Использование нейросетевого подхода в задаче оценки характеристик систем массового обслуживания

П.В. Антонова[1], Р.М. Гиздатуллин[2], М.В. Казаков[2], А.С. Титовцев[2]

[1]Национальный банк Республики Татарстан Волго-Вятского ГУ Банка России
[2]Казанский национальный исследовательский технологический университет

Аннотация: В данной работе предложен способ оценки ключевых показателей многоканальной системы массового обслуживания с неограниченной очередьюи многофазным обслуживанием эрланговского типа. Показано, что переход к многоканальному случаю приводит к резкому увеличению размерности пространства состояний и усложнению системы уравнений Колмогорова, из-за чего прямой аналитический расчёт часто становится недоступен. Предлагается метамодель на основе методов машинного обучения, обучаемая на данных дискретно-событийного имитационного моделирования, для приближённого прогноза среднего времени ожидания, средней длины очереди и доли обслуженных заявок. Выполнено сравнение базовых регрессионных и нейросетевых моделей и рассмотрена устойчивость аппроксимации при изменении коэффициента загрузки.
Ключевые слова: система массового обслуживания, очередь, имитационное моделирование, метамодель, машинное обучение, нейронная сеть, многоканальное обслуживание, эрланговское распределение, нетерпеливая заявка, уравнение Колмогорова, регрессия, градиентный бустинг, случайный лес, перцептрон.

Введение

В условиях наступившей четвертой промышленной революции, предполагающей новый подход к производству, важнейшими становятся задачи, связанные с повышением показателей эффективности различного рода технологических процессов и систем. Исследования в направлениях теории массового обслуживания не только улучшают существующие технологии и процессы, но и открывают новые возможности для будущих исследований, способствуя развитию как научных, так и практических дисциплин [1,2].

В последние годы наблюдается рост интереса к использованию методов машинного обучения в задачах теории массового обслуживания (ТМО). Это связано с тем, что для немарковских и многоканальных систем классические аналитические методы сталкиваются с ограничениями: рост размерности пространства состояний, вычислительная неустойчивость

матричных процедур и высокая стоимость имитационного моделирования. В этих условиях нейросетевые модели рассматриваются как перспективный инструмент для приближенной оценки характеристик систем массового обслуживания (СМО).

Ранние результаты применения нейросетевых методов в задачах моделирования и анализа СМО представлены в работах [3–5]. Практическое применение нейросетевых моделей для прогнозирования поведения очередей также рассматривается в работе [6], где искусственные нейронные сети используются для предсказания характеристик многоканальных систем типа M/M/n в классификации Кендалла. В последние годы также появились работы, направленные на разработку универсальных глубоких нейросетевых моделей для анализа широкого класса очередей [7]. Таким образом, современная литература подтверждает, что применение методов искусственного интеллекта в теории массового обслуживания является развивающимся направлением. Это особенно актуально для систем смешанного типа с многофазным обслуживанием, для которых аналитические подходы не дают замкнутного решения.

Одноканальная версия рассматриваемой системы была подробно исследована ранее в работе [8]. Однако переход к многоканальному случаю приводит к существенному усложнению динамики очереди и росту размерности пространства состояний, что делает невозможным прямое использование результатов, полученных для одноканального случая.

Формализация модели для многоканальной системы массового обслуживания

Рассмотрим систему массового обслуживания с с параллельными каналами обслуживания ($c \in N$). Входной поток — пуассоновский с интенсивностью λ. Время обслуживания имеет эрланговское распределение порядка r, которое можно представить в виде суммы r независимых

экспоненциальных фаз с параметром μ. Буфер (очередь) неограниченной ёмкости. Каждая заявка в очереди, ожидающая начала обслуживания, покидает систему через независимое экспоненциально распределённое время с параметром θ. Предполагается, что уход из обслуживания отсутствует. Далее используется фазовое представление, которое позволяет обеспечивать марковость рассматриваемого процесса [8].

В многоканальном случае одной координаты «оставшихся фаз» недостаточно, поскольку одновременно обслуживаются до с заявок. Введём:

- $q(t)$— число заявок в очереди в момент времени t;
- $n_k(t)$— число заявок, находящихся в обслуживании на фазе $k, k = 1, \dots, r$.

Определим состояние системы:

$$X(t) = \left(q(t), n_1(t), \dots, n_r(t)\right),$$

Где $q(t) \in Z_{\geq 0}, n_k(t) \in Z_{\geq 0}$. и выполняется ограничение по числу занятых каналов

$$m(t) = \sum_{k=1}^{r} n_k(t) \leq c.$$

Обозначим $\boldsymbol{m} = \sum_{k=1}^{r} \boldsymbol{n_k}$ — текущее число занятых каналов.

1. Поступление заявки (интенсивность λ)

- Если $m < c$ (есть свободный канал), заявка немедленно начинает обслуживание с 1-й фазы:

$$(q, n_1, \dots, n_r) \to (q, n_1 + 1, n_2, \dots, n_r);$$

- Если $m = c$ (все каналы заняты), заявка поступает в очередь:

$$(q, n_1, \dots, n_r) \to (q + 1, n_1, \dots, n_r).$$

2. Завершение фазы обслуживания $k \to k+1$(интенсивность μ)

Для $k = 1, \dots, r-1$ каждая из n_k заявок завершает фазу с интенсивностью μ поэтому суммарная интенсивность перехода:

$$(q, \dots, n_k, \dots, n_{k+1}, \dots) \to (q, \dots, n_k - 1, \dots, n_{k+1} + 1, \dots), \text{интенсивность } \mu n_k.$$

3. Завершение последней фазы r и немедленный «подхват» из очереди

Суммарная интенсивность завершения последней фазы равна μn_r. Далее:

- Если $q \geq 1$, освободившийся канал немедленно занимает одна заявка из очереди, стартуя с фазы 1 с интенсивностью μn_r :

$$(q, n_1, \dots, n_r) \to (q-1, n_1+1, n_2, \dots, n_r - 1);$$

- Если $q = 0$, канал становится свободным:

$$(0, n_1, \dots, n_r) \to (0, n_1, \dots, n_r - 1).$$

4. Уход из очереди (интенсивность θ_q)

Так как каждая из q заявок уходит с интенсивностью θ, суммарно:

$$(q, n_1, \dots, n_r) \to (q-1, n_1, \dots, n_r), q \geq 1.$$

Пусть

$$p_{q,n}(t) = P\{X(t) = (q, n_1, \dots, n_r)\}, \quad n = (n_1, \dots, n_r).$$

Тогда вектор вероятностей $p(t)$ удовлетворяет системе прямых уравнений Колмогорова:

$$\dot{p}(t) = p(t)Q,$$

где Q – матрица, формируемая из интенсивностей переходов, перечисленных выше.

Введем $m = \sum_{k=1}^{r} \quad n_k \leq c$. Тогда система уравнений Колмогорова будет иметь вид:

$$\frac{d}{dt}p_{q,n}(t) = \lambda 1_{\{m=c,q\geq 1\}}p_{q-1,n}(t) + \lambda 1_{\{m\leq c-1,n_1\geq 1\}}p_{q,n-e_1}(t)$$
$$+\sum_{k=1}^{r-1} \quad \mu(n_k+1)1_{\{n_{k+1}\geq 1\}}p_{q,n+e_k-e_{k+1}}(t)$$
$$+\mu(n_r+1)1_{\{n_1\geq 1\}}p_{q+1,n-e_1+e_r}(t)$$

$$+\mu(n_r+1)1_{\{q=0\}}p_{0,n+e_r}(t) + \theta(q+1)p_{q+1,n}(t) -$$
$$(\lambda+\theta q+\mu\sum_{k=1}^{r} \quad n_k)p_{q,n}(t)$$

где e_k — единичный вектор по компоненте n_k. При этом пространство состояний системы будет иметь вид:

$$S = \{(q,n): q \geq 0, n_k \geq 0, \sum \quad n_k \leq c, n\left(q > 0 \Rightarrow \sum \quad n_k = c\right)\}.$$

Построение метамодели

Входные параметры

В качестве исходных входов используются параметры СМО:

$$\lambda, \mu, r, c, \theta$$

где λ — интенсивность входного потока, μ — интенсивность экспоненциальной фазы обслуживания, r — число фаз (порядок распределения Эрланга), c — число параллельных каналов обслуживания, θ — интенсивность ухода заявок из очереди (экспоненциальная «терпеливость»).

Практика построения регрессионных и нейросетевых аппроксиматоров показывает, что использование безразмерных комбинаций параметров улучшает устойчивость обучения и обобщающую способность модели за счёт нормализации масштабов. Поэтому наряду с исходными параметрами вычисляются следующие безразмерные характеристики [9-11]:

Относительная нагрузка (коэффициент загрузки):

$$\rho = \frac{\lambda}{c\mu/r} = \frac{\lambda r}{c\mu}$$

Где $c\mu/r$ соответствует суммарной средней производительности системы (с учётом среднего времени обслуживания $E[S] = r/\mu$);

Относительная интенсивность ухода:

$$\alpha = \frac{\theta}{\mu}$$

В экспериментах метамодель обучается на расширенном векторе признаков

$$x = (\lambda, \mu, \theta, r, c, \rho, \alpha)$$

При этом (ρ, α) используются как основные «режимные» признаки, а (r, c) – как структурные параметры системы

Выходные характеристики

Метамодель аппроксимирует следующие показатели:

1. среднее время ожидания в очереди для обслуженных заявок;
2. средняя длина очереди;
3. доля обслуженных заявок, равная отношению числа заявок, начавших обслуживание и завершивших его, к числу поступивших заявок за период наблюдения (в условиях отсутствия потерь по буферу).

Для получения целевых значений используется дискретно-событийная имитационная модель, реализованная в среде Python с применением библиотеки SimPy [12]. Для каждого набора параметров (λ,μ,r,c,θ) проводится серия независимых прогонов с различными генераторами случайных чисел. Оценивание метрик производится на стационарном участке.

Модели метамоделирования и протокол сравнения

Поскольку задача аппроксимации носит регрессионный характер и включает как непрерывные, так и дискретные признаки , рассматриваются несколько семейств моделей: (i) базовая линейная регрессия/ридж-регрессия как простой эталон, (ii) ансамблевые методы (градиентный бустинг/случайный лес) как устойчивые нелинейные аппроксиматоры, (iii) многослойный перцептрон (Multi-Layer Perceptron - MLP) как нейросетевая реализация метамодели.

Данные разделяются на обучающую и тестовую части. Помимо случайного разбиения применяется проверка обобщающей способности для различных входных параметров.

Для построения метамодели были протестированы несколько стандартных семейств регрессионных алгоритмов:

- **Ridge Regression** - линейная модель с L2-регуляризацией;
- **Gradient Boosting Regressor (GBR)** — ансамблевый метод на решающих деревьях;
- **Hist Gradient Boosting (HGB)** — ускоренный вариант бустинга;
- **Random Forest (RF)** — ансамбль деревьев;
- **MLP Regressor** — многослойный перцептрон (нейросетевая модель).

Во всех экспериментах использовались реализации из библиотеки scikit-learn с типовыми гиперпараметрами по умолчанию. Для нейросетевой

модели применялась предварительная стандартизация признаков (StandardScaler).

Обучение проводилось на случайном разбиении выборки в пропорции 80/20.

Таблица 1. Результаты эксперимента

Модель	**MAE** $W_q^{(serv)}$	**MAE** L_q	**MAE** P_{serv}	**Train time (s)**	**Pred time 10k (s)**
Ridge	0.212	0.256	0.0335	0.008	0.026
GBR	0.113	0.236	**0.0158**	0.205	0.058
HGB	0.119	0.226	0.0204	1.135	0.124
RF	0.127	0.197	0.0275	0.876	0.218
MLP	**0.0838**	**0.134**	0.0784	1.691	0.103

Полученные результаты показывают:

- Наилучшую точность в прогнозировании характеристик ожидания и длины очереди продемонстрировала нейросетевая модель **MLP**, обеспечив минимальные значения средней абсолютной ошибки (Mean Absolute Error - MAE) для $W_q^{(serv)}$ и L_q.

- Для доли обслуженных заявок Pserv наиболее устойчивым оказался ансамблевый метод **градиентного бустинга**, обеспечивающий минимальную ошибку MAE =0.0158.

- Линейная модель Ridge существенно уступает нелинейным методам, что подтверждает сложный характер зависимости характеристик СМО от входных параметров.

Дополнительно был проведен анализ качества моделей в зависимости от коэффициента загрузки ρ. Показано, что при переходе системы в область перегрузки (ρ>1) ошибка аппроксимации для $W_q^{(serv)}$ и L_q возрастает для всех моделей. Для проверки обобщающей способности метамоделей было проведено дополнительное исследование: обучение выполнялось на системах с $c \in \{1,2,4\}$, а тестирование — при $c=8$.

Результаты показали, что:

- нейросетевая модель сохраняет конкурентоспособность по времени ожидания,
- однако ансамблевые методы (HGB, GBR) оказываются более устойчивыми при переносе на новые структурные параметры системы.

Таким образом, построенная метамодель позволяет с высокой точностью аппроксимировать ключевые показатели эффективности многоканальной СМО смешанного типа. Наиболее точные результаты для временных характеристик достигнуты с использованием нейросетевого перцептрона, тогда как для вероятностной метрики обслуживания оптимальным оказался градиентный бустинг.

Литература

1. Ахметшин Д.А., Титовцев А.С. Система массового обслуживания с взаимопомощью между каналами и ограниченным временем пребывания в очереди // Инженерный вестник Дона. 2024. №12. URL: ivdon.ru/ru/magazine/archive/n12y2024/9688

2. Царькова Е. Г. Математическая модель управления системой массового обслуживания с динамической дисциплиной обслуживания заявок // Инженерный вестник Дона. 2022. №5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7638

3. Лабинский А.Ю. Моделирование системы массового обслуживания с использованием нейронной сети. 2019. URL: cyberleninka.ru/article/n/modelirovanie-sistemy-massovogo-obsluzhivaniya-s-ispolzovaniem-neyronnoy-seti

4. Dieleman N.A. et al. A neural network approach to performance analysis of tandem queuing lines: The value of analytical knowledge. Computers & Operations Research. 2023. 152. URL: sciencedirect.com/science/article/pii/S0305054822003549

5. Efrosinin D. Use Cases of Machine Learning in Queueing Theory Based on a GI/G/K System. Mathematics. 2025. 13(5). URL: mdpi.com/2227-7390/13/5/776

6. Norrman F. Prediction of queuing behaviour through the use of artificial neural networks. 2017. URL: diva-portal.org/smash/get/diva2:1111289/fulltext01.pdf

7. Sherzer E. et al. Can machines solve general queueing systems? 2022. URL: arxiv.org/pdf/2202.01729

8. Титовцев А.С., Антонова П.В. Характеристики системы массового обслуживания с ограниченным временем пребывания заявки в очереди и временем обслуживания, распределенным по закону Эрланга. Научно-технический вестник Поволжья. 2021. № 8. С. 79-82.

9. Bengio Y. Practical Recommendations for Gradient-Based Training of Deep Architectures. Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. 2012. 7700. URL: doi.org/10.1007/978-3-642-35289-8_26

10. Heaton J., Goodfellow I., Bengio Y., Courville A. Deep learning. Genet Program Evolvable Mach. 2018. 19. URL: doi.org/10.1007/s10710-017-9314-z

11. Ioffe S., Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015. URL: arxiv.org/pdf/1502.03167

12. Zinoviev D. Discrete Event Simulation: It's Easy with SimPy! 2024. URL: arxiv.org/pdf/2405.01562

References

1. Akhmetshin D.A., Titovtsev A.S. Inzhenernyj vestnik Dona. 2024. No. 12. URL: ivdon.ru/ru/magazine/ archive/n12y2024/9688

2. Tsarkova E. G. Inzhenernyj vestnik Dona, 2022, No 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7638

3. Labinskiy. A.Yu. 2019. URL: cyberleninka.ru/article/n/modelirovanie-sistemy-massovogo-obsluzhivaniya-s-ispolzovaniem-neyronnoy-seti

4. Dieleman N.A. et al. Computers & Operations Research. 2023. 152. URL: sciencedirect.com/science/article/ pii/S0305054822003549

5. Efrosinin D. Mathematics. 2025. 13(5). URL: mdpi.com/2227-7390/13/5/776

6. Norrman F. 2017. URL: diva-portal.org/smash/get/diva2:1111289/ fulltext01.pdf

7. Sherzer E. et al. 2022. URL: arxiv.org/pdf/2202.01729

8. Titovtsev A.S., Antonova P.V. Scientific and Technical Volga region Bulletin. 2021. No 8. P. 79-82.

9. Bengio Y. Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science. 2012. 7700. URL: doi.org/10.1007/978-3-642-35289-8_26

10. Heaton J., Goodfellow I., Bengio Y., Courville A. Genet Program Evolvable Mach. 2018. 19. URL: doi.org/10.1007/s10710-017-9314-z

11. Ioffe S., Szegedy C. 2015. URL: arxiv.org/pdf/1502.03167

12. Zinoviev D. 2024. URL: arxiv.org/pdf/2405.01562

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 6.01.2026

Дата публикации: 3.03.2026