ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2023 Управление, вычислительная техника и информатика № 62

Tomsk State University Journal of Control and Computer Science

Научная статья
УДК 519.688
doi: 10.17223/19988605/62/5

# Бинарное прогнозирование динамических показателей на основе методов машинного обучения

**Юрий Мечеславович Краковский[1], Ольга Константиновна Куклина[2]**

*[1] Иркутский государственный университет путей сообщения, Иркутск, Россия, 79149267772@yandex.ru*
*[2] Читинский институт (филиал) Байкальского государственного университета, Чита, Россия, kuklinaok@bgu-chita.ru*

**Аннотация.** Рассмотрена задача бинарного прогнозирования динамических показателей на основе машинного обучения с приложением к задаче перевозки грузов железнодорожным транспортом. В качестве методов выбраны вероятностная нейронная сеть и логистическая регрессия. Бинарное прогнозирование заключается в оценке прогнозных значений показателя на основе вероятностей принадлежности одному из двух интервалов. Так как при такой процедуре определяется не само будущее значение показателя, а то, в каком интервале оно будет находиться, такое прогнозирование называют бинарным, или интервальным. Программное обеспечение разработано на языке программирования Python с применением сторонних библиотек с открытым исходным кодом. Тестирование созданного программно-алгоритмического обеспечения по реальным исходным данным перевозочного процесса показало высокую точность бинарного прогнозирования и на основе вероятностной нейронной сети, и на основе логистической регрессии.

**Ключевые слова:** бинарное прогнозирование; вероятностная нейронная сеть; логистическая регрессия; динамические показатели.

***Для цитирования*:** Краковский Ю.М., Куклина О.К. Бинарное прогнозирование динамических показателей на основе методов машинного обучения // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2023. № 62. С. 50–55. doi: 10.17223/19988605/62/5

Original article
doi: 10.17223/19988605/62/5

# The binary forecasting of dynamic indicators based on machine learning methods

**Yuri M. Krakovsky[1], Olga K. Kuklina[2]**

*[1] Irkutsk State Transport University, Irkutsk, Russian Federation, 79149267772@yandex.ru*
*[2] Chita Institute of Baikal State University, Chita, Russian Federation, kuklinaok@bgu-chita.ru*

**Abstract.** The problem of binary forecasting of dynamic indicators based on machine learning methods in relation to the problem of cargo transportation by railway transport is considered. The probabilistic neural network and logistic regression were chosen as the methods. The binary forecasting consists on evaluating predictive values of the indicator which is based on the belonging probabilities to one of two intervals. The forecasting is called binary or interval as on this process is calculated interval for the indicator value where it will be, not the predicted value of the indicator. The software is developed using the Python programming language with open source libraries. The software and algorithm test were done on the examples of real values of railway transportation process and shown its high accuracy of binary forecasting both on the probabilistic neural network and logistic regression methods.

**Keywords:** binary forecasting; probabilistic neural network; logistic regression; dynamic indicators.

***For citation*:** Krakovsky, Y.M., Kuklina, O.K. (2023) The binary forecasting of dynamic indicators based on machine learning methods. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 62. pp. 50–55. doi: 10.17223/19988605/62/5

## Введение

Как правило, динамические процессы проявляются в виде ряда последовательно расположенных в хронологическом порядке значений того или иного показателя, характеризующего динамику развития изучаемого явления. В связи с этим важным направлением математического моделирования является прогнозирование показателей динамических процессов. Данному направлению посвящено значительное число работ, но среди них преобладают работы, основанные на регрессионных или авторегрессионных моделях. Для увеличения точности прогноза в условиях неопределенности в последние годы происходит совершенствоввование методов прогнозирования [1–3], и здесь следует отметить широкое применение ансамблей и методов на основе машинного обучения [4, 5].

Методы прогнозирования показателей можно разделить на две группы: а) точечные, когда определяется будущее значение; б) вероятностные, когда определяется не само значение, а вероятности того, в каком интервале это значение находится. В последнее время наблюдается возрастающий интерес пользователей именно к вероятностным методам прогнозирования [6, 7]. Это можно объяснить тем, что на практике часто не требуется знания самого значения показателя, а необходимо лишь знание о том в каком интервале оно окажется и с какой вероятностью.

Среди методов вероятностного прогнозирования, включая бинарное прогнозирование, можно выделить следующие: вероятностные кластерные методы, логистической регрессии, на основе вероятностных нейронных сетей и др. [8, 9].

В данной работе рассматривается задача прогнозирования динамических показателей, характеризующих перевозочный процесс грузов железнодорожным транспортом. Для России перевозка грузов по железной дороге существенно влияет на ее экономику. Это особенно важно для восточного полигона железнодорожной сети ОАО «РЖД», расположенного в границах четырех дорог: Красноярской, Восточносибирской, Забайкальской и Дальневосточной. Особая роль среди этих дорог принадлежит Дальневосточной, так как она является конечной и взаимодействует через морские порты и погранпереходы с другими странами. В связи с этим для повышения эффективности управления перевозкой грузов прогнозирование динамических показателей перевозочного процесса для этой дороги имеет существенное значение. В условиях неопределенности перевозочного процесса такая задача является актуальной и нетривиальной.

Цель работы – реализация и апробация методов на основе машинного обучения для бинарного прогнозирования динамических показателей с приложением к задаче перевозки грузов железнодорожным транспортом. В качестве методов выбраны вероятностная нейронная сеть (ВНС) и логистическая регрессия (ЛР).

Пусть известен временной ряд некоторого показателя $Q=\{q_t : t\in T\}$. Здесь $q_t$ – значения показателя, заданные в моменты времени *t*; время *t* принимает значения из множества $T = 0,...,n-1$ ; *n* – количество значений показателя; все значения показателя $q_t > 0$. Обозначим интервал возможных значений показателя в будущем $(c_1;c_2)$, где $c_1 > 0$, $c_2 < \infty$; введем внутреннюю точку *c*: $c_1 < c < c_2$. Это позволяет разделить этот интервал на два других^

$$I_a = (c_1;c], \quad I_b = (c;c_2). \tag{1}$$

Тогда бинарное (интервальное) прогнозирование заключается в оценке прогнозных значений показателя на основе вероятностей принадлежности одному из двух интервалов (1). Так как при такой процедуре определяется не само будущее значение показателя, а то, в каком интервале оно будет находиться, такое прогнозирование названо бинарным? или интервальным [1].

Значение внутренней точки *c* можно определять различными способами. В данной работе она определяется так^

$$c = q_{n-1} + \Delta, \quad \Delta = \alpha \cdot \left( \sum_{t=1}^{n-1} |q_t - q_{t-1}| \right) / (n-1), \tag{2}$$

где $\alpha \in [-1;\ 1]$ – коэффициент, который задается в исходных данных.

В момент времени $t = n-1$ необходимо определить, в каком из интервалов (1) будет находиться будущее (неизвестное) значение $q_{t+p}$ на основе оценок вероятностей $\rho^a_{t+p}$ и $\rho^b_{t+p}$, где $p = 1,...,r$ – время упреждения; $\rho^a_{t+p}$ – вероятность того, что $q_{t+p} \in I^a$; $\rho^b_{t+p}$ – вероятность того, что $q_{t+p} \in I^b$; $\rho^a_{t+p} + \rho^b_{t+p} = 1$.

Бинарное прогнозирование проводится по правилу:

будущее значение $q_{t+p} \in I^a$, если $\rho^a_{t+p} \geq \rho^b_{t+p}$; будущее значение $q_{t+p} \in I^b$, если $\rho^b_{t+p} > \rho^a_{t+p}$. (3).

## 1. Математическое описание задачи

### *1.1. Случай вероятностной нейронной сети*

Архитектура ВНМ была предложена в 1988 г. Д. Спехтом [10] для проведения классификации векторов (образов) с неизвестной классификацией. Перед проведением такой классификации ВНМ должна быть обучена на множестве векторов с известной классификацией. Достоинством ВНМ является простота построения и высокая скорость обучения.

Под вектором ВНМ понимается набор значений $V = v_j, j \in J$, где $j$ принимает значения из множества $J = 0,...,m-1$, при этом $m > 0$. Часть векторов ($u$) – обучающие, другая часть – для тестирования. Для всех векторов ВНМ должно выполняться условие

$$\sum_{j \in J} v_j^2 = 1. \quad (4)$$

Условие (4) обеспечивается специальной обработкой реальных значений.

ВНС включают четыре слоя: входной слой, в котором количество нейронов определяется количеством введенных исходных значений (размерность векторов $m$); скрытый слой, в котором каждый нейрон имеет $m$ входов; суммирующий слой, где определяются вероятности принадлежности входного вектора к тому или иному классу; выходной слой, на котором проводится сравнение вероятностей и формулируется результат.

На этапе обучения создается матрица весов: $w_{z,j} = v_{z,j}$, где $z$ принимает значения из множества $Z = 0,...,u-1$, $u = n^A + n^B$, где $n^A$ – число векторов при обучении из класса $A$ (прогнозное значение исследуемого показателя попадает в интервал $I_a$); $n^B$ – число векторов при обучении из класса $B$, когда прогнозное значение показателя попадает в интервал $I_b$.

Рассмотрим алгоритм классификации векторов с неизвестной классификацией посредством ВНМ, а также формализуем вид активационной функции нейронов скрытого слоя.

Пусть имеется произвольный вектор $V_h = \{v_{h,j}, j \in J\}$ с неизвестной классификацией, для которого также выполняется условие (4). Этот вектор подается на входы входного слоя ВНМ.

Далее в скрытом слое вычисляется $u$ значений с помощью нелинейной активационной функции

$$H_{h,z} = \exp\left(\frac{-\sum_{j \in J}(v_{h,j} - w_{z,j})^2}{2\sigma^2}\right). \quad (5)$$

Здесь $\sigma$ – параметр ВНМ, который задается в исходных данных.

С учетом (5) нейроны суммирующего слоя реализуют вычисления

$$l_h^A = \frac{\sum_{z \in M^A} H_{h,z}}{n^A}, \quad l_h^B = \frac{\sum_{z \in M^B} H_{h,z}}{n^B}, \quad (6)$$

где $M^A$ – множество номеров нейронов скрытого слоя, созданных для векторов класса $A$; $M^B$ – множество номеров нейронов скрытого слоя, созданных для векторов класса $B$. Эти множества определяются на этапе обучения.

Выходной нейрон ВНМ вычисляет бинарное значение

$$\Omega_h = \begin{cases} 1, & \tilde{\rho}^b_{t+p} > \tilde{\rho}^a_{t+p}, \\ 0, & \tilde{\rho}^a_{t+p} \geq \tilde{\rho}^b_{t+p}. \end{cases} \tag{7}$$

В функции (7) приведены оценки вероятностей $\rho^a_{t+p}$ и $\rho^b_{t+p}$:

$$\tilde{\rho}^a_{t+p} = l^A_h \big/ \left(l^A_h + l^B_h\right),\ \tilde{\rho}^b_{t+p} = l^B_h \big/ \left(l^A_h + l^B_h\right). \tag{8}$$

Само бинарное прогнозирование осуществляется по правилу (3): значение 1 в (7) означает, что будущее значение показателя попадет в интервал $I_b$, а при значении 0 – в интервал $I_a$. Оценки вероятностей определяются по формуле (8) с учетом (6).

### *1.2. Случай логистической регрессии*

Введем: а) бинарное значение

$$y_{t+p} = \begin{cases} 1, & q_{t+p} > c, \\ 0, & q_{t+p} \leq c, \end{cases} \tag{9}$$

где $c$ – внутренняя точка (2); б) линейную регрессионную функцию

$$s_t = a_0 + \sum_{i=1}^{f} a_i \cdot q_{t-i+1}, \tag{10}$$

где $a_0, ..., a_f$ – коэффициенты; $f$ – число регрессоров. Следуя рекомендациям, будем предполагать, что оценка вероятности первого события

$$\tilde{\rho}^b_{t+p}\ \ y_{t+p} = 1 | s_t\ = \sigma\ s_t\ , \tag{11}$$

где $\sigma\ s_t\ = 1/(1+e^{-s_t})$ – логистическая (сигмоидальная) функция; величина $s_t$ определяется согласно выражению (10). Так как $y_{t+p}$ принимает только два возможных значения (9), то оценка вероятности наступления второго события равна

$$\tilde{\rho}^a_{t+p}\ \ y_{t+p} = 0 | s_t\ = 1 - \sigma\ s_t\ . \tag{12}$$

При бинарном прогнозировании искомые вероятности заменяются оценками (11) и (12). Само бинарное прогнозирование осуществляется по правилу (3): значение 1 в (9) означает, что будущее значение показателя попадет в интервал $I_b$, а при значении 0 – в интервал $I_a$.

## 2. Программное обеспечение и обсуждение результатов

Программное обеспечение бинарного прогнозирования разработано на языке программирования Python с применением сторонних библиотек с открытым исходным кодом. Для разработки выбрана среда PyCharm – интеллектуальная Python IDE с полным набором средств для эффективной разработки на языке Python. Пользователю предлагается выбрать подготовленные исходные данные; способ прогнозирования показателей – на основе вероятностной нейронной модели; на основе логистической регрессии (вкладка Load data). На вкладке Plot пользователю доступны инструменты визуализации данных. Результаты интервального прогнозирования доступны на вкладке Result.

Например, построение ВНС выполнено с использованием библиотеки neupy и sklearn; для обработки и анализа данных – библиотека pandas. Для визуализации машинного обучения предложено использовать yellowbrick.classifier и библиотеку интерактивной визуализации данных cufflinks. По-

строение ВНС выполнено с использованием модуля neupy.algorithms.rbfn.pnn (библиотека для искусственных нейронных сетей и глубокого обучения).

Для визуализации машинного обучения предложено использовать платформу Yellowbrick – проект Python с открытым исходным кодом, который объединяет API-интерфейсы scikit-learn и matplotlib.

На рис. 1 приведены вкладки разработанного программного обеспечения для бинарного прогнозирования.

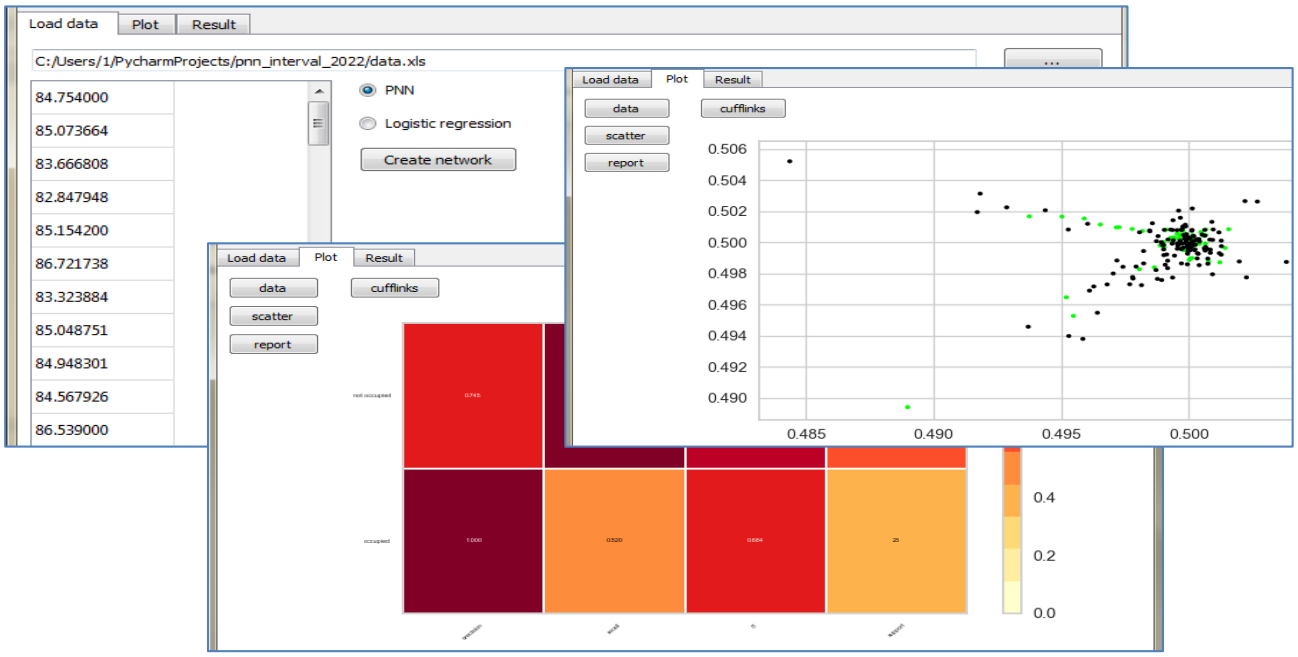


Рис. 1. Вкладки программы для бинарного прогнозирования
Fig. 1. Program tabs of binary forecasting

В результате тестирования созданного программно-алгоритмического обеспечения бинарного прогнозирования базового показателя перевозочного процесса в виде грузооборота на основе ВНС показано, что в 93,3% случаев получены правильные результаты. Применительно к ЛР правильные результаты оказались еще выше и составили 95,1%.

## Заключение

Предложено и апробировано программно-алгоритмическое обеспечение бинарного прогнозирования динамических показателей с приложением к задаче перевозки грузов железнодорожным транспортом. В качестве методов машинного обучения выбраны вероятностная нейронная сеть и логистическая регрессия. Тестирование созданного обеспечения по реальным исходным данным перевозочного процесса в виде грузооборота показало его высокую точность и на основе вероятностной нейронной сети, и на основе логистической регрессии.

### Список источников

1. Shumway R.H. Time series analysis and its applications with R examples. Springer, 2011. 609 p.
2. Mitrea C.A. A Comparison between neural networks and traditional forecasting methods: a case study // International Journal of Engineering Business Management. 2009. V. 1 (2). P. 19–24.
3. Vernay M., Lafaysse M., Merindol L. Ensemble forecasting of snowpack conditions and avalanche hazard // Cold Regions Science and Technology. 2015. V. 120. P. 251–262.
4. Краковский Ю.М., Курчинский Ю.В., Лузгин А.Н. Интервальное прогнозирование интенсивности кибератак на объекты критической информационной инфраструктуры // Доклады ТУСУР. 2018. Т. 21, № 1. С. 71–79.

5. Wang H., Li G., Wang H. Deep learning based ensemble approach for probabilistic wind power forecasting // Applied Energy. 2017. V. 188. P. 56–70.
6. Yoder M., Cering A.S., Navidi W.C. Short-term forecasting of categorical changes in wind power with Markov chain models // Wind Energy. 2014. V. 17. P. 1425–1439.
7. Krakovsky Y., Luzgin A. Robust interval forecasting algorithm based on a probabilistic cluster model // Journal of Statistical Computation and Simulation. 2018. V. 88 (12). P. 2309–2324.
8. Ivanyo Y.M., Krakovsky Y.M., Luzgin A.N. Interval forecasting of cyber-attacks on industrial control systems // IOP Conference Series: Materials Science and Engineering (Simulation and Automation of Production Engineering). 2018. V. 327. Art. 022044.
9. Munkhdorj B., Yuji S. Cyber attack prediction using social data analysis // Journal of High Speed Networks. 2017. V. 23 (2). P. 109–135.
10. Spetch D.F. Probabilistic Neural networks // Neural Networks. 1990. V. 3. P. 109–118.

**References**

1. Shumway, R.H. (2011) *Time Series Analysis and Its Applications with R Examples.* Springer.
2. Mitrea, C.A. (2009) A Comparison between neural networks and traditional forecasting methods: a case study. *International Journal of Engineering Business Management*. 1(2). pp. 19–24.
3. Vernay, M., Lafaysse, M. & Merindol, L. (2015) Ensemble forecasting of snowpack conditions and avalanche hazard. *Cold Regions Science and Technology*. 120. pp. 251–262.
4. Krakovsky, Y.M., Kurchinsky, B.V. & Luzgin, A.N. (2018) Cyber-attack intensity interval forecasting on objects of critical information infrastructure. *Doklady TUSUR – Proceedings of the TUSUR University*. 21(1). pp. 71–79.
5. Wang, H., Li, G. & Wang, H. (2017) Deep learning based ensemble approach for probabilistic wind power forecasting. *Applied Energy*. 188. pp. 56–70.
6. Yoder, M., Cering, A.S. & Navidi, W.C. (2014) Short-term forecasting of categorical changes in wind power with Markov chain models. *Wind Energy*. 17. pp. 1425–1439.
7. Krakovsky, Y. & Luzgin, A. (2018) Robust interval forecasting algorithm based on a probabilistic cluster model. *Journal of Statistical Computation and Simulation*. 88(12). pp. 2309–2324. DOI: 10.1080/00949655.2018.1462809
8. Ivanyo, Y.M., Krakovsky, Y.M. & Luzgin, A.N. (2018) Interval forecasting of cyber-attacks on industrial control systems. *IOP Conference Series: Materials Science and Engineering (Simulation and automation of production engineering)*. 327. Art. 022044. DOI:10.1088/1757-899X/327/2/022044
9. Munkhdorj, B. & Yuji, S. (2017) Cyber attack prediction using social data analysis. *Journal of High Speed Networks*. 23(2). pp. 109–135.
10. Spetch, D.F. (1990) Probabilistic Neural networks. *Neural Networks*. 3. pp. 109–118.

***Информация об авторах*:**

**Краковский Юрий Мечеславович** – профессор, доктор технических наук, профессор кафедры информационных систем и защиты информации Иркутского государственного университета путей сообщения (Иркутск, Россия). E-mail: 79149267772@yandex.ru

**Куклина Ольга Константиновна** – старший преподаватель кафедры информационных технологий и высшей математики Читинского института (филиала) Байкальского государственного университета (Чита, Россия). E-mail: kuklinaok@bgu-chita.ru

***Вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.***

***Information about the authors:***

**Krakovsky Yuri Mecheslavovich** (Professor, Doctor of Engineering sciences, Professor of the Information Systems and Information Security Department, Irkutsk State University of Railway Transport, Irkutsk, Russian Federation). E-mail: 79149267772@yandex.ru

**Kuklina Olga Konstantinovna** (Senior Professor of the Information Technology and Higher Mathematics Department, Chita Institute of Baikal State University, Chita, Russian Federation). E-mail: kuklinaok@bgu-chita.ru

***Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.***

*Поступила в редакцию 16.08.2022; принята к публикации 01.03.2023*

*Received 16.08.2022; accepted for publication 01.03.2023*