Исследование методов AutoML в задаче классификации волновых данных

Е. А. Неверов[1], И. И. Викснин[1], С. С. Чупров[2,3]
[1] *Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)*
[2] *Колледж компьютерных и информационных наук им. Б. Томаса Голизано*
[3] *Рочестерский институт технологий*
datnever@ya.ru

***Аннотация*. В данной работе рассматриваются методы автоматического машинного обучения для решения задачи классификации объектов. Automated machine learning представляет собой процесс автоматического создания и оптимизации моделей машинного обучения, способен производить автоматический отбор признаков, оптимизацию гиперпараметров, выбор модели и оценку модели на основе заданных метрик качества. В задаче классификации ряда сигналов автоматизация этого процесса может быть особенно полезной, поскольку этот тип задач может быть сложным в декомпозиции и требует большого объема работы по выделению значимых признаков. Кроме того, AutoML имеет высокую применимость в задаче построения нейронных сетей, например, нейронных сетей прямого распространения, и оптимизации таких гиперпараметров, как количество слоев, количество нейронов. В работе были проанализированы различные подходы к AutoML. Наилучший результат показали методы на основе MLJAR AutoML, AutoKeras и TPOT.**

Ключевые слова: AutoML; машинное обучение; нейронные сети; анализ данных

I. Введение

Для решения задачи классификации волновых данных авторами данного исследования был рассмотрен ряд библиотек автоматизированного машинного обучения. В целом, что нейронные сети, что классические методы машинного обучения справляются с задачей, однако обозначенной осталась проблема подбора оптимальных параметров для различных условий, будь то подбор k-ближайших соседей, глубина случайного леса, количество слоев и связи в архитектуре нейронной сети, а также прочие характеристики, связанные с формированием алгоритма. Помимо этого, дополнительную наибольшую сложность несет экспертный выбор подходящего алгоритма машинного обучения или топологии нейронной сети под те или иные данные, что может быть затруднительным и требует экспертного мнения в области машинного обучения. Потребность в обработке сигналов возникает в различных сферах, таких как обработка изображений, акустических данных, временных рядов. Для решения всех вышеперечисленных проблем, в контексте поставленной ранее задачи, необходимо использовать программное обеспечение, основанное на технологии AutoML.

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации "Госзадание" №075-01024-21-02 от 29.09.2021 (проект FSEE-2021-0014).

II. Обзор существующих решений

Фреймворк MLJAR AutoML [1] включает такие функции, как:

1. автоматический анализ данных – анализ полей загруженного набора данных;
2. выбор алгоритма машинного обучения и настройка гиперпараметров;
3. создание отчетов с подробной информацией обо всех моделях;
4. сохранение, перезапуск и загрузка моделей.

Данный фреймворк поддерживает 4 режима работы:

1. режим анализа набора данных;
2. построение последовательностей методов машинного обучения;
3. соревновательный режим, в котором обучаются комбинации хорошо настроенных моделей машинного обучения;
4. режим «Оптуна», который может использоваться для поиска хорошо гиперпараметров моделей машинного обучения.

Фреймворк использует различные методы оптимизации для поиска оптимальных гиперпараметров, и один из них – поиск путем восхождения к вершине. Он заключается в итеративной максимизации целевой функции f(x). На каждой итерации восхождения изменяется один элемент в x и определяется, улучшают ли внесенные коррективы значение f(x).

Фреймворк TPOT строит комбинации методов предварительной обработки и машинного обучения и подбирает оптимальные гиперпараметры для этих комбинаций. Для этого метод использует генетический алгоритм [2]. Он основан на концепции естественного отбора, в которой наиболее приспособленные особи отбираются для воспроизводства, чтобы произвести потомство следующего поколения. Для решения задачи оптимизации генетический алгоритм рассматривает множество решений задачи и выбирает набор лучших решений. Генетический алгоритм рассматривает пять этапов:

1. начальная популяция;
2. целевая функция – функция для оценки пригодности того или иного решения;

3. селекция – отбор наиболее приспособленных особей для формирования следующего поколения;

4. скрещивание (кроссовер) – наиболее важная фаза генетического алгоритма; для каждой пары родителей, подлежащих скрещиванию, случайным образом выбирается точка кроссовера в генах; потомство создается путем обмена генами родителей между собой до достижения точки кроссовера; пример образования потомства при кроссовере, равном 3, представлен в (1)

$$(0,0,0,0,0,0) + (1,1,1,1,1,1) \rightarrow (1,1,1,0,0,0), \quad (0,0,0,1,1,1) \tag{1}$$

5. мутация: в некоторых из образовавшихся новых потомков некоторые из их генов могут претерпеть мутацию с небольшой случайной вероятностью; пример мутации дается (2):

$$(1,1,1,0,0,0) \rightarrow (1,1,0,1,1,1) \tag{2}$$

Фреймворк AutoKeras [3] позволяет подобрать оптимальную архитектуру и параметры нейронных сетей. Для этого используется Байесовская оптимизация – метод оптимизации «черного ящика» с шумом. Этот метод основан на итеративном варьировании параметров с целью выявления дополнительной информации об оптимизируемой функции, а также определения потенциального оптимального места. AutoKeras использует реализацию байесовской оптимизации в виде гауссовского процесса.

Модификация нейронной сети происходит следующим образом: сетевые слои f_a сопоставляются со слоями f_b. На основе результатов сопоставления слои f_a модифицируются, чтобы быть более похожими на слои f_b. Архитектура f_a сначала расширяется, а затем в расширенную архитектуру вставляется новый узел.

Для того чтобы определить, как должна быть изменена нейронная сеть, минимизируется функция, представленная в (3):

$$D_1(L_a, L_b) = min \sum_{i=1}^{L_a} \vee\, d_l\left(l_a^{(i)}, \varphi_l\left(l_a^{(i)}\right)\right) + \left||L_b| - |L_a|\right|, |L_a| < L_b \tag{3}$$

где, D_1 – метрика для вычисления оптимального преобразования слоев двух архитектур f_a и f_b; $\varphi_i - L_a \rightarrow L_b$ инъективная функция сопоставления слоев $\left(\forall i < j, \varphi_l\left(l_a^{(i)}\right) \prec \varphi_l\left(l_a^{j}\right)\right)$, если слои в L_a и в L_b отсортированы в топологическом порядке; d_l – расстояние редактирования, вычисляемое по (4):

$$d_l(l_a, l_b) = w(l_a) - w(l_b) \vee ma\,x[w(l_a) - w(l_b)], \tag{4}$$

где *w(l)* – ширина слоя *l*.

В исследовании [3] говорится, что в отличие от существующих сред AutoML, которые в основном сосредоточены на выборе модели / гиперпараметров, AutoGluon-Tabular успешно объединяет несколько моделей и размещает их в нескольких слоях. Эксперименты показывают, что послойное объединение множества моделей позволяет лучше использовать выделенное время на обучение, чем поиск лучших моделей.

Тесты набора из 50 задач классификации и регрессии из источников Kaggle и OpenML AutoML Benchmark показывают, что фреймворк AutoGluon быстрее, надежнее и намного точнее других алгоритмов. Результаты сравнения представлены в таблице 1. В списке указано количество наборов данных, на которых каждый из фреймворков выдал:

- лучшие прогнозы, чем AutoGluon (победы);
- худшие прогнозы (проигрыши);
- сбои системы во время обучения (неудачи);
- более точные прогнозы, чем все остальные 5 фреймворков (чемпионы).

Последние 3 столбца показывают среднее значение: ранг фреймворка (среди 6 фреймворков AutoML, примененных к каждому набору данных), (пересчитанные) потери на тестовых данных и фактическое время обучения (лимит времени 4 часа).

Преимущества и недостатки подходов представлены в табл. 1.

ТАБЛИЦА I. СРАВНЕНИЕ ФРЕЙМВОРКОВ AUTOML НА 11 СОРЕВНОВАНИЯХ KAGGLE

Фреймворк	+	-	О	Ч	Средний ранг	Средний %	Среднее время, мин
AutoGluon	-	-	0	7	1.7143	0.7041	202
GCP-Tables	3	7	1	3	2.2857	0.6281	222
H2O AutoML	1	7	3	0	3.4286	0.5129	227
TPOT	1	9	1	0	3.7143	0.4711	380
auto-sklearn	3	8	0	1	3.8571	0.4819	240
Auto-WEKA	0	10	1	0	6.0000	0.2056	221

где, "+" – число побед, "-" – число поражений, "О" – число отказов и "Ч" – число чемпионств.

В [4] представлен сравнительный анализ инструментов контролируемого машинного обучения. Сначала анализируются характеристики восьми инструментов машинного обучения с открытым исходным кодом (Auto-Keras, Auto-PyTorch, Auto-Sklearn, AutoGluon, H2O AutoML, rminer, TPOT и TransmogrifAI) и описываются 12 популярных наборов данных OpenML, используемых в тестировании, разделенных на задачи регрессии, бинарной и многоклассовой классификации. Затем проводится сравнительное исследование с сотнями вычислительных экспериментов на основе трех сценариев: общего машинного обучения, глубокого обучения и XGBoost. В ходе работы было установлено, что фреймворки H2O AutoML и AutoGluon дают наилучший предсказательный эффект в значительной части сценариев.

В [5] был проведен сравнительный анализ подходов AutoML для классификации сигналов электрокардиограммы (ЭКГ). Авторы протестировали

следующие фреймворки на несбалансированном наборе данных:

- auto-sklearn;
- AutoKeras;
- Tree-Based Pipeline Optimization Tool (TPOT).

Предварительная обработка данных и выбор признаков осуществлялись с помощью встроенных методов. Для количественной оценки влияния количества используемых электродов на конечную эффективность классификации рассчитывалась кумулятивная сумма оценок, полученных для каждой электродной системы. Модели обучались в течение 70 часов и получили следующие оценки на трех отведениях ЭКГ (табл. 2).

ТАБЛИЦА II. СРАВНЕНИЕ ФРЕЙМВОРКОВ AUTOML НА ДАННЫХ ЭКГ

Фреймворк	Результат
auto-sklearn	0.32
AutoKeras	0.33
TPOT	0.35

III. НАБОР ДАННЫХ

Для исследования были выбраны наборы данных содержащие различные ряды сигналов. Сравнение наборов представлено в табл. 3.

ТАБЛИЦА III. СРАВНЕНИЕ НАБОРОВ ДАННЫХ

Набор данных	Признаки	Отклики (классы)
Sonar: Набор данных для классификации подводных препятствий [6]	Энергия в заданных 60 полосах частот	мина камень
Doppler: Набор данных для распознавания автомобиля, человека, дрона [7]	Признаки матрицы радара (по вертикали – допплеровские частоты, dBm, по горизонтали – ячейки расстояния)	автомобиль человек дрон
Winnipeg: Набор данных для классификации пахотных земель [8]	Данные радара, радио-метрическая информация	кукуруза горох рапс соя овес пшеница широко-лиственные культуры

В наборе данных [6] содержится два класса: камень и подводная мина. В качестве признаков набор данных содержит показатели отклика (энергия) для 60 отдельных частот сонара. Набор данных включает в себя 208 наблюдений и 60 признаков. Количество наблюдений для класса «камень» (rock – R) равно 97, для класса «подводная мина» (mine – M) – 111.

В наборе данных [7] представлены измерения данных, полученных от радиолокационной системы. Набор содержит 17485 наблюдений. В наборе данных содержится 5720 наблюдений для класса «автомобиль», 6700 наблюдений для класса «человек» и 5065 наблюдений для класса «дрон».

Набор данных [8] содержит двухвременные данные оптического радара для классификации пахотных земель в табличной форме, полученные из изображений, собранных спутниками RapidEye, и поляриметрической радиолокационной информации, собранной беспилотными летательными аппаратами над сельскохозяйственным регионом недалеко от Виннипега (Канада), Набор включает в себя 325834 наблюдений и 7 классов сельскохозяйственных культур.

IV. РЕЗУЛЬТАТЫ

На основе собранных данных была произведена проверка фреймворков AutoML. В табл. 4 представлены результаты оценки фреймворков AutoML на тестовых выборках данных. В первой строке указаны наборы данных, в первом столбце – библиотеки AutoML.

Расчет производился по метрике *Accuracy* в силу сбалансированности целевых классов в представленных выше наборах данных.

Оценка методов производилась на тестовой выборке. Соотношение обучающей и тестовой выборок составило 80 к 20.

ТАБЛИЦА IV. ТОЧНОСТЬ БИБЛИОТЕК AUTOML НА ТЕСТОВЫХ НАБОРАХ ДАННЫХ

	Sonar	Doppler	Winnipeg	Среднее значение Accuracy
H2O	0.85	0.93	0.94	0.91
AutoKeras	0.85	0.91	0.95	0.91
MLBox	0.83	0.86	0.83	0.84
Auto-Sklearn	0.78	0.95	0.9	0.88
TPOT	0.85	0.92	0.85	0.87
AutoGluon	0.91	0.97	0.98	0.96
MLJAR AutoML	0.79	0.92	0.96	0.9

Исходя из данных таблицы, определено лучшее значение Accuracy для каждого набора данных. Самый низкий результат показал фреймворк MLBox, самый высокий – AutoGluon. В целом, данные точности отличаются незначительно для различных фреймворков.

Поскольку оценка велась не на реальных данных, а только на подобранных наборах данных из открытых источников, при оценке фреймворков было решено принимать во внимание не только точность, но и базовые аспекты функционала, удобство пользования, а также множество имплементированных во фреймворк методов машинного обучения и метрик их оценки.

V. ЗАКЛЮЧЕНИЕ

Существующие методы автоматического машинного обучения были оценены для выбранных данных. В качестве наиболее перспективных методов для дальнейшего анализа были выбраны фреймворки MLJAR AutoML, AutoKeras и TPOT. Их выбор был обусловлен возможностью работы как с традиционными методами машинного обучения, так и с нейронными сетями; возможностью работы как в автоматическом, так и в ручном режиме; реализацией большого количества различных методов машинного обучения, а также метриками для оценки их качества.

Список литературы

[1] MLJAR Automated Machine Learning for Humans [Электронный ресурс]. – Режим доступа: https://github.com/mljar/mljar-supervised (дата обращения: 10.03.2023).

[2] Mitchell, M. (1998). An introduction to genetic algorithms. MITpress.

[3] Jin, H., Song, Q., & Hu, X. (2019, July). Auto-keras: An efficient neural architecture search system. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 1946-1956).

[4] Ferreira L. et al. A comparison of AutoML tools for machine learning, deep learning and XGBoost //2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021. С. 1-8.

[5] Bodini, Matteo & Rivolta, Massimo & Sassi, Roberto. (2021). Classification of ECG Signals with Different Lead Systems Using AutoML. 1-4. 10.23919/CinC53138.2021.9662802.

[6] Sonar Dataset suitable for classification [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/datasets/mortrest/sonar-dataset-suitable-for-classification, свободный. Яз. англ. (дата обращения 25.03.2022).

[7] RealDopplerRAD-DARdatabase [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/datasets/iroldan/real-doppler-raddar-database, свободный. Яз. англ. (дата обращения 25.03.2022).

[8] CroplandMapping [Электронный ресурс]. – Режим доступа: https://www.kaggle.com/datasets/pcbreviglieri/cropland-mapping, свободный. Яз. англ. (дата обращения 25.03.2022).