Научная статья
Original article
УДК 004.622.0

СИСТЕМЫ AUTOML КАК СОВРЕМЕННЫЙ ИНСТРУМЕНТ ДЛЯ ПОСТРОЕНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

AUTOMNL SYSTEMS AS A MODERN TOOL FOR BUILDING MACHINE LEARNING MODELS

Попова Инна Андреевна, аспирант, МГТУ имени Н.Э. Баумана (национальный исследовательский университет) (105005 Россия, г. Москва, ул. 2-я Бауманская, д. 5, стр. 1), тел. +7(977) 954-47-73, popovai1@student.bmstu.ru

Popova Inna Andreevna, graduate student, Moscow State Technical University named after N.E. Bauman (National Research University) (2nd Baumanskaya st., 5, b. 1, Moscow, 105005 Russia), tel. +7(977) 954-47-73, popovai1@student.bmstu.ru

Аннотация. Специалисты в области машинного обучения сталкиваются с задачей выбора подходящего алгоритма с оптимальными гиперпараметрами для описания исследуемого набора данных. Для этого они обычно выполняют и оценивают множество конфигураций методом проб и ошибок. Однако для начинающих аналитиков данных это трудоемкая задача. Последние достижения в области исследований AutoML решают эту проблему путем автоматического поиска подходящего алгоритма с соответствующими гиперпараметрами. Тем не менее, эти системы применимы только для задач машинного обучения обучения с учителем.

AutoML является современной тенденцией в области машинного обучения. В статье рассмотрен ряд систем автоматизированного машинного обучения, описаны их ключевые особенности, решаемые задачи.

Abstract. Specialists in the field of machine learning are faced with the task of choosing a suitable algorithm with optimal hyperparameters to describe the data set under study. To do this, they usually perform and evaluate many configurations by trial and error. However, for novice data analysts, this is a time-consuming task. Recent advances in automated machine learning research solve this problem by automatically searching for a suitable algorithm with appropriate hyperparameters. However, these systems are only applicable for machine learning tasks of learning with a teacher.

AutoML is a modern trend in the field of machine learning. The article considers a number of automated machine learning systems, describes their key features, solved tasks.

Ключевые слова: машинное обучение, AutoML системы, алгоритм, метод, модель, гиперпараметры.

Keywords: machine learning, AutoML systems, algorithm, method, model, hyperparameters.

Автоматизированное машинное обучение или AutoML - это термин, предложенный сообществом специалистов в области машинного обучения для обозначения методов, направленных на автоматизацию проектирования, разработку систем и приложений машинного обучения. Данное направление в области машинного обучения развивается стремительными темпами, о чем свидетельствует большое число публикаций и разработок зарубежных и отечественных исследователей.

AutoML предоставляет методы и процессы, позволяющие сделать машинное обучение доступным для специалистов, не занимающихся машинным обучением, повысить эффективность машинного обучения и ускорить исследования в области машинного обучения. Данные системы можно использовать, не обладая экспертными знаниями в области машинного обучения. Так как современные приложения работают зачастую с большим

объемом данных, возникает потребность в системах, которые эти данные могут быстро обрабатывать. Для этого и создавались AutoML системы.

Зачастую процесс обучения модели машинного обучения состоит из ряда этапов:

- предварительная обработка данных;
- выбор нужных признаков;
- выбор подходящей модели;
- оптимизация гиперпараметров модели;
- обучение модели с настроенными гиперпараметрами;
- анализ полученных результатов.

Системы AutoML помогают автоматизировать ряд этапов, при этом предсказательная точность целевой модели машинного обучения не будет потеряна.

При помощи систем AutoML можно выполнять следующие операции:

- автоматическая очистка данных (Auto Clean)
- автоматический подбор признаков (Auto FE)
- оптимизация гиперпараметров (HPO)
- мета-обучение
- определение нейронной архитектуры (NAS).

Можно интуитивно описать AutoML как задачу поиска f, которое лучше обобщает любое возможное T с меньшим человеческим вмешательством. Где f может быть композицией нескольких функций, которые могут преобразовывать входное пространство, подвыборку данных, комбинировать несколько предикторов и т. д. Функции вида $f(x) = \nu_{\theta_\nu}(\Phi_{\theta_\Phi}(x))$ называются полными моделями [3] или конвейерами [2], поскольку они включают в себя все процессы, которые должны применяться к данным, чтобы получить модель обучения с учителем. AutoML можно рассматривать как поиск функций υ и Ф с соответствующими гиперпараметрами θ_ν и θ_Φ с использованием D.

На рисунке 1 приведен пример общей архитектуры системы AutoML, которая построена на основе системы Auto-Sklearn. Цель данных систем

заключается в подборе подходящей конфигурации (модели машинного обучения) для тренировочного набора данных на основе выделенных ресурсов (Budget).

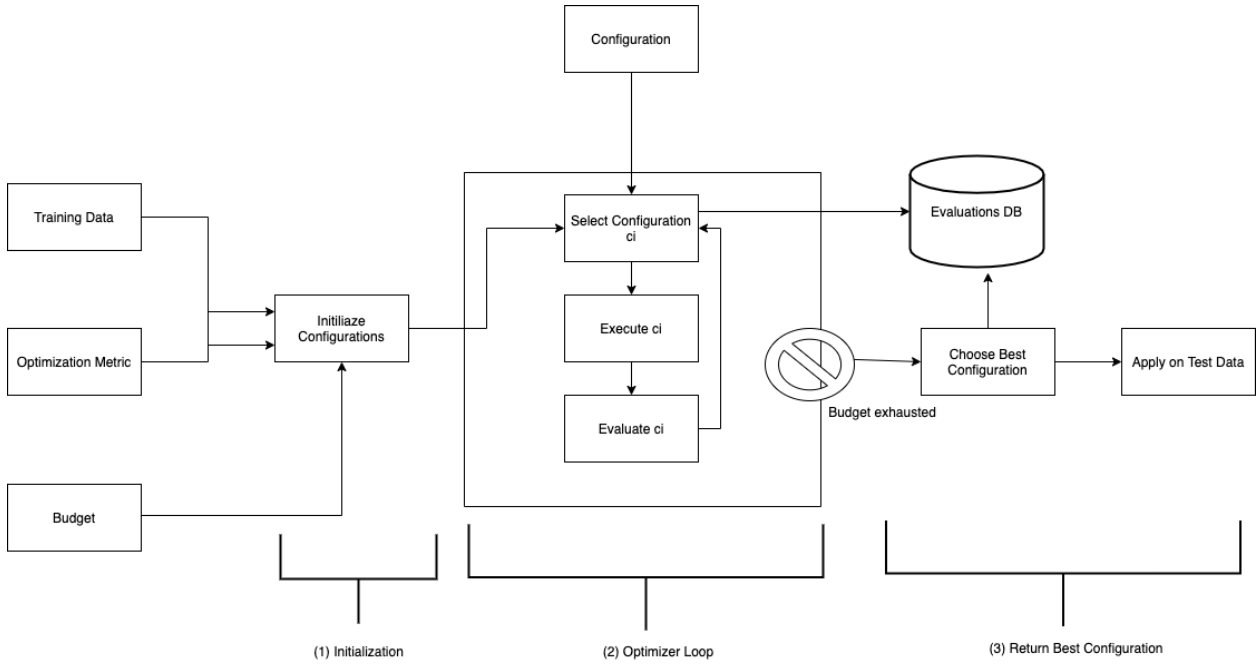


Рисунок 1. Общая архитектура системы AutoML, адаптировано из [11].

На этапе инициализации оптимизируется конфигурация для оптимизаторов. Цикл оптимизации запускает итерации оптимизатора пока не будет исчерпан бюджет (Budget) [1]. Этап возврата лучшей конфигурации возвращает наиболее подходящую конфигурацию модели.

Рассмотрим ряд систем AutoML, которые используются в современных проектах машинного обучения.

Auto-Weka

Разработчики проекта Auto-Weka предложили выбор алгоритма и оптимизацию гиперпараметров рассматривать как общую задачу, назвав ее C.A.S.H [5]. Алгоритм использует методы байесовской оптимизации, которые итеративно строят модели ландшафта алгоритмов/гиперпараметров и используют эти модели для определения новых точек в пространстве, которые стоит исследовать. Основное внимание в Auto-Weka уделяется классификации. Предлагаемый метод может быть применен к любой задаче, которая может рассматриваться как задача оптимизации. Auto-Weka решает проблему C.A.S.H (Combination and Selection and Hyperparameter optimization). с помощью

конфигурации алгоритма последовательной модели - SMAC, метода байесовской оптимизации со случайными лесами, который обрабатывает дискретные входные данные большой размерности.

Из-за перебора всех возможных моделей и их гиперпараметров алгоритм работает достаточно долго.

Auto-Sklearn

Система написана на Python и использует алгоритмы и методы из Scikit-Learn (15 алгоритмов классификации, 14 методов обработки признаков, 4 метода предварительной обработки данных). Auto-Sklearn реализует алгоритм SMAC для настройки гиперпараметров. Система умеет генерировать признаки, выбирать модель, настраивать гиперпараметры. Auto-Sklearn представляет два новшества: использование метаобучения для горячего запуска процедуры байесовской оптимизации и шаг построения ансамбля, который использует более одной конфигурации, найденной в процедуре оптимизации. Auto-Sklearn запускает процедуру байесовской оптимизации, предоставляя начальные экземпляры из конфигураций, которые дали наилучшие результаты для аналогичных наборов данных. В частности, метаобучение показало значительные улучшения во всех оценках конфигурации в процедуре байесовской оптимизации, а построение ансамбля оказалось выгодным с помощью метаобучения, поскольку для построения ансамбля были выбраны более эффективные модели [6].

TPOT

TPOT (Tree-based Pipeline Optimization Tool) является проектом на Python с открытым исходным кодом, который альтернативен байесовским методам оптимизации, автоматизирует конвейеры машинного обучения с помощью генетического программирования (GP), хорошо известного метода автоматического построения программ. То есть в данной системе конвейер машинного обучения полностью автоматизирован и для определения оптимальной модели применяется генетический алгоритм. Основное внимание в этом проекте уделяется обучению с учителем, а именно задаче классификации с поддержкой ста пятидесяти алгоритмов ScikitLearn [4], включая алгоритмы

предварительной обработки. Система, как и Auto-Sklearn представляет собой надстройку над библиотекой Sklearn, однако использует собственные алгоритмы регрессии и классификации.

Случайный набор из 100 древовидных конвейеров создает начальное поколение, как указано в генетическом программировании, и дополнительно оптимизируется в соответствии с пакетом Python DEAP. Отобраны двадцать лучших конвейеров с точки зрения максимальной точности перекрестной проверки и минимизации количества процессов, которые видоизменяются для создания конвейеров нового поколения. Каждый из конвейеров производит еще пять с помощью перекрестных методов или случайных вставок и усадок. Алгоритм выполняется для 100 поколений для каждого из обновляемых решений.

TPOT не может работать с естественным языком и категориальными признаками.

Google Vizier

Google Vizier — это новейшее исследование, посвященное оптимизации черного ящика. Он предоставляет основные компоненты для подсистемы облачного машинного обучения Google HyperTune, обладает высокой масштабируемостью и реализован на Python, C ++ и Golang.

Ключевой особенностью Google Vizier является динамический выбор алгоритмов оптимизации. По умолчанию система использует в качестве алгоритма оптимизации метод Gaussian Process Bandits. Данный метод оптимизации, обеспечивает отличные результаты, но масштабируется как O(3) [6] с количеством обучающих точек, поэтому, если Gaussian Process Bandits выбран для запуска процедуры оптимизации, позже он может быть динамически заменен более масштабируемыми алгоритмами. Важной особенностью является автоматическая ранняя остановка. Данный метод, направлен на прекращение подбора параметров, если в следующей конфигурации, которая должна быть оценена, не предполагается, что будут какие-либо улучшения по сравнению с последней оцененной конфигурацией.

Упомянутый метод реализован с помощью двух различных методов: первый - правило остановки кривой производительности, где регрессор гауссовского процесса обучается на частичной кривой производительности и параметрах процедуры оптимизации, чтобы предсказать, является ли оптимальное значение целевой функции, найденное до любой заданной точки, достаточно низким для завершения процесса.

Второе правило остановки медианы, правило, которое завершает процесс оптимизации, если на любом этапе найденная цель строго хуже, чем среднее значение предыдущих.

В таблице 1 представлены отличительные черты рассмотренных AutoML систем.

	Auto-Sklearn	Auto-Weka	TPOT	Google Vizier
Класс решаемых задач	Обучение с учителем	Обучение с учителем	Обучение с учителем	Обучение с учителем
Оптимизация гиперпараметров (HPO)	Байесовская оптимизация	Байесовская оптимизация	Генетическая оптимизация	Различные алгоритмы оптимизации, зачастую применяется Batched Gaussian Process
Платформа раелизации	Python	Weka	Python	Python, C#, Golang
Ключевые особенности	Предварительный запуск HPO с помощью Мета-функций	Первая завершенная AutoML система	Генетическая оптимизация, минимизация количества процессов	Можно выбрать алгоритм динамической оптимизации, Transfer learning

Таблица 1. Отличительные черты рассмотренных AutoML систем.

Рассмотренные системы различаются алгоритмами, используемыми для настройки гиперпараметров, масштабируемостью и платформами, на которых реализованы.

Байесовская оптимизация является наиболее популярным выбором для настройки гиперпараметров, но доказано, что такие подходы, как генетическое программирование и процессы Bandit, также обеспечивают хорошие результаты. Кроме того, можно заметить, что большинство систем реализованы на Python, и этот выбор не является необоснованным. Scikit Learn предоставляет широкий спектр алгоритмов машинного обучения наряду со многими алгоритмами предварительной обработки, а библиотеки, такие как Keras и Pytorch, являются одними из наиболее популярных вариантов для проектирования современных нейронных сетей.

Все современные AutoML системы поддерживают генерацию моделей машинного обучения с учителем, это связано со сложностями выбора целевой функции для оптимизации. Мета-обучение используется только в AutoSKlearn и в качестве подсистемы Google Vizier.

Можно отметить, что AutoML системы все же не смогут полностью вытеснить человека как участника процесса обработки данных, так как функцию целеполагания они обеспечить не могут [8].

Вывод

В статье рассмотрены основные тенденции в области машинного обучения - AutoML системы. Автоматизированное машинное обучение заменяет традиционный подход к построению систем машинного обучения, при использовании которого требуется знать практически все существующие алгоритмы машинного обучения, уметь их правильно применить и сконфигурировать.

В ходе обзора были выделены следующие перспективные направления развития в области AutoML систем:

- использование сложных нелинейных конвейеров обработки данных;

- внедрение алгоритмов машинного обучения без учителя для предварительной обработки данных;
- подбор алгоритма оптимизации гиперпараметров выбранной оптимальной модели.

Таким образом, автоматизация машинного обучения активно развивается и является передовой областью исследований, о чем свидетельствует большое число разработок и исследований в данной области машинного обучения.

Литература

1. Dai J., He K, J. Sun. Instance-aware semantic segmentation via multi-task network cascades. 2014. Режим доступа: https://www.cv-foundation.org/openaccess/content_cvpr_2016/app/S14-02.pdf (дата обращения: 05.01.2022).
2. Koh P.W., Liang P. Understanding black-box predictions via influence functions // Proceedings of the 34th International Conference on Machine Learning, PMLR. Vol. 70. P. 1885–1894.
3. Olson R.S. TPOT: A Python tool for automating data science. 2016. Режим доступа: https://www.kdnuggets. com/2016/05/tpot-python-automating-data-science.html/2 (дата обращения: 05.01.2022).
4. Olson R.S., Moore J. H. TPOT: A tree-based pipeline optimization tool for automating machine learning. P. 66–74.
5. Thornton C. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms ACM, P. 847–855.
6. Zhang Y., Yang Q. A survey on multi-task learning. Режим доступа: https://arxiv.org/abs/1707.08114 (дата обращения: 06.09.2018).
7. Caruana R. Multitask learning springer, Boston, MA, 1998, P. 95–133.
8. Коротеев М. В. Обзор некоторых современных тенденций в технологии машинного обучения // E-Management. 2018. №1. URL: https://cyberleninka.ru/article/n/obzor-nekotoryh-sovremennyh-tendentsiy-v-tehnologii-mashinnogo-obucheniya (дата обращения: 12.01.2022).

9. Snoek J., Larochelle H., Ryan P. A. Practical bayesian optimization of machine learning algorithms, 2012.
10. Allingham J.U.. Unsupervised automatic dataset repair. Master's thesis, Computer Laboratory, University of Cambridge, 2018.
11. Tschechlov D. Analysis and transfer of AutoML concepts of clustering algorithms. Master's thesis, Computer Laboratory, Institute of parallel and distributed systems, University of Stuttgart, 2019.

References

1. Dai J., He K, J. Sun. Instance-aware semantic segmentation via multi-task network cascades. 2014. Режим доступа: https://www.cv-foundation.org/openaccess/content_cvpr_2016/app/S14-02.pdf (дата обращения: 05.01.2022).
2. Koh P.W., Liang P. Understanding black-box predictions via influence functions // Proceedings of the 34th International Conference on Machine Learning, PMLR. Vol. 70. P. 1885–1894.
3. Olson R.S. TPOT: A Python tool for automating data science. 2016. Режим доступа: https://www.kdnuggets. com/2016/05/tpot-python-automating-data-science.html/2 (дата обращения: 05.01.2022).
4. Olson R.S., Moore J. H. TPOT: A tree-based pipeline optimization tool for automating machine learning. P. 66–74.
5. Thornton C. Auto-WEKA: combined selection and hyperparameter optimization of classification algorithms ACM, P. 847–855.
6. Zhang Y., Yang Q. A survey on multi-task learning. Режим доступа: https://arxiv.org/abs/1707.08114 (дата обращения: 06.09.2018).
7. Caruana R. Multitask learning springer, Boston, MA, 1998, P. 95–133.
8. Koroteev M. V. Review of some modern trends in machine learning technology // E-Management. 2018. No. 1. URL: https://cyberleninka.ru/article/n/obzor-nekotoryh-sovremennyh-tendentsiy-v-tehnologii-mashinnogo-obucheniya (date of access: 12.01.2022).

9. Snoek J., Larochelle H., Ryan P. A. Practical bayesian optimization of machine learning algorithms, 2012.
10. Allingham J.U.. Unsupervised automatic dataset repair. Master’s thesis, Computer Laboratory, University of Cambridge, 2018.
11. Tschechlov D. Analysis and transfer of AutoML concepts of clustering algorithms. Master’s thesis, Computer Laboratory, Institute of parallel and distributed systems, University of Stuttgart, 2019.

© Попова И.А., 2022 Научно-образовательный журнал для студентов и преподавателей «StudNet» №1/2022.

Для цитирования: Попова И.А. СИСТЕМЫ AUTOML КАК СОВРЕМЕННЫЙ ИНСТРУМЕНТ ДЛЯ ПОСТРОЕНИЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ// Научно-образовательный журнал для студентов и преподавателей «StudNet» №1/2022.