УДК 004.89

# ПАЙПЛАЙН АВТОМАТИЧЕСКОГО МАШИННОГО ОБУЧЕНИЯ ДЛЯ ОБРАБОТКИ ЛОГОВ В ОБОБЩЕННОМ ПРИЗНАКОВОМ ПРОСТРАНСТВЕ

**Муратов С.Ю.** (ИТМО)

**Научный руководитель – кандидат технических наук, доцент Муравьёв С.Б.** (ИТМО)

**Введение.** Современные подходы мониторинга программных систем и информационной инфраструктуры подразумевают оптимальное сочетание скорости реакции на инциденты и эффективности их решения. Внедрение инструментов мониторинга, использующих машинное обучение, позволяет достичь данного сочетания. Однако, существует ряд проблем, сужающих область применения подобных инструментов, а именно: высокий порог интеграции, недостаточная гибкость и скорость работы. Данные проблемы в определенной мере решены услугами крупнейших облачных провайдеров, но их применение ограничено в силу невозможности внедрения их продуктов на корпоративном уровне. Кроме того, эффективность подобных инструментов мониторинга, в частности обработки логов, зависит от степени доступности облачных технологий для заданных программных комплексов и информационной инфраструктуры. Одним из способов решения проблемы является применение автоматического машинного обучения [1].

**Основная часть.** Пайплайн автоматического машинного обучения предоставляет подготовленные на данных логов, перешедших в целевое признаковое пространство, модели машинного обучения. Модели для каждой из задач машинного обучения (обнаружение аномалий, классификация и кластеризация) выбраны в соответствии с актуальными данными бенчмарков в области обработки логов для различных наборов данных [2]. При отборе моделей наиболее важным критерием стало время вывода, тогда как значение метрик считалось допустимым в рамках 85% верхнего уровня значений. Наиболее распространенные большие языковые модели не прошли ни одного эксперимента в силу большого времени вывода. Несмотря на способность больших языковых моделей работать с большим количеством контекста, длительное время ожидания их вывода, свыше 10 секунд на запрос, приводит к полной невозможности применения для решения текущей задачи. На вход конвейеру необходимо предоставить либо файловый объект с необработанными логами, либо точку потокового вывода логов. Далее производятся операции по сборке и подготовке данных логов. После этого применяется алгоритм автоматического формирования наборов данных логов в обобщенном признаковом пространстве. С помощью полученных наборов данных можно либо запустить автоматический тюнинг гиперпараметров одной или нескольких моделей, при условии возможного горизонтального масштабирования, либо использовать уже готовые модели. Основные критерии те же, что и при отборе: первичный – время вывода модели, вторичный – значение метрик качества модели. По результатам бенчмарков и последующего отбора выбраны следующие модели обнаружения аномалий: изолирующий лес, метод опорных векторов с одним классом, эллипсоидальная аппроксимация данных, feature/rotated bagging, модели классификации: к-ближайших соседей, градиентный бустинг, логистическая регрессия, случайный лес, модели кластеризации: к-средних, DBSCAN, MeanShift, HDBSCAN. Реализация внутреннего конвейера для моделей обнаружения аномалий основана на Automatic Outlier Detection [3], для моделей классификации основана на LightAutoML [4], а для моделей кластеризации с помощью библиотеки автоматической кластеризации мультимодальных данных на Apache Spark с открытым исходным кодом – Sparkling [5].

**Выводы.** Проведён отбор моделей обнаружения аномалий, классификации и кластеризации логов. Разработан пайплайн автоматического машинного обучения для обработки логов в обобщенном признаковом пространстве.

**Список использованных источников:**


1. Jiang Z., Liu J., Huang j., Li Y., Huo Y., Gu J., Chen Z., Zhu J., Lyu M.R. A Large-scale Evaluation for Log Parsing Techniques: How Far are We? // 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis. – ISSTA. 2024. – P. 223 – 234.
2. Khan Z.A, Shin D., Bianculli D., Briand L. Guidelines for Assessing the Accuracy of Log Message Template Identification Techniques // International Conference on Software Engineering. – 2022. – P. 1095–1106.
3. Bahri M., Salutari F., Putina A., Sozio M. AutoML: state of the art with a focus on anomaly detection, challenges, and research directions // International Journal of Data Science and Analytics. – Springer. 2022. – P. 113–126.
4. Vakhrushev A., Ryzhkov A., Simakov D., Damdinov R., Savchenko M., Tuzhilin A. LightAutoML: AutoML Solution for a Large Financial Services Ecosystem. – 2021. – P. 55–67.
5. Muravyov S.B., Usov I.S. An opensource library for automl multimodal clustering on apache spark // Записки научных семинаров ПОМИ. 2024. Vol. 540, no. 0. P. 178–193.