Построение модели прогнозирования временных рядов средствами автоматизированного машинного обучения

В. Э. Ковалевский

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

darkeol@ieee.org

Н. А. Жукова

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

nazhukova@mail.ru

Аннотация. **Область машинного обучения включает в себя различные алгоритмы, с помощью которых возможно построение моделей, позволяющих извлекать знания из большого объема данных. Отдельной задачей машинного обучения является задача прогнозирования временных рядов. Особенность этой задачи заключается в том, что в этом случае необходимо учитывать взаимосвязь измерений со временем, а не только разнообразие и другие статистические характеристики выборки. Выбор подходящего к конкретным данным алгоритма и настройка его гиперпараметров являются непростой задачей. Автоматизация данной задачи относится к сфере автоматизированного машинного обучения (AutoML). Большинство AutoML инструментов решают задачу автоматизированного построения моделей для классификации и прогнозирования данных на основе атрибутов, без учета изменений параметра во времени. Однако уже появилось и несколько систем позволяющих автоматизированное построение моделей для задачи прогнозирования временных рядов.**

В данной работе мы рассматриваем AutoML системы, работающие с временными рядами. Приводится процесс автоматизированного создание модели для оценки временного ряда с помощью системы AutoGluon, и оценивается влияние различных параметров данной системы на получаемую модель.

Ключевые слова: автоматизированное машинное обучение, машинное обучение, прогнозирование временных рядов

I. Введение

Область машинного обучение включает в себя различные алгоритмы, с помощью которых возможно построение моделей, позволяющих извлекать знания из большого объема данных [1]. Такие модели, в частности позволяют решать задачи классификации и прогнозирования на основе исторических данных. Используя большой набор данных можно обучить модель, применяя в качестве входных данных значения атрибутов. Отдельно стоящей задачей в этом списке является задача прогнозирования временных рядов. Особенность этой задачи заключается в том, что, во-первых, в этом случае необходимо учитывать взаимосвязь измерений со временем, а не только разнообразие и другие статистические характеристики выборки, а во-вторых, тот же самый атрибут является и входным, и целевым.

Каждый алгоритм машинного обучения, на основе которого строится модель, имеет свои гиперпараметры - значение задающие его работу. Выбор подходящего алгоритма для создания модели и настройка его гиперпараметров являются непростой задачей. Для того чтобы автоматизировать данную трудоемкую работу был предложен ряд решений, получивших общее название Автоматизированного машинного обучения (Automated Machine Learning – AutoML)[2]. Большинство AutoML инструментов решают задачу автоматизированного построения моделей для классификации и прогнозирования данных, в которых не учитывается взаимосвязь параметров со временем. Однако было разработано и несколько систем позволяющих автоматизированное построение моделей для задачи прогнозирования временных рядов.

II. Прогнозирование временных рядов

Временной ряд – это собранные с равными промежутками времени значения каких-либо параметров, в простейшем случае одного параметра. Каждое значение называется измерением, отсчётом, либо уровнем на указанный момент времени. Во временном ряде для каждого отсчёта должно быть указано время измерения или номер измерения по порядку. Временной ряд существенно отличается от простой выборки данных, так как при анализе учитывается взаимосвязь измерений со временем, а не только статистическое разнообразие и статистические характеристики.

Анализ временный рядов – это совокупность математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и их прогнозирования [3]. Выявление структуры временного ряда служит основой для построения математической модели процесса, являющегося источником анализируемого временного ряда. Прогноз будущих значений временного ряда помогает эффективному принятию решений.

Примерами временных рядов являются:

- электрическая активность мозга;
- измерения количества осадков;
- цены на акции;

Работа выполнена в рамках бюджетной темы FFZF-2022-0006

- годовые розничные продажи;
- количество ударов сердца в минуту.

Особенностью временных рядов являются:

- зависимость данных от времени;
- необходимость равного интервала измерений;
- невозможность менять порядок следования данных;
- использование тех же данных в качестве входных и целевых значений.

III. AutoML решения для прогнозирования временных рядов

В настоящие время существует ряд решений, предназначенных для автоматизированного построения моделей прогнозирования временных рядов. Некоторые из этих решений являются расширением существующих AutoML подходов к задаче прогнозирования временных рядов, другие являются специализированными инструментами. К первым относятся такие решения, как AutoGluon, AutoKeras, Auto-PyTorch и FEDOT имеющие в своем арсенале как инструменты для построения моделей на основе данных, не имеющих зависимость от времени, так и для временных рядов. Ко вторым носится система AutoTS предназначенная исключительно для построения моделей для временных рядов.

Система AutoGluon[4] позволяет прогнозировать будущие значения нескольких временных рядов с учетом исторических данных и других связанных ковариат. AutoGluon сочетает использование различных алгоритов машинного обучения, таких как ETS и ARIMA, LightGBM, а также моделей глубокого обучения, таких как DeepAR и Temporal Fusion Transformer. AutoGluon основан на библиотеке GluonTS, которая в свою очередь основывается на библиотеках PyTorch и MXNet.

AutoKeras[5] является AutoML системой, производящей поиск по алгоритмам глубокого обучения из библиотеки Keras, которая базируется на библиотеке машинного обучения TensorFlow от компании Google.

Система Auto-PyTorch[6] также производит поиск подходящих моделей среди алгоритмов из библиотеки PyTorch. Её особенностью является использование Портфолио – набора заранее отобранных перспективных конфирураций нейронных сетей служащих отправной точкой для дальнейших поисков.

AutoTS[7] это специализированная AutoML система, предназначенная для построения моделей для временных рядов. Данная система производит поиск по алгоритмам из целого ряда библиотек, в частности StatsModels, GluonTS, Sklearn и TensorFlow. Выбор подходящей модели основан на использовании генетического алгоритма.

Ещё одним AutoML решением, использующим для отбора моделей эволюционный подход, является система FEDOT[8].

Сравнение указанных AutoML решений приведено в табл. 1.

ТАБЛИЦА I. Сравнение AutoML решений

Решение	**Библиотека**	**ОС**	**Распред.вычисл**
AutoGluon	GluonTS, PyTorch, MXNet	Linux, MacOS, Windows	На GPU
AutoKeras	Keras, TensorFlow	Linux, Windows	На GPU
Auto-PyTorch	PyTorch	Linux	-
AutoTS	StatsModels, GluonTS, Scikit-Learn, TensorFlow	Linux, Windows	На CPU
FEDOT		Linux, MacOS, Windows	На GPU

Для проверки возможностей построения моделей прогнозирования временных рядов средствами AutoML была выбрана система AutoGluon, для проведения на ней экспериментов с использованием подходящего набора данных.

IV. Данные

A. *Исходные данные*

Открытые репозитории данных, такие как Kaggle[9] и OpenML[10] содержат большое количество свободно распространяемых наборов данных предназначенных для обучения и тестирования различных подходов машинного обучения. Однако лишь малая часть этих данных является временными рядами и может быть использована для проверки соответствующих алгоритмов. Существуют также небольшие репозитории, в которых хранятся исключительно данные временных рядов, как например Time Series Classification[11].

При выборе набора данных для целей тестирования предпочтение отдавалось большим наборам данных (>100 000 экземпляров) содержащим одновременно несколько временных рядов. В качестве таких данных был выбран набор Daily Temperature of Major Cities содержащий сведения об изменении температуры в различных городах мира. Данный набор содержит информацию о средней температуре по Фаренгейту 7 регионов (Африка, Азия, Австралия, Европа, Ближний Восток, Северная Америка, Южная/Центральная Америка), 53 штатах США, 321 городе и имеет следующие характеристики:

- количество экземпляров – 2906327;
- количество атрибутов – 8;
- интервал – 1 день;
- сроки измерения – 26 лет (1995–2020).

B. *Предобработка данных*

Используемый набор данных был предварительно очищен. Были удалены строки, содержащие некорректные даты. Также с помощью Z-оценки были убраны данные содержащие выбросы. Это уменьшило набор данных на 83990 записей. Пример измерений температуры в трех странах в течение двух лет приведен на рис. 1. На рисунке видно, как меняется средняя

температура в указанных странах в зависимости от месяца.

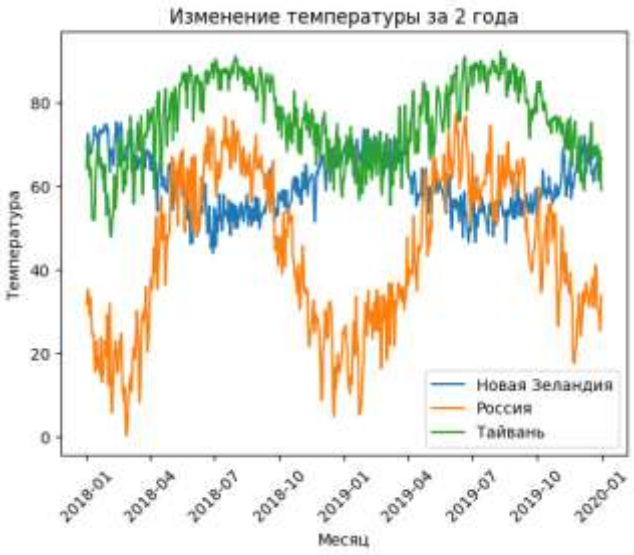

Рис. 1. Данные об изменении температуры в трех странах

Затем данные были приведены к виду, подходящему для использования в рамках выбранной AutoML системы. Был удален столбец «State», содержащий пустые значения за пределами США, а столбцы, содержащие по отдельности значения дня, месяца и года были объединены в один столбец «Date». Поскольку система AutoGluon требует, чтобы входные данные имели числовой формат, то категориальные столбцы «Region» и «Country» были преобразованы с помощью функции LabelEncoder из библиотеки Scikit-Learn.

C. Дополнительные преобразования

Для работы AutoGluon требуется задать столбец, содержащий значения даты-времени и столбец с идентификаторами временных рядов, в случае если набор данных содержит несколько. В качестве столбца даты был выбран сформированный столбец «Date», а в качестве идентификатора временного ряда был выбран столбец «City», таким образом, было объявлено, что измерения температуры в каждом городе представляют собой отдельный временной ряд.

V. Эксперименты

В качестве начальных установок системе AutoGluon необходимо указать целевой столбец, содержащий измеряемое значение, определить горизонт прогнозирования, и задать метрику, по которой решения будут оцениваться. В случае если шаг временного ряда не был распознан автоматически, то его можно указать вручную. В качестве целевого столбца был выбран столбец содержащий значение температуры, а интервалом шага указан день.

A. Предустановки

Процесс поиска также дополнительно настраивается заданием максимального времени, которое отводится на поиск, и заданием предустановок, влияющих на качество поиска. Система предлагает четыре варианта предустановок, см. табл. 2.

ТАБЛИЦА II. Предустановки системы AutoGluon

Предустановка	**Описание**	**Особенности**
fast_training	Подбор простых статистических и базовых моделей + быстрые модели на основе деревьев	Быстрое обучение, но может быть не точным
medium_quality	Как предыдущий пункт, но дополнительно использование модели глубокого обучения TemporalFusion Transformer	Хорошие прогнозы с умеренным временем обучения
high_quality	Использование более мощных моделей глубокого обучения, машинного обучения и статистического прогнозирования.	Точнее, чем medium_quality, но обучение занимает больше времени.
best_quality	Те же модели, что и в предыдущем пункте, но больше окон перекрестной проверки.	Обычно более точный, чем high_quality, особенно для наборов данных с небольшим количеством временных рядов.

B. Метрики оценки

Для оценки моделей в процессе поиска AutoGluon может использовать следующие метрики, см. табл. 3.

ТАБЛИЦА III. Метрики оценки

Метрика	Название (англ)	Название (рус)
SQL	Scaled quantile loss	Масштабированная квантильная потеря
WQL	Weighted quantile loss	Взвешенная квантильная потеря
MAE	Mean absolute error	Средняя абсолютная ошибка
MAPE	Mean absolute percentage error	Средняя абсолютная ошибка в процентах
MASE	Mean absolute scaled error	Средняя абсолютная масштабированная ошибка
MSE	Mean squared error	Среднеквадратичная ошибка
RMSE	Root mean squared error	Корень среднеквадратичной ошибки
RMSSE	Root mean squared scaled error	Корень среднеквадратичной масштабированной ошибки
SMAPE	Symmetric mean absolute percentage error	Симметричная средняя абсолютная процентная ошибка
WAPE	Weighted absolute percentage error	Взвешенная абсолютная процентная ошибка

Для вычисления данных метрик используется следующая информация об обрабатываемых временных рядах и прогнозируемых значениях:

$y_{i,t}$ – значение временного ряда i, в момент времени t;

$f_{i,t}$ – прогнозируемое значение временного ряда i, в момент времени t;

N – количество временных рядов в наборе данных;

T – длина временного ряда;

H – горизонт планирования.

Для целей эксперимента был осуществлен поиск моделей с заданием горизонта планирования 60 (2 месяца) и использованием различных предустановок и метрик оценки. Были использованы следующие метрики MAE, RMSE, WQL.

$$MAE = \frac{1}{N}\frac{1}{H}\sum_{i=1}^{N}\sum_{t=T+1}^{T+H}|y_{i,t} - f_{i,t}|$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\frac{1}{H}\sum_{i=1}^{N}\sum_{t=T+1}^{T+H}(y_{i,t} - f_{i,t})^2}$$

$$\text{WQL} = \frac{1}{\sum_{i=1}^{N}\sum_{t=T+1}^{T+H}|y_{i,t}|}\sum_{i=1}^{N}\sum_{t=T+1}^{T+H}\sum_{q} p_q(y_{i,t}, f_{i,t}^{q})$$

где $f_{i,t}^{q}$ – прогнозируемый квантиль q временного ряда i в момент времени t;

$p_q(y,f)$ – квантильная потеря на уровне q.

Поиск моделей был проведен с лимитом времени 5, 10, 15 и 30 мин. Результаты экспериментов приведены в табл. 4.

ТАБЛИЦА IV. СРАВНЕНИЕ ПРЕДУСТАНОВОК

Установка	Время	MAE	RMSE	WQL
fast_training	300	5.9093	7,7755	0.0741
fast_training	600	5.9929	7,8625	0.0747
fast_training	900	6.0363	7,8827	0.0754
fast_training	1800	6.0363	7.8827	0.0753
medium_quality	300	5.3815	7,3873	0.0986
medium_quality	600	5.3626	7,3514	0.0782
medium_quality	900	**5.3540**	7,3326	**0.07**
medium_quality	1800	5.3829	7.3205	0.0703
high_quality	300	5.9197	7,4377	0.1153
high_quality	600	5.4345	7,3975	0.0967
high_quality	900	5.4202	7,3617	0.0837
high_quality	1800	5.4057	**7.3162**	0.07
best_quality	300	8.9054	11,601	0.1158
best_quality	600	5.8145	7,8126	0.1158
best_quality	900	5.4362	7,4078	0.1158
best_quality	1800	5.4096	7.3562	0.0842

Пример графика прогноза для Москвы, сформированного моделью найденной за 15 минут с использованием предустановки medium_quality и метрикой оценки MAE приведен на рис. 2.

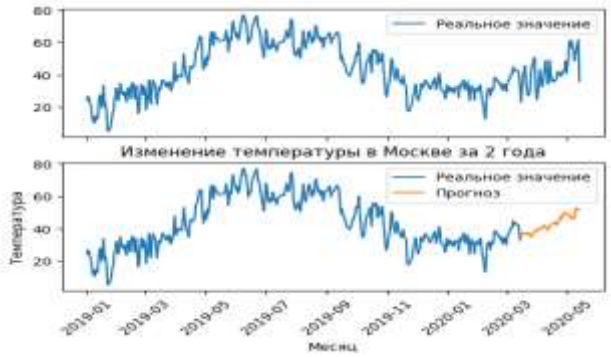

Рис. 2. Реальное и прогнозируемое значение временного ряда

VI. ЗАКЛЮЧЕНИЕ

В результате проведенного исследования была показана применимость моделей, построенных с помощью автоматизированного машинного обучения к задаче прогнозирования временных рядов.

Для экспериментальных исследований была использована система автоматизированного машинного обучения AutoGluon и набор данных Daily Temperature of Major Cities, содержащий данные о температуре в 321 городах мира.

По результатам проведенных экспериментов было выявлено, что для данного набора данных наиболее эффективной метрикой оценки модели является WQL, а предустановка fast_training показывает ухудшение результатов при увеличении лимита времени. При заданных лимитах времени, не больше 30 минут на поиск, предустановкой, дающей наиболее точную модель, является medium_quality.

СПИСОК ЛИТЕРАТУРЫ

[1] Бишоп К.М. Распознавание образов и машинное обучение = Pattern Recognition and Machine Learning / пер. с англ. и редакция Д. А. Клюшина. Санкт-Петербург: Диалектика, 2020. 960 с.

[2] He X. AutoML: A Survey of the State-of-the-Art / X. He, K. Zhao, X. Chu // Knowledge-Based Systems. 2021. Vol. 212. Art. No. 106622. 27 p. DOI: 10.1016/j.knosys.2020.106622.

[3] Мишулина О.А. Статистический анализ и обработка временных рядов. М.: МИФИ, 2004. С. 180. ISBN 5-7262-0536-7.

[4] Erickson, Nick, et al. / AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data.// arXiv preprint arXiv:2003.06505 (2020)

[5] AutoKeras: An AutoML Library for Deep Learning / H. Jin, F. Chollet, Q. Song, X. Hu // Journal of Machine Learning Research. 2023. Vol. 24. Art. No. 6, 6 p.

[6] Auto-Pytorch: Multi-Fidelity MetaLearning for Efficient and Robust AutoDL / Zimmer L., Lindauer M., & Hutter F. (n.d.) // IEEE Transactions On Pattern Analysis and Machine Intelligence, vol. 43, no. 9, pp. 3079-3090, 1 Sept. 2021, doi: 10.1109/TPAMI.2021.3067763

[7] AutoTS - time series package for Python, https://github.com/winedarksea/AutoTS (дата обращения 11.03.2024)

[8] Automated machine learning approach for time series classification pipelines using evolutionary optimization / Revin I., Potemkin V.A., Balabanov N. R., Nikitin N. O.//Knowledge-Based Systems, vol. 268, 2023, ISSN 0950-7051

[9] Kaggle: Your Machine Learning and Data Science Community. URL: http://www.kaggle.com (дата обращения 27.04.2023).

[10] OpenML. A worldwide machine learning lab. https://www.openml.org (дата обращения 27.04.2023).

[11] Time Series Classification Website. https://www.timeseriesclassification.com (дата обращения 27.04.2023)