

УДК 519.234

DOI 10.34822/1999-7604-2021-4-48-53

НЕПАРАМЕТРИЧЕСКИЙ НЕЙРОСЕТЕВОЙ ГЕНЕРАТОР ПСЕВДОСЛУЧАЙНЫХ ВЕЛИЧИН

О. С. Черепанов ✉, **А. В. Маер**

Курганский государственный университет, Курган, Россия

✉ E-mail: ocherepanov@inbox.ru

Цель работы – построение генератора в нейросетевом базисе и создание алгоритма обучения многослойного перцептрона для решения задачи генерации псевдослучайных величин. По результатам исследования на различных распределениях и объемах выборки с использованием компьютерного моделирования показана высокая эффективность предложенной архитектуры и схемы обучения алгоритма.

Ключевые слова: искусственные нейронные сети, генератор псевдослучайных величин, непараметрический уровень.

NON-PARAMETRIC NEURAL NETWORK GENERATOR OF PSEUDO RANDOM VALUES

O. S. Cherepanov ✉, **A. V. Maer**

Kurgan State University, Kurgan, Russia

✉ E-mail: ocherepanov@inbox.ru

The aim of the study is to build a generator in a neural network basis and to create an algorithm for training a multilayer perceptron to solve the problem of pseudo-random values generation. Based on the results of the study on various distributions and sample sizes using computer modeling, the high efficiency of the proposed architecture and training scheme for the algorithm is shown.

Keywords: artificial neural networks, pseudo-random variable generator, non-parametric level.

Введение

Метод статистических испытаний (МСИ) с момента своего появления получил широкое применение в различных областях науки и техники. Например, на его основе удается оценить надежность сложных систем, гиперпараметры математических моделей в машинном обучении и эффективность алгоритмов. Важную роль в методе статистических испытаний отводят генераторам псевдослучайных величин, которые используются для порождения значений случайных факторов в математических моделях процессов или явлений. В классической постановке МСИ распределение случайных факторов полностью известно или задано с точностью до конечного числа неизвестных параметров, значения которых оцениваются по исходной выборке. Но на практике априорные знания о виде распределений часто отсутствуют. В этом случае мы находимся на непараметрическом уровне априорной информации, на котором нам известны лишь исходные выборки и, возможно, некоторая дополнительная информация о виде распределения, например симметричность, унимодальность, значения некоторых квантилей и т. д. Генераторы на таком уровне будем называть непараметрическими.

Интерес к непараметрическим генераторам случайных величин значительно возрос с появлением бутстреп-методов [1–3], которые позволяют оценивать различные характеристики моделей при ограниченных объемах выборки. В бутстреп-процедурах формируются серии ре-выборок с использованием непараметрических генераторов, построению которых посвящено много работ [4–7]. Основной подход непараметрической генерации случайных величин без повторений – построение сглаженной непараметрической оценки интегральной функции распределения случайного фактора с дальнейшим ее обращением.

Благодаря развитию вычислительной техники и методов глубокого обучения искусственные нейронные сети стали одним из доминирующих инструментов при решении большого круга задач высокой размерности. В данной работе предлагается подход к построению генератора псевдослучайных величин в нейросетевом базисе в виде многослойного перцептрона.

1. Постановка задачи

Пусть X – случайная величина с функцией распределения $F(x)$ (исходная функция распределения). Обозначим $(x_1, x_2, \dots, x_n)$ – исходную выборку случайной величины X . Для построения непараметрического генератора случайных величин из распределения $F(x)$ воспользуемся известным обратным преобразованием случайных величин [6]:

$$X = F^{-1}(U). \quad (1)$$

Как известно, случайная величина U должна быть распределена равномерно на отрезке $[0;1]$, тогда случайная величина X будет иметь распределение $F(x)$.

Так как мы находимся на непараметрическом уровне априорной информации, то согласно выражению (1) для построения генератора необходимо решить две задачи:

1. Построение непараметрической оценки $\tilde{F}_n(x)$ распределения $F(x)$. В качестве оценки $\tilde{F}_n(x)$ могут выступать, например, ядерные оценки вида:

$$\tilde{F}_n(x) = \frac{1}{n} \sum_{i=1}^n G\left(\frac{x-x_i}{h_n}\right), \quad (2)$$

где $G(x)$ – ядерная функция;

h_n – параметр размытости.

Примеры ядерных оценок функций распределения, их свойства и ограничения на ядерные функции и параметр размытости приведены в [8].

2. Обращение полученной оценки интегральной функции распределения. Обратная функция оценки (2) не может быть представлена в явном виде, поэтому нелинейное уравнение (1) решают численными методами.

Постараемся решить задачу построения непараметрического генератора без решения промежуточной задачи оценки интегральной функции распределения. Вместо этого обучим многослойный перцептрон таким образом, чтобы он аппроксимировал сразу обратную оценку функции распределения случайной величины по выборке.

2. Модель непараметрического генератора псевдослучайных величин

Задачу построения непараметрического генератора можно свести к решению задачи непараметрической регрессии (1), когда вид обратной интегральной функции неизвестен. Для оценивания $F^{-1}(U)$ будем использовать многослойный перцептрон, типовая архитектура которого изображена на рис. 1.

В качестве обучающей выборки возьмем список следующих пар:

$$[(F_n(x_1), x_1), (F_n(x_2), x_2), \dots, (F_n(x_n), x_n)],$$

где x_i – i -й элемент исходной выборки;

$F_n(x_i)$ – значение эмпирической функции распределения в точке x_i .

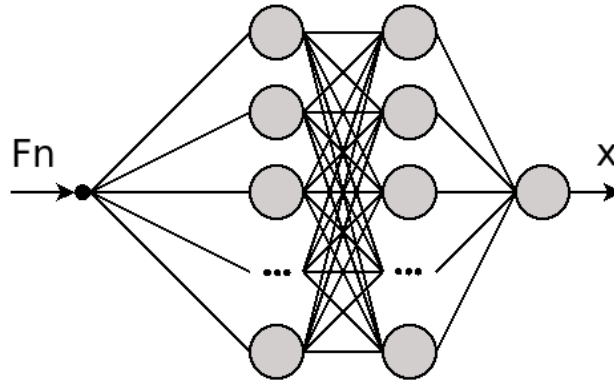


Рис. 1. Архитектура непараметрического нейросетевого генератора псевдослучайных чисел

Примечание: рисунок авторов.

На входной слой перцептрона подаются значения эмпирической функции распределения, а нейрон выходного слоя пытается предсказать ее аргумент, аппроксимируя обратную функцию.

3. Моделирование

Проводилось исследование качества предложенного непараметрического генератора псевдослучайных чисел на следующих распределениях:

1. Нормальное распределение со средним 3 и среднеквадратичным отклонением 0,5.
2. Экспоненциальное распределение с параметром масштаба 0,75.
3. Смесь нормальных распределений. Основное стандартное нормальное распределение с долей 0,9 и дополнительное со средним 8 и масштабом 0,5.

Для каждого распределения исследование проводилось при объеме выборки 50, 100 и 500.

Рассмотрим этапы построения непараметрического генератора и компьютерного моделирования по оценке его эффективности:

1. Получение исходной выборки $X = (x_1, \dots, x_n)$ из заданного распределения $F(x)$ с использованием стандартных генераторов из библиотеки *numpy.random*.
2. Проверка качества исходной выборки X с использованием одновыборочного критерия согласия Колмогорова – Смирнова [9] с уровнем значимости 0,01. Если выборка не подчиняется заданному распределению (нулевая гипотеза отклоняется), то происходит возврат к шагу 1.
3. Для каждого элемента исходной выборки $x_i \in X, i = 1 \dots n$ оцениваем значение эмпирической функции распределения $F_n(x_i), i = 1 \dots n$ и формируем обучающую выборку $(F, X) = [(F_n(x_1), x_1), (F_n(x_2), x_2), \dots, (F_n(x_n), x_n)]$.
4. Обучение многослойного перцептрона с использованием обучающей выборки (F, X) .
5. Генерация выборки $U = (u_1, \dots, u_n)$ из равномерного распределения $[0;1]$ объема n с применением функций библиотеки *numpy.random*.
6. Последовательная подача элементов выборки U на входной слой многослойного перцептрона. На выходе перцептрона формируется ре-выборка $R = (r_1, \dots, r_n)$.
7. Проверка нулевой гипотезы «Ре-выборка принадлежит исходному распределению с указанными параметрами» с использованием одновыборочного критерия согласия Колмогорова – Смирнова.
8. Проверка нулевой гипотезы «Исходная выборка X и ре-выборка R подчиняются одному распределению» с использованием двухвыборочного критерия согласия Колмогорова – Смирнова [9].
9. Повторение шагов 1–8 $M = 100$ раз с подсчетом значений соотношений:

$$v_1 = \frac{k_1}{M}, \quad (3)$$

$$v_2 = \frac{k_2}{M}, \quad (4)$$

где k_1 – количество неотвергнутых нулевых гипотез при оценивании качества ре-выборок на шаге 7;

k_2 – количество неотвергнутых нулевых гипотез при оценивании качества ре-выборок на шаге 8.

Соотношения (3), (4) показывают, какая доля M ре-выборок подчиняется исходному распределению.

При моделировании многослойный перцептрон состоял из двух скрытых слоев с 64 нейронами. Нейроны в скрытых слоях имели функцию активации ReLU. Обучение модели осуществлялось с использованием стохастического градиентного спуска с уменьшающейся скоростью обучения. В качестве критерия минимизации использовалась функция среднеквадратической ошибки. Для предотвращения переобучения использовалась L_2 -регуляризация и слои Dropout с вероятностью 0,3.

Обучение сети проводилось с использованием алгоритма обратного распространения ошибки [10]. При моделировании использовалась реализация многослойного перцептрона и алгоритма обучения библиотеки TensorFlow [11].

На рис. 2–4 представлены ядерные оценки плотности вероятности по исходной выборке и ре-выборке, а в таблице – значения соотношения (3) и (4) для разных распределений и объемов выборки.

Оценка плотности вероятности определена в следующем виде:

$$\tilde{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x-x_i}{h_n}\right), \quad (5)$$

где $K(x)$ – ядерная функция;

h_n – параметр размытости.

Выбор ядерной функции $K(x)$ и параметра размытости h_n исследован в рамках непараметрической статистики (см., например, [8, 12]). В данном исследовании в качестве $K(x)$ выбрана нормальная ядерная функция, параметр h_n оценивался путем минимизации энтропии [8].

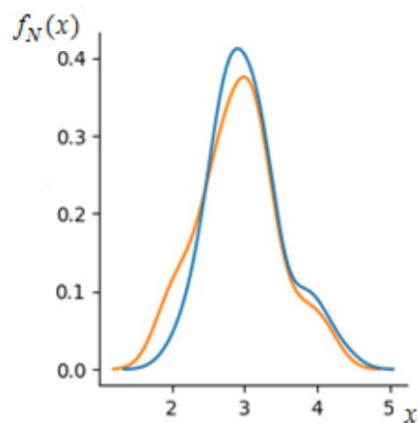
Не сложно предположить, что качество предложенного генератора случайных чисел будет зависеть от объема исходной выборки (с увеличением исходной выборки расстояние $\tilde{f}_n(x)$ для исходной и ре-выборки будет уменьшаться). Период повторения комбинации значений в ре-выборке будет зависеть от качества генератора случайных чисел из равномерного распределения.

Таблица

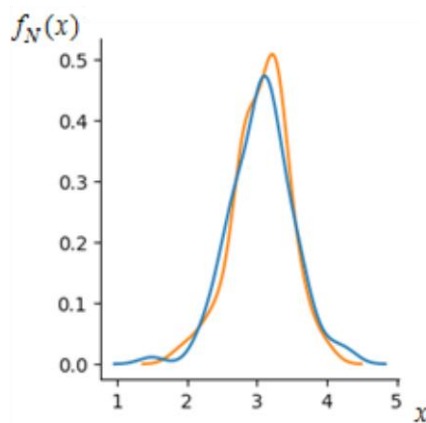
Значения соотношения (2) при оценке качества непараметрического нейросетевого генератора

Объем выборки	50	100	500
Нормальное распределение			
v_1	0,91	0,91	0,92
v_2	1,00	1,00	1,00
Экспоненциальное распределение			
v_1	0,90	0,92	0,90
v_2	0,99	0,99	1,00
Смесь нормальных распределений			
v_1	0,92	0,90	0,91
v_2	1,00	1,00	1,00

Примечание: составлено авторами на основе результатов компьютерного моделирования.

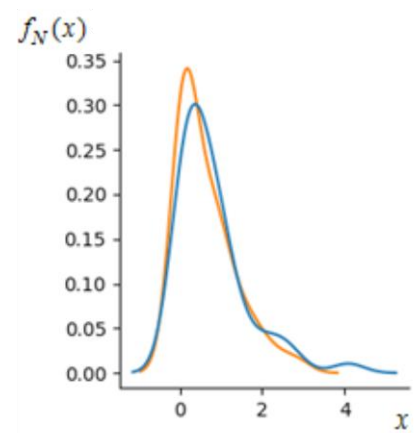


Объем выборки 50

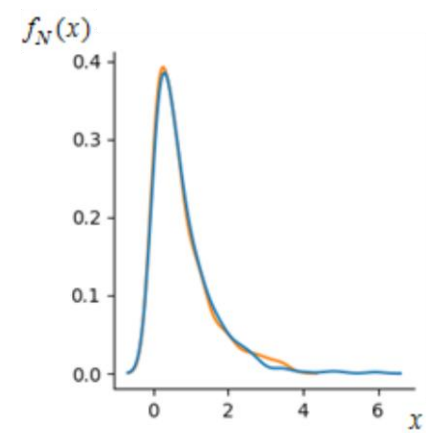


Объем выборки 500

Рис. 2. Графики ядерных оценок плотности вероятности по выборкам из нормального распределения: оранжевый – исходная выборка, синий – ре-выборка
Примечание: рисунок авторов.

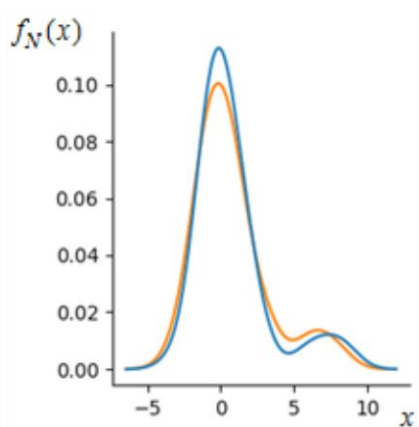


Объем выборки 50

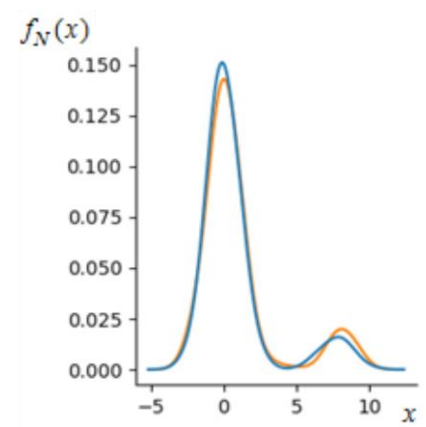


Объем выборки 500

Рис. 3. Графики ядерных оценок плотности вероятности по выборкам из экспоненциального распределения: оранжевый – исходная выборка, синий – ре-выборка
Примечание: рисунок авторов.



Объем выборки 50



Объем выборки 500

Рис. 4. Графики ядерных оценок плотности вероятности по выборкам из смеси нормальных распределений: оранжевый – исходная выборка, синий – ре-выборка
Примечание: рисунок авторов.

Заключение

По результатам проведенной работы можно сделать следующие выводы:

1. Удалось предложить непараметрический генератор псевдослучайных чисел на основе многослойного перцептрона.
2. Предложенный подход позволяет решить поставленную задачу генерации без решения промежуточной задачи оценивания интегральной функции распределения случайной величины.
3. Таблица показывает высокое качество предложенного генератора (меньше 10 % ошибок) на разных распределениях и объемах выборок.

Литература

1. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. М. : Финансы и статистика, 1998. 263 с.
2. Davison A. C., Hinkley D. V. Bootstrap Methods and Their Application. Cambridge University Press, 1997.
3. Bickel P., Freedman D. Some Asymptotic Theory for the Bootstrap // Ann Statist. 1981. Vol. 9, Is. 6. P. 1196–1217.
4. Tereshchenko E. P., Teskin O. I., Simakhin V. A. Bootstrap Models for Interval Estimation of Longevity Characteristics of Sequential Systems From Small // J Math Sci. 1996. Vol. 81, No. 4. P. 2811–2817.
5. Маер А. В., Симахин В. А. Непараметрические датчики случайных векторов // Оптика атмосферы и океана. Физика атмосферы : сб. материалов XVI Междунар. симпозиума. Иркутск, 12–15 октября 2009 г. Томск, 2009. С. 403–407
6. Devroye L. Non-uniform Random Variate Generation. N. Y. : Springer-Verlag, 1986. 843 p.
7. Батраков П. А., Маер А. В., Симахин В. А. Одномерный непараметрический датчик с учетом априорной информации // Омский науч. вестник. 2012. № 2. С. 42–47.
8. Васильев В. А., Добровидов А. В., Кошкин Г. М. Непараметрическое оценивание функционалов от распределений стационарных последовательностей. М. : Наука, 2004. 510 с.
9. Лемешко Б. Ю., Лемешко С. Б., Чимитова С. Н. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход. Новосибирск : Изд-во НГТУ, 2011. 888 с.
10. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. М. : ДМК Пресс, 2018. 652 с.
11. Джулли А. Библиотека Keras – инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и Tensorflow. М. : ДМК Пресс, 2018. 294 с.
12. Parzen E. On Estimation of Probability Density Function and Mode // Ann Math Statist. 1962. Vol. 33, Is. 3. P. 1065–1076.