

МЕТОДЫ ПРОГНОЗИРОВАНИЯ ВРЕМЕНИ ВЫПОЛНЕНИЯ ЗАДАЧ В ГЕТЕРОГЕННОЙ СЕТИ ВЫЧИСЛИТЕЛЕЙ

В. О. Писковский^{1,}, Д. Р. Лысенко¹*

¹ Московский государственный университет им. М. В. Ломоносова,
Москва, Россия

Авторы рассматривают применимость методов машинного обучения для прогнозирования времени выполнения задач на узлах высокопроизводительных вычислительных устройств с целью эффективного использования их ресурсов и, следовательно, оптимального планирования. Цель работы состоит в том, чтобы найти методы и ансамбли методов, позволяющие минимизировать расхождение между прогнозируемыми и фактически израсходованными ресурсами, например временем выполнения данной задачи. Рассмотренные методы машинного обучения основаны на журналах работы систем управления заданиями, содержащих данные об использовании запрошенных вычислительных ресурсов. Оценки сопровождаются расчетной дисперсией и доверительным интервалом прогнозируемой величины.

We consider an applicability of machine learning methods to predict task execution time on nodes of high-performance computing devices in order to guarantee an efficient usagen of their resources and therefore optimal planning. The purpose is to find methods and their ensemble to minimize discrepancy between predicted and actually used up resources, e.g., the time to perform a given task. Considered machine learning methods are based on data logging computational resource consumptions for a pretty long time period. This assessment goes along with estimated dispersion to assure a proper confidence interval.

PACS: 89.20.Ff

ВВЕДЕНИЕ

Эффективность вычислительных систем зависит от того, насколько оптимально распределяются задачи и используются ресурсы компьютера. Планировщики задач используются как средство регистрации и планирования использования ресурсов и управления заданиями на высокопроизводительных компьютерах. На основе логов таких планировщиков, которые учитывают использование ресурсов для выполнения вычислительных задач, были обучены и верифицированы модели машинного обучения. Целью машинного обучения и его ансамблей является задача прогнозирования времени выполнения, требуемого узлом в гетерогенной

* E-mail: vpiskovski@asvk.cs.msu.ru

сети высокопроизводительных компьютеров с заданными характеристиками (типом и количеством ядер, объемом памяти и т. д.).

ПОСТАНОВКА ЗАДАЧИ

Статья продолжает предыдущие работы [1–4]. В [4] задача была математически сформулирована следующим образом.

Даны временные ряды журнала использования: $X = \{\mathbf{x}^n | \mathbf{x} \in R^D, n = \overline{1, N}\}$, где $\mathbf{x}^n$ — n -й вектор анализируемого временного ряда ($n = \overline{1, N}$); R^D — пространство зарегистрированных атрибутов процесса; D — количество зарегистрированных атрибутов; N — количество записей.

Задача состоит в том, чтобы определить оптимальный метод прогноза $F_k^*(X)$ следующим образом: $F_k^*(X) = \operatorname{argmin}_{p=\overline{1, P}} \|F_k^p(X) - y_k\|_{l_1}$, где k —

индекс прогнозируемого компонента; y_k — k -е значения прогнозируемого компонента ($y \in R^D$); $F_k = \{F_k^p | k = \overline{1, K}, p = \overline{1, P}\}$ — набор применимых методов; K — количество прогнозируемых компонентов. Поскольку в данной работе авторы рассматривают единственный прогнозируемый компонент, т. е. задачу времени выполнения, то $K = 1$; P — количество применимых методов.

НАБОРЫ ДАННЫХ

Наборы данных содержат статистические данные о работе суперкомпьютерной системы, основанной на системе управления заданиями SLURM [5]. Источники наборов данных представлены в табличном формате, где каждая строка соответствует одной вычислительной задаче.

Исследование проводилось с использованием ряда наборов данных:

- инфраструктуры совместного исследовательского центра «Высокопроизводительные вычисления и большие данные» (ЦКП «Информатика») ФИЦ КНЦ РАН (Федеральный исследовательский центр «Информатика и управление» РАН, Москва) [6];
- Межведомственного суперкомпьютерного центра РАН [7]. Набор содержит записи о задачах, выполняемых в суперкомпьютерном центре;
- Суперкомпьютерного центра «Политехнический» Санкт-Петербургского политехнического университета Петра Великого [8].

Объем, характеристики наборов данных показаны ниже (таблица).

Наборы данных

Характеристика наборов данных	ФИЦ РАН [6]	МСЦ [7]	СПбПУ [8]
N (записи)	48 289	2 304 380	1 545 793
D (атрибуты)	106	13	22
T (период)	01.01.23–27.11.24	24.04.07–17.06.21	22.08.21–31.08.23

МЕТОД ИССЛЕДОВАНИЯ

Методология исследования состояла из следующих этапов:

- предварительная обработка данных;
- обучающая модель;
- модель проверки.

Суть предварительной обработки данных, по аналогии с формированием корпоративных хранилищ данных [9], состоит из следующих этапов.

- Очистка данных: удаление повторяющихся записей и записей, содержащих ошибочные или пустые ячейки.
- Выбор записей, соответствующих нормально выполненным задачам, для выполнения которых было выделено достаточно ресурсов, включая процессорное время.
- Классификация вычислительных задач по времени выполнения.
- Уменьшение объема набора данных за счет сохранения только наиболее важных компонентов с помощью метода анализа главных компонент (PCA) или усеченной сингулярной декомпозиции (TSVD). Статьи [4, 10] посвящены этим процедурам уменьшения размерности. Как показано в упомянутых работах, количество учитываемых атрибутов можно уменьшить в несколько раз при ухудшении оценки примерно на 5 %.
- Преобразование данных в единый формат.
- Добавление дополнительных избыточных атрибутов: час, день недели, дата, месяц и год.

Если для обучения не хватает данных, предлагается применять методы, широко используемые в рекомендательных системах, такие как [1–3], наряду с применением квантования, если это необходимо.

Использовались следующие модели или методы сборки:

- Lasso regression;
- Ridge regression;
- Decision tree;
- Random forest;
- Gradient boosting;
- K-nearest neighbors.

РЕЗУЛЬТАТЫ

В ходе экспериментальной работы были реализованы механизмы управления размером выборки и ее временными границами.

Эксперименты проводились в два этапа: базовый этап и этап оптимизации. Оптимизация гиперпараметров проводилась с использованием Optuna [11] — автоматизации поиска наилучших комбинаций параметров моделей.

Результаты сравнительного анализа применения разных моделей показали, что ансамблевые методы демонстрируют наилучшую производительность. Линейные методы (Linear Regression, Lasso, Ridge) оказались наименее эффективными, с низкими значениями коэффициента детерминации R^2 .

После оптимизации параметров были получены следующие результаты:

- ФИЦ ИУ РАН: оптимизация снизила ошибки, но R^2 остался на уровне базовых моделей;
- МСЦ РАН: оптимизация улучшила метрики для большинства моделей, особенно для Gradient boosting;
- СПбПУ: оптимизация повысила производительность Random forest.

ВЫВОДЫ

1. Разработаны модели машинного обучения для прогнозирования выполнения задач.
2. Градиентный бустинг показал баланс между точностью и скоростью обучения.
3. Наилучшие результаты получены для групповых методов: случайного леса и градиентного бустинга.
4. Ансамблевые методы демонстрируют наилучшую производительность.

СПИСОК ЛИТЕРАТУРЫ

1. Antonenko V., Petrov I., Smeliansky R. L. et al. MC2E: Meta-Cloud Computing Environment for HPC // Performance Evaluation Models for Distributed Service Networks. Cham, Switzerland: Springer Nature Switzerland AG, 2019.
2. Chupakhin A., Kolosov A., Bahmurov A. et al. Application of Recommender Systems Approaches to the MPI Program Execution Time Prediction // 2020 Intern. Sci. Techn. Conf. "Modern Computer Network Technologies" (MoNeTeC). IEEE. 2020. P. 1–7.
3. Smeliansky R. Network Powered by Computing // 2022 Intern. Conf. Modern Network Technologies (MoNeTec). IEEE. 2022. P. 1–5.
4. Grusho A. A. et al. Predicting the Network Service Quality via the Log of Hardware Usage // J. Comput. Systems Sci. Intern. Pleiades Publ., Ltd, 2025. V. 64, No. 3. P. 452–459.
5. SLURM Workload Manager. 2025. <https://slurm.schedmd.com> (accessed 20.05.2025).
6. Федеральный исследовательский центр «Информатика и управление» Российской академии наук. <https://www.frccsc.ru/> (дата обращения 20.05.2025).
7. Joint Supercomputer Center of the Russian Academy of Sciences. <https://rscgroup.ru/en/project/jscc/> (accessed 14.09.2025).

8. Суперкомпьютерный центр «Политехнический» Санкт-Петербургского политехнического университета Петра Великого. <https://scc.spbstu.ru/> (дата обращения 20.05.2025).
9. Enterprise Data Warehouse: Components, Basic Concepts, and Types of Architectures EDW. <https://habr.com/ru/articles/693360/> (accessed 14.09.2025).
10. Писковский В. О., Лычева Е. О., Могиленец В. М. Прогнозирование временных характеристик сетевых сервисов приложений // Моск. ун-т компьютерной математики и кибернетики. 2025. Т. 49, №. 3. С. 238–248.
11. *Takuya Akiba, Shotaro Sano, Toshihiko Yanase et al.* Optuna: A Next-Generation Hyperparameter Optimization Framework // Proc. 25th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining. 2019.