностей (ПРВ), которые требуется оценить.

Таким образом, параметры и шумы полученной *рандомизированной модели* (РМ) будут иметь вид:

$$\mathbf{a} \sim P(\mathbf{a}), \quad \mathbf{a} \in \mathcal{A} \in R^d, \tag{4}$$

$$\boldsymbol{\xi} \sim Q(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \Xi \in R^m, \tag{5}$$

где d — размерность пространства параметров, m — количество точек оценивания и, соответственно, размерность пространства шумов. Здесь и далее полужирным шрифтом будет обозначать векторные величины из соответствующих пространств.

Интервалы параметров в общем случае предполагаются разными, что касается шумов, то в зависимости от интерпретации механизма возникновения шума, они также могут быть различными:

$$a_i \in A_i = [a_i^-, a_i^+], \quad \mathcal{A} = \bigotimes_{i=1}^{d} A_i, \tag{6}$$

$$\xi_j \in \Xi_j = [\xi_j^-, \xi_j^+], \quad \Xi = \bigotimes_{j=1}^{m} \Xi_j. \tag{7}$$

Предполагая независимость измерений выхода модели, совместное распределение шумов во всех точках реализуется произведением отдельных распределений (функций плотности), в то время как предполагать независимость каждой компоненты параметров в общем случае нельзя:

$$Q(\boldsymbol{\xi}) = \prod_{j=1}^{m} q_j(\xi). \tag{8}$$

В результате, модель (1) представляется в виде:

$$\mathbf{v} = F(\mathbf{x}, \mathbf{a}) + \boldsymbol{\xi}, \tag{9}$$

где $\mathbf{v} \in R^m$ — зашумленный выход модели, $\boldsymbol{\xi} \in R^m$ — вектор шумов.

Следуя концепции рандомизированного машинного обучения (РМО) [9], основанной на энтропийной оценке параметров модели, которая состоит в максимизации информационной энтропии распределений параметров и шумов при условии их нормировки и балансе среднего выхода с наблюдаемым выходом модели (реальными данными), сформулируем соответствующую задачу оптимизации:

$$H(P,Q) = -\left[\int_{\mathcal{A}} P(\mathbf{a}) \ln P(\mathbf{a}) d\mathbf{a} + \int_{\Xi} Q(\boldsymbol{\xi}) \ln Q(\boldsymbol{\xi}) d\boldsymbol{\xi}\right] \to \max_{P,Q}, \tag{10}$$

$$\int_{\mathcal{A}} P(\mathbf{a}) d\mathbf{a} = 1, \quad \int_{\Xi_j} q_j(\xi) d\xi = 1, \tag{11}$$

$$\int_{\mathcal{A}} F(\mathbf{x}_j,\mathbf{a}) P(\mathbf{a}) d\mathbf{a} + \int_{\Xi} \xi q_j(\xi) d\xi = \hat{y}_j, \quad j = \overline{1,m}, \tag{12}$$

где $\hat{y}_j$ — наблюдаемый выход модели в j-й точке (реальные данные).

Решая эту задачу методом множителей Лагранжа, получим энтропийно-оптимальные функции плотности параметров и шумов:

$$P^*(\mathbf{a},\lambda) = \frac{\exp\left(-\sum\limits_{j=1}^{m} \lambda_j F(\mathbf{x}_j,\mathbf{a})\right)}{\int_{\mathcal{A}} \exp\left(-\sum\limits_{j=1}^{m} \lambda_j F(\mathbf{x}_j,\mathbf{a})\right) d\mathbf{a}}, \tag{13}$$

$$q_j^*(\xi,\lambda) = \frac{\exp\left(-\lambda_j \xi\right)}{\int_{\Xi_j} \exp\left(-\lambda_j \xi\right) d\xi}, \quad j = \overline{1,m}, \tag{14}$$

где $\boldsymbol{\lambda} = \{\lambda_j\}$.

Подставляя выражения (13)-(14) в (12), получим уравнение для определения множителей:

$$\frac{\int_{\mathcal{A}} F(\mathbf{x}_j,\mathbf{a}) \exp\left(-\sum_{j=1}^{m} \lambda_j F(\mathbf{x}_j,\mathbf{a})\right) d\mathbf{a}}{\int_{\mathcal{A}} \exp\left(-\sum_{j=1}^{m} \lambda_j F(\mathbf{x}_j,\mathbf{a})\right) d\mathbf{a}} + \frac{\int_{\Xi_j} \xi_j \exp\left(-\lambda_j \xi\right) d\xi}{\int_{\Xi_j} \exp\left(-\lambda_j \xi\right) d\xi} = \hat{y}_j. \tag{15}$$

Учитывая сложную нелинейную структуру этого уравнения и наличие интегральных компонент, решать ее на практике необходимо численно. Следует также отметить, что в большинстве программных средств, реализующих численные методы, существует возможность вычисления интегралов размерности не более 3. В условиях, когда модель содержит параметры размерности больше 3, с точки зрения эффективности вычислений следует переходить к дискретным распределениям и соответствующим вычислительным задачам [13].

4. Прогнозирование общего количества инфицированных с использованием динамической эпидемиологической модели. Для демонстрации предлагаемого подхода рассмотрим задачу прогнозирования общего количества инфицированных при распространении инфекционного заболевания, которая становится особенно актуальной в настоящее время в связи с продолжающейся пандемией COVID-19. Одним из основных подходов к моделированию и последующему прогнозированию этого процесса является использование динамических моделей, основанных на дифференциальных уравнениях и имеющих длинную историю [2,14]. Этой класс моделей активно используется в настоящее время на уровне правительств для оценки и прогнозирования развития эпидемии COVID-19. Основная идея, на которых построены эти модели, состоит в разделении популяции на непересекающиеся группы (компартменты) с дальнейшим построением и оценкой характеристик перехода членов популяции из одной группы в другую. В настоящее время существуют модели со множеством различных групп, например, госпитализированных, госпитализированных в отделение реанимации и интенсивной терапии (ОРИТ), легких и среднетяжелых больных и т.п. Здесь рассмотрим наиболее распространенную модель с тремя компартментами.

4.1. Модель. Для моделирования развития эпидемии COVID-19 будем использовать модель SIR, являющуюся «золотым стандартом» современной эпидемиологии [2, 15, 16]. Она реализуется системой нелинейных дифференциальных уравнений:

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \gamma I, \quad \frac{dR}{dt} = \gamma I, \quad S(0) = 1. \tag{16}$$

Модель основана на трех группах (компартментах): уязвимых (Susceptible), инфицированных (Infected), удаленных (Removed). В группе S находятся люди, не имеющие иммунитета к инфекции, в группу I попадают заболевшие (инфицированные), в группу R — умершие и выздоровевшие.

В модели есть два параметра: β, называемый transmission rate, характеризует скорость передачи инфекции от одного члена популяции к другому и определяется как среднее количество контактов человенка за единицу времени $\times$ вероятность передачи инфекции в результате контакта; в итоге он определяет среднее количество зараженных от одного инфицированного, и γ, называемый recovery rate, который характеризует перемещение из группы I в группу R.

На основе этих параметров определяются основные индикаторы эпидемии: $d = 1/\gamma$ — средний инфекционный период (mean infectious period), в течении которого человек может распространять инфекцию, и основное репродуктивное число (basic reproduction number) $R_0 = \beta/\gamma$. Последний показывает, сколько человек каждый инфицированный может заразить за период d.

Параметр γ является характеристикой самой инфекции, в то время как β зависит не только от свойств инфекции, но и от структуры общества, плотности населения, структуры рынка труда, потоков общественного транспорта и также большого количества иных факторов, которые составляют типичную жизнь современного городского жителя. Также этот показатель можно использовать, чтобы анализировать развитие эпидемии со временем, а также оценивать эффективность принимаемых для борьбы с ней мер.

Очевидный подход, который стал применяться в начале эпидемии COVID-19 практически по всему миру, состоял как в оценке β, так и в оценке γ, т.к. инфекция, с которой столкнулось человечество, была неизвестна [17, 18]. К настоящему времени, после появления и распространения уже нескольких отдельных штаммов вируса SARS-CoV-2 параметр γ в среднем был вполне надежно оценен для каждого из штаммов [19–22]. Однако, параметр β по-прежнему требуется оценивать по наблюдениям за оперативными данными с тем, чтобы повысить точность модели (16). В

этой связи, далее будем рассматривать параметр γ фиксированным и равным 0.1, что соответствует инфекционному периоду продолжительностью 10 дней (этот период был установлен для первого, уханьского штамма).

Задача состоит в том, чтобы оценить (обучить модель) параметр β по наблюдениям за общим количеством инфицированных.

Для перехода к модели в виде «вход-выход» используем схему Эйлера с шагом h и получим следующую систему разностных уравнений:

$$\begin{aligned} S[k] &= S[k-1] - h\beta S[k-1]I[k-1], \\ I[k] &= I[k-1] + h(\beta S[k-1]I[k-1] - \gamma I[k-1]), \\ R[k] &= R[k-1] + h\gamma I[k-1], \end{aligned} \tag{17}$$

где k обозначает индекс узла сетки.

Эти выражения будем использовать в дальнейшем для обучения методом наименьших квадратов и построения и обучения рандомизированной модели.

4.2. Данные. В вычислительных экспериментах используются оперативные данные по Германии, собираемые сервисом Data Hub [23,24]. Самыми надежными оперативными данными об эпидемии являются дневные данные о регистрируемых случаях и смертях. Далее будем использовать следующие обозначения для наборов данных, полученных из первичных оперативных данных:

– Дневные абсолютные;
– Confirmed (`Cd`) — Количество инфицированных (случаев);
– Deaths (`Dd`) — Количество умершие;
– Recovered (`Rd`) — Количество выздоровевших;
– Дневные 7-дневные средние `Cd_avg`, `Dd_avg`, `Rd_avg`;
– Накопленные (общие, куммулятивные) `C`, `D`, `R`;
– Дневные относительные `Cd/N`, `Dd/N`, `Rd/N` и общие относительные `C/N`, `D/N`, `R/N`, где `N` — численность населения в изучаемой стране, на момент исследования для Германии $N = 82{,}905{,}782$ человек.

На рисунке 1 представлены используемые оперативные данные по Германии с начала эпидемии COVID-19. На рисунке слева представлены дневные данные с 7-дневным скользящим средним, справа — общие (куммулятивные) данные.

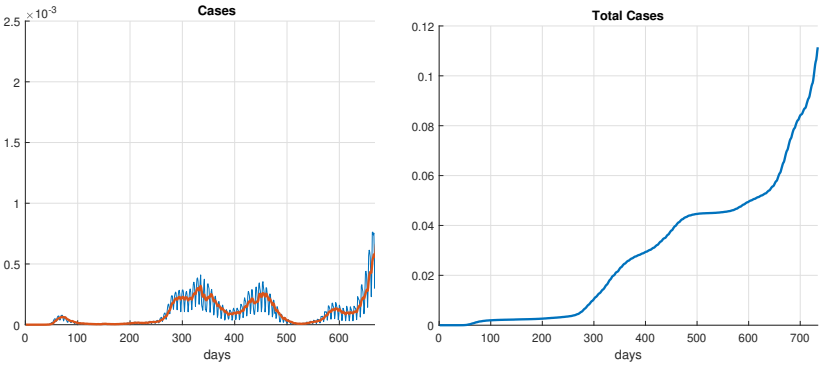

Рис. 1. Оперативные данные о случаях заболевания в Германии с начала эпидемии COVID-19

Модель SIR показывает динамику соответствующих частей популяции, следовательно, для обучения модели необходимо использовать общие (накопленные) данные на душу населения (относительные), которые в принятой концепции являются наблюдаемыми данными по траектории I в (16). Использование относительных данных позволяет работать в масштабе $[0,1]$, а также сравнивать динамику в разных регионах (странах) из-за различий в численности населения.

Общие данные по дням вычисляются по 7-дневному среднему из дневных данных следующим образом:

$$\mathtt{C}[n] = \frac{1}{N}\sum_{k=1}^{n-1} \mathtt{Cd\_avg}[k],$$

где n — индекс дня наблюдений, и выравниваются таким образом, чтобы начало данных приходилось на понедельник (первый день недели).

Необходимо отметить, что в настоящее время эпидемиологические данные и модели как правило оперируют именно недельными показателями вследствие специфики сбора данных в течении недели: выписка пациентов происходит во многих странах в воскресенье или понедельник, тестирование более активное в конце недели, в некоторых странах пятница и суббота — выходные дни и т.д.

В обучении использовались 4 точки данных (5 для обучения МНК, см. ниже), соответствующие началу недели, начиная с 230 дня от начала эпидемии (33 неделя).

4.3. Обучение модели методом наименьших квадратов. Для реализации сравнительного исследования проводится обучение модели на том же наборе данных с помощью нелинейного метода наименьших квадратов (МНК) [6–8, 25].

Оценивание МНК производится в окне длиной 4 недели, используются недельные данные, соответствующие началу недели (5 точек данных), параметр $\gamma = 0.1$, оценивается только β. Задача оптимизации β с использованием средне-абсолютной процентной ошибки MAPE (Mean Absolute Percentage Error) формулируется следующим образом:

$$\beta^* = arg\min_{\beta} \sum_{n=1}^{M} \left(\frac{I[n] - C[n]}{C[n]} \right)^2. \tag{18}$$

Для вычисления выхода модели в точках оценивания требуется решать систему ОДУ в окне каким-либо численным методом. Здесь будем использовать метод Эйлера с параметром $h = 0.1$, который определен экспериментально с помощью сравнения решения данной задачи на используемом наборе данных методом Рунге-Кутта, реализованным функцией `ode45` на платформе MATLAB.

Необходимо отметить, что сетка решения ОДУ существенно мельче дневной шкалы данных, но обучение производится только в требуемых точках, соответствующих началу недели.

4.4. Рандомизированое обучение и прогнозирование. Для построения *рандомизированной модели SIR (R-SIR)*, используем модель (17) для I с рандомизированным параметром β. Обозначим выход модели через y, вектор входа $\mathbf{x} = (S,I)$ и $y = I$, тогда:

$$y = F(\mathbf{x},\beta) = x_2 + h(\beta x_1 x_2 - \gamma x_2). \tag{19}$$

Обучение модели производится в тех же точках, что и при обучении МНК, следовательно входами РМ будут значения траекторий S и I в предыдущих точках на сетке решения ОДУ (17).

Для всех точек на интервале обучения будем использовать шум в пределах 30% во всех точках, а интервалы параметров установим в пределах 50% от оптимального β^*, полученного с помощью МНК (далее будем использовать обозначение β_{ols} для этого значения).

Следуя (10)-(12) сформулируем задачу энтропийной оптимизации для вычисления оценок распределений (функций ПРВ) $P(\beta)$ параметра β и шумов измерений $q_j(\xi)$:

$$H(P,Q) = -\left[\int_{\mathcal{A}} P(\beta)\ln P(\beta)d\beta + \int_{\Xi} Q(\boldsymbol{\xi})\ln Q(\boldsymbol{\xi})d\boldsymbol{\xi}\right] \to \max_{P,Q}, \quad (20)$$

$$\int_{\mathcal{A}} P(\beta)d\beta = 1, \quad \int_{\Xi} q_j(\xi)d\xi = 1, \quad (21)$$

$$\int_{\mathcal{A}} F(x_j,\beta)P(\beta)d\beta + \int_{\Xi} \xi q_j(\xi)d\xi = C_j, \quad (22)$$

где через C_j обозначены значения общих случаев заболевания в точках обучения.

Решая эту задачу методом множителем Лагранжа, получим энтропийно-оптимальные распределения параметра β и шумов:

$$P^*(\beta,\lambda) = \frac{\exp\left(-\sum\limits_{j=1}^{m}\lambda_j F(\mathbf{x}_j,\beta)\right)}{\int\limits_{\mathcal{A}}\exp\left(-\sum\limits_{j=1}^{m}\lambda_j F(\mathbf{x}_j,\beta)\right)}, \quad (23)$$

$$q_j^*(\xi,\lambda) = \frac{\exp(-\lambda_j\xi)}{\int\limits_{\Xi}\exp(-\lambda_j\xi)}, \quad j=\overline{1,m}, \quad (24)$$

где:

$$F(\mathbf{x}_j,\beta) = F(\mathbf{x}[k],\beta) = I[k-1] + h(\beta S[k-1]I[k-1] - \gamma I[k-1]). \quad (25)$$

Согласно принятой в машинном обучении методике, связанной с тестированием модели на тестовом наборе данных, построенную и обученную модель необходимо протестировать на наборе данных, не использовавшимся при обучении. Рассматриваемая здесь задача характеризуется малым набором данных, используемым для обучения, к тому же, сама постановка задачи и подход к ее решению, демонстрируемый здесь, не предполагает, что отдельные данные для тестирования перед прогнозированием доступны. В этой связи, под тестированием здесь будем понимать реализацию модели на интервале обучения с вычислением ошибок MAPE.

Прогнозирование с использованием рандомизированной модели осуществляется методом рандомизированного прогнозирования, описанным в [9, 13]. Он состоит в сэмплировании соответствующих распределений с последующим построением ансамбля прогнозных траекторий. Здесь существует несколько подходов к ее построению, в настоящей работе используется средняя по ансамблю траектория. Сэмплирование непрерывного распределения $P(\beta)$ осуществляется методом Acceptance-Rejection (AR) [26]. Распределения шумов, получаемых при обучении РМ, при прогнозе не используются, таким образом осуществляется фильтрация шумов в данных, реализующих неопределенность в них.

5. Результаты. Реализация всех вычислительных экспериментов была проведена на платформе MATLAB 9.7 (2019b) с использованием пакетов Optimization и Curve Fitting соответствующих версий.

В результате обучения МНК оптимальное значение $\beta_{ols} = 0{,}1070$. Интервалы параметров, использующиеся для рандомизированного обучения, показаны в таблице 1.

Таблица 1. Интервалы параметров и шумов рандомизированной модели

β_{ols}	β^-	β^+	ξ^-	ξ^+
0.1070	0.0538	0.1613	-0.3	0.3

Результаты обучения приведены на рисунке 2, реальные данные (наблюдения) с меткой `real`, траектории, полученные при обучении МНК с меткой `ols`, средняя рандомизированная траектория с меткой `avg`, также представлен ансамбль траекторий и область стандартного отклонения, вычисленная по нему.

Результаты прогнозирования представлены на рисунке 3. Показан рандомизированный прогноз общего количества заболевших I_{mee} и прогноз МНК I_{ols}. Среднее по распределению значение параметра $\beta_{mean} = 0.1091$, что говорит о том, что рандомизированный прогноз отличается от стандартного, полученного МНК. Ошибки на интервале обучения составили $\delta_{ols}^{train} < 0.01$, $\delta_{mee}^{train} = 0{,}11$, на интервале прогноза: $\delta_{ols}^{pred} = 0{,}29$, $\delta_{ols}^{pred} = 0{,}03$.

6. Заключение. Полученные результаты показывают, что предлагаемый в работе подход, основанный на энтропийном оценивании параметров и рандомизированном прогнозировании, является эффективным в задачах, подобных рассмотренной.

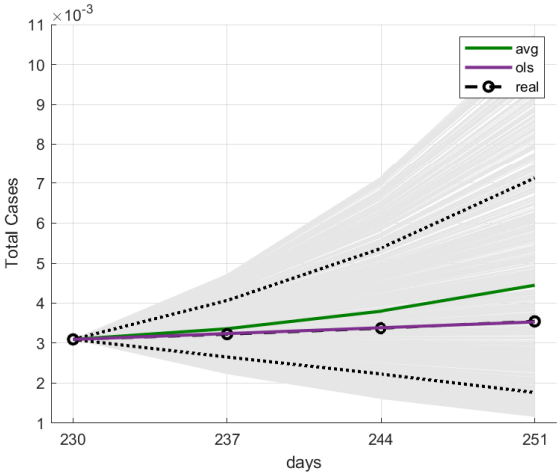

Рис. 2. Результаты обучения

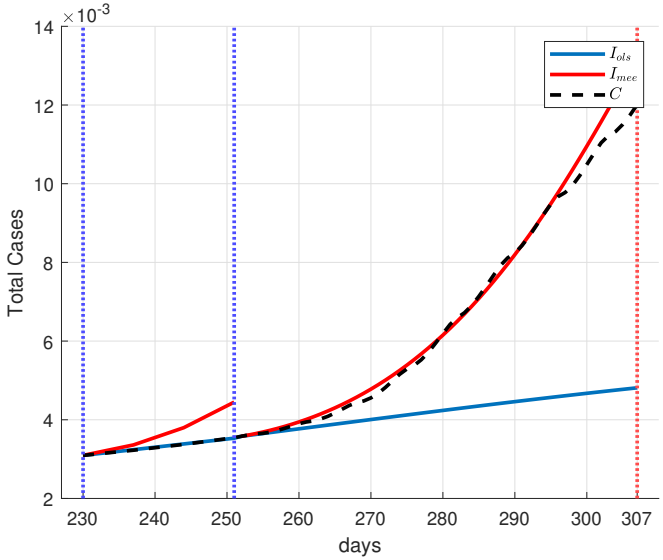

Рис. 3. Результаты прогнозирования

Эти задачи характеризуются, с одной стороны, малым количеством наблюдений, с другой стороны — процессом непознанной природы, вследствие чего данные наблюдений не могут надежно считаться качественными с статистическом смысле. Описание подобных процессов с помощью динамических моделей является популярным и эффективным подходом, однако, необходимо качественно настроить их параметры по наблюдениям. В отличие от подобных задач из других областей, рассматриваемая в работе задача характеризуется существенной неопределенностью в доступных оперативных данных, а сама модель очень чувствительна к начальным условиям. В этих условиях, предлагаемый в работе подход позволяет получить оценки параметров моделей в виде их распределений, которые вычисляются с достаточно высокой точностью в результате решения строго поставленной оптимизационной задачи. Кроме этого, моделирование неопределенности в данных с помощью аддитивного в каждой точке шума со своим распределением, повышает гибкость модели и позволяет осуществить его фильтрацию при прогнозировании.

Литература

1. Попков Ю.С., Дубнов Ю.А., Попков А.Ю. Прогнозирование развития эпидемии COVID-19 в странах Европейского союза с использованием энтропийно-рандомизированного подхода // Информатика и автоматизация, 2021, Т. 20, № 5, с. 1010-1033, https://doi.org/10.15622/ia.20.5.1.
2. van den Driessche P. Mathematical Epidemiology / ed. by Brauer F., van den Driessche P., Wu J. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. Vol. 1945 of Lecture Notes in Mathematics. P. 147–157. https://doi.org/10.1007/978-3-540-78911-6.
3. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. — М.:Юнити, 1998.
4. Лагутин М.Б. Наглядная математическая статистика. — Бином. Лаб. знаний, 2013.
5. Боровков А.А. Математическая статистика. — М.:Наука, 1984.
6. Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics), 2006. Springer, New York, 2006.
7. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer New York, 2009.
8. Мерков А.Б. Распознавание образов. Введение в методы статистического обучения. M. : URSS, 2010.
9. Попков Ю.С., Попков А.Ю., Дубнов Ю.А. Рандомизированное машинное обучение при ограниченных наборах данных: от эмпирической вероятности к энтропийной рандомизации. — М.: ЛЕНАНД, 2019. ISBN: 978-5-9710-5908-0.
10. Попков Ю.С., Дубнов Ю.А. Энтропийно-робастное рандомизированное прогнозирование при малых объемах ретроспективных данных // Автоматика и телемеханика. 2016. № 5. С. 109–127.
11. Попков А.Ю. Рандомизированное машинное обучение нелинейных моделей с применением к прогнозированию развития эпидемического процесса // Автоматика и телемеханика. 2021. № 6. С. 149–168. https://doi.org/10.31857/S0005231021060064.
12. Popkov Y.S., Dubnov Y.A., Popkov A.Y. Introduction to the Theory of Randomized Machine Learning // Learning Systems: From Theory to Practice / ed. by Sgurev V., Piuri

V., Jotsov V. Cham: Springer International Publishing, 2018. P. 199–220. ISBN: 978-3-319-75181-8. https://doi.org/10.1007/978-3-319-75181-$8_1 0$.

13. Попков Ю.С., Попков А.Ю., Дубнов Ю.А. Элементы рандомизированного прогнозирования и его применение для предсказания суточной электрической нагрузки энергетической системы // Автоматика и телемеханика. 2020. С. 148–172. https://doi.org/10.1134/S0005231019070107.
14. Kermack W.O., McKendrick A.G. Contributions to the Mathematical Theory of Epidemics // Proceedings of the Royal Society. 1927. Vol. 115A. P. 700–721.
15. Müller G.R. Zeitschrift für allgemeine Mikrobiologie / In: The Population Dynamics of Infectious Diseases: Theory and Applications. 368 S., 135 Abb., 104 Tab. London-New York, Chapman and Hall, 1984, Vol. 24, no. 2. pp. 76–76. https://doi.org/10.1002/jobm.19840240203.
16. Hethcote H.W. Three Basic Epidemiological Models // Applied Mathematical Ecology. Springer Berlin Heidelberg, 1989. pp. 119–144. https://doi.org/10.1007/978-3-642-61317-3_5.
17. Peng L., Yang W., Zhang D., Zhuge C., Hong L. Epidemic analysis of COVID-19 in China by dynamical modeling // arXiv, 2020. 10.48550/ARXIV.2002.06563.
18. Yang W., Zhang D., Peng L., Zhuge C., Hong L. Rational evaluation of various epidemic models based on the COVID-19 data of China // Epidemics, 2021. Vol. 37. p. 100501. https://doi.org/10.1016/j.epidem.2021.100501.
19. Cheng C., Zhang D., Dang D., Geng J., Zhu P., Yuan M., Liang R., Yang H., Jin Y., Xie J., Chen S., Duan G. The incubation period of COVID-19: a global meta-analysis of 53 studies and a Chinese observation study of 11 545 patients // Infectious Diseases of Poverty, 2021. Vol. 10, no. 1. https://doi.org/10.1186/s40249-021-00901-9.
20. Huang S., Li J., Dai C., Tie Z., Xu J., Xiong X., Hao X., Wang Z., Lu C. Incubation period of coronavirus disease 2019: New implications for intervention and control // International Journal of Environmental Health Research, 2021. P. 1–9. https://doi.org/10.1080/09603123.2021.1905781.
21. Li Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus — Infected Pneumonia // New England Journal of Medicine, 2020. Vol. 382, no. 13. P. 1199–1207. https://doi.org/10.1056/nejmoa2001316.
22. Nie X. et al. Epidemiological Characteristics and Incubation Period of 7015 Confirmed Cases With Coronavirus Disease 2019 Outside Hubei Province in China // The Journal of Infectious Diseases, 2020. Vol. 222, no. 1. pp. 26–33. https://doi.org/10.1093/infdis/jiaa211.
23. Guidotti E., Ardia D. COVID-19 Data Hub // Journal of Open Source Software. 2020. Vol. 5, no. 51. P. 2376. https://doi.org/10.21105/joss.02376.
24. COVID-19 Data Hub. https://www.covid19datahub.io. 2021. Accessed: 2022-06-20.
25. Флах П. Наука и искусство построения алгоритмов, которые извлекают знания из данных. ДМК Пресс, 2015.
26. Rubinstein R.Y., Kroese D.P. Simulation and the Monte Carlo method. John Wiley & Sons, 2007. Vol. 707.

Попков Алексей Юрьевич — канд. техн. наук, ведущий научный сотрудник, Федеральный исследовательский центр "Информатика и управление"Российской академии наук. Область научных интересов: энтропийные методы, рандомизированное машинное обучение, интеллектуальный анализ данных, разработка программного обеспечения. Число научных публикаций — 48. apopkov@isa.ru; улица Вавилова, 44-2, 119133, Москва, Россия; р.т.: +7(499)135-6260.

Дубнов Юрий Андреевич — научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук. Область научных интересов: машинное обучение, байесовское оценивание. Число научных публикаций — 32. yury.dubnov@phystech.edu; улица Вавилова, 44-2, 119133, Москва, Россия; р.т.: +7(499)135-6260.

Попков Юрий Соломонович — академик РАН, главный научный сотрудник, Федеральный исследовательский центр «Информатика и управление» Российской академии наук. Область научных интересов: энтропийные методы, макросистемы, рандомизированное машинное обучение. Число научных публикаций — 224. popkov@isa.ru; улица Вавилова, 44-2, 119133, Москва, Россия; р.т.: +7(499)135-6260.

Поддержка исследований. Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проекты 20-07-00683, 20-04-60119).

DOI 10.15622/ia.21.4.1

A. POPKOV , Yu. DUBNOV , Yu. POPKOV

RANDOMIZED MACHINE LEARNING AND FORECASTING OF NONLINEAR DYNAMIC MODELS APPLIED TO SIR EPIDEMIOLOGICAL MODEL

Popkov A., Dubnov Yu., Popkov Yu. **Randomized Machine Learning and Forecasting of Nonlinear Dynamic Models Applied to SIR Epidemiological Model.**

Abstract. We propose an approach to the estimation of the parameters of non-linear dynamic models using the concept of Randomized Machine Learning (RML), based on the transition from deterministic models to random ones (with random parameters), followed by estimation of the probability distributions of parameters and noises on real data. The main feature of this method is its efficiency in conditions of a small amount of real data. The paper considers models formulated in terms of ordinary differential equations, which are converted to a discrete form for setting and solving the problem of entropy optimization. The application of the proposed approach is demonstrated on the problem of predicting the total number of infected COVID-19 using a dynamic SIR epidemiological model. To do this, we construct a randomized SIR model (R-SIR) with one parameter, the entropy-optimal estimate of which is realized by its probability density function, as well as the probability density functions of the measurement noise at the points where training is performed. Next, the technique of randomized prediction with noise filtering is applied, based on the generation of the corresponding distributions and the construction of an ensemble of predictive trajectories with the calculation of the trajectory averaged over the ensemble. The paper implements a computational experiment using real operational data on the infection cases in the form of a comparative study with a well-known method for estimating model parameters based on the least squares method. The results obtained in the experiment demonstrate a significant decrease in the mean absolute percentage error (MAPE) with respect to real observations in the forecast interval, which shows the efficiency of the proposed method and its effectiveness in problems of the type considered in the work.

Keywords: randomized machine learning, entropy, entropy estimation, forecasting, randomized forecasting.

References

1. Popkov Y.S., Dubnov Y.A., Popkov A.Y. Forecasting Development of COVID-19 Epidemic in European Union Using Entropy-Randomized Approach // Informatics and Automation, 2021, Vol. 20, No. 5, pp. 1010-1033, https://doi.org/10.15622/ia.20.5.1.
2. van den Driessche P. Mathematical Epidemiology / ed. by Brauer F., van den Driessche P., Wu J. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. Vol. 1945 of Lecture Notes in Mathematics. P. 147–157. https://doi.org/10.1007/978-3-540-78911-6.
3. Aivazyan S.A., Mhitaryan V.S. Prikladnaya statistika i osnovy ekonometriki. — Moscow, Unity, 1998.
4. Lagutin M.B. Naglyadnaya matematicheskaya statistika. — Binom, 2013.
5. Borovkov A.A. Matematicheskaya statistika. — Moscow, Nauka, 1984.
6. Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics), 2006. Springer, New York, 2006.
7. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data mining, Inference, and Prediction. Springer New York, 2009.

8. Merkov A.B. Raspoznavanie obrazov. Vvedenie v metody statisticheskogo obuchenia. Moscow, URSS, 2010.
9. Popkov Y.S., Popkov A.Y., Dubnov Y.A. Randomizirovannoe mashinnoe obuchenie pri ogranichennych naborach dannykh: ot emp[iricheskoy veroyatnosti k entropyinoy randomizacii. Moscow: LENAND, 2019. ISBN: 978-5-9710-5908-0.
10. Popkov, Y.S., Dubnov, Y.A. Entropy-robust randomized forecasting under small sets of retrospective data. Automation and Remote Control. 2016. Vol. 77, pp. 839–854. https://doi.org/10.1134/S0005117916050076.
11. Popkov A.Y. Randomized Machine Learning of Nonlinear Models with Application to Forecasting the Development of an Epidemic Process // Automation and Remote Control. 2021. Vol. 82, pp. 1049–1064. https://doi.org/10.1134/S0005117921060060.
12. Popkov Y.S., Dubnov Y.A., Popkov A.Y. Introduction to the Theory of Randomized Machine Learning // Learning Systems: From Theory to Practice / ed. by Sgurev V., Piuri V., Jotsov V. Cham: Springer International Publishing, 2018. P. 199–220. ISBN: 978-3-319-75181-8. https://doi.org/10.1007/978-3-319-75181-$8_1$0.
13. Popkov Y., Popkov A., Dubnov Y. Elements of Randomized Forecasting and Its Application to Daily Electrical Load Prediction in a Regional Power System // Automation and Remote Control. 2020. Vol. 81, pp. 1286–1306. https://doi.org/10.1134/S0005117920070103.
14. Kermack W.O., McKendrick A.G. Contributions to the Mathematical Theory of Epidemics // Proceedings of the Royal Society. 1927. Vol. 115A. P. 700–721.
15. Müller G.R. Zeitschrift für allgemeine Mikrobiologie / In: The Population Dynamics of Infectious Diseases: Theory and Applications. 368 S., 135 Abb., 104 Tab. London-New York, Chapman and Hall, 1984, Vol. 24, no. 2. pp. 76–76. https://doi.org/10.1002/jobm.19840240203.
16. Hethcote H.W. Three Basic Epidemiological Models // Applied Mathematical Ecology. Springer Berlin Heidelberg, 1989. pp. 119–144. https://doi.org/10.1007/978-3-642-61317-3_5.
17. Peng L., Yang W., Zhang D., Zhuge C., Hong L. Epidemic analysis of COVID-19 in China by dynamical modeling // arXiv, 2020. 10.48550/ARXIV.2002.06563.
18. Yang W., Zhang D., Peng L., Zhuge C., Hong L. Rational evaluation of various epidemic models based on the COVID-19 data of China // Epidemics, 2021. Vol. 37. p. 100501. https://doi.org/10.1016/j.epidem.2021.100501.
19. Cheng C., Zhang D., Dang D., Geng J., Zhu P., Yuan M., Liang R., Yang H., Jin Y., Xie J., Chen S., Duan G. The incubation period of COVID-19: a global meta-analysis of 53 studies and a Chinese observation study of 11 545 patients // Infectious Diseases of Poverty, 2021. Vol. 10, no. 1. https://doi.org/10.1186/s40249-021-00901-9.
20. Huang S., Li J., Dai C., Tie Z., Xu J., Xiong X., Hao X., Wang Z., Lu C. Incubation period of coronavirus disease 2019: New implications for intervention and control // International Journal of Environmental Health Research, 2021. P. 1–9. https://doi.org/10.1080/09603123.2021.1905781.
21. Li Q. et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus — Infected Pneumonia // New England Journal of Medicine, 2020. Vol. 382, no. 13. P. 1199–1207. https://doi.org/10.1056/nejmoa2001316.
22. Nie X. et al. Epidemiological Characteristics and Incubation Period of 7015 Confirmed Cases With Coronavirus Disease 2019 Outside Hubei Province in China // The Journal of Infectious Diseases, 2020. Vol. 222, no. 1. pp. 26–33. https://doi.org/10.1093/infdis/jiaa211.
23. Guidotti E., Ardia D. COVID-19 Data Hub // Journal of Open Source Software. 2020. Vol. 5, no. 51. P. 2376. https://doi.org/10.21105/joss.02376.

24. COVID-19 Data Hub. https://www.covid19datahub.io. 2021. Accessed: 2022-06-20.
25. Flach P. Nauka и iskusstvo postroenia algoritmov, kotorie izvlekaiut znania iz dannykh. Moscow, DMK Press, 2015.
26. Rubinstein R.Y., Kroese D.P. Simulation and the Monte Carlo method. John Wiley & Sons, 2007. Vol. 707.

Popkov Alexey — Ph.D., Leading researcher, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences. Research interests: entropy methods, machine learning, data mining, software development. The number of publications — 48. apopkov@isa.ru; 44-2, Vavilov St., 119133, Moscow, Russia; office phone: +7(499)135-6260.

Dubnov Yuri — Researcher, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences. Research interests: machine learning, bayessian estimation. The number of publications — 32. yury.dubnov@phystech.edu; 44-2, Vavilov St., 119133, Moscow, Russia; office phone: +7(499)135-6260.

Popkov Yuri — Academician of Russian Academy of Sciences, Chief researcher, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences. Research interests: entropy, macrosystems, randomized machine learning, optimization. The number of publications — 224. popkov@isa.ru; 44-2, Vavilov St., 119133, Moscow, Russia; office phone: +7(499)135-6260.

Acknowledgements. This work was supported by Russian Foundation for Basic Research (projects 20-07-00683, 20-04-60119).