ISSN 2658-5014 (Print)
2021. Т. 3, № 1 (5). С. 50–57

СИИТ
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

ISSN 2686-7044 (Online)
http://siit.ugatu.su

УДК 004.89

АВТОМАТИЗИРОВАННОЕ МАШИННОЕ ОБУЧЕНИЕ: ОБЗОР ВОЗМОЖНОСТЕЙ СОВРЕМЕННЫХ ПЛАТФОРМ АНАЛИЗА ДАННЫХ

И. П. Болодурина[1], Д. И. Парфенов[2], А. Е. Шухман[3], Л. С. Забродина[4]

[1] prmat@mail.osu.ru, [2]parfenovdi@mail.ru, [3]shukhman@gmail.com, [4]zabrodina97@inbox.ru

Оренбургский государственный университет

Поступила в редакцию 10 марта 2021 г.

Аннотация. Методы автоматического машинного обучения (AutoML) играют важнейшую роль в работе с огромным объемом данных и используются практически во всех возможных областях. Использование инструментов AutoML в прикладных задачах анализа данных позволяет извлекать новые знания из исходной информации, выявлять взаимосвязи внутри данных и решать задачи классификации, кластеризации, регрессии, прогнозирования и др. В рамках данной работы проводится обзор существующих подходов и систем построения гибко настраиваемых конвейеров AutoML, использующих интеллектуальные алгоритмы оптимизации структуры и параметров. Для расширения применяемых подходов байесовской оптимизации в рамках реализации процесса автоматического машинного обучения, включен этап метаобучения, а также автоматизированного построения ансамбля для повышения эффективности получаемых результатов. Формализуются этапы процесса оптимизации конвейеров данных и настройки алгоритмов машинного обучения, а также сформулирована постановка задачи оптимизации выбора комбинированного алгоритма и настройки гиперпараметров (CASH). Задача CASH является важнейшим элементом систем AutoML, от методов и алгоритмов решения которой зависит производительность и эффективность конечных моделей обучения. В связи с этим, следующим этапом данного исследования является разработка и модификация подходов к решению данной задачи, а также планируется подобрать наиболее эффективные алгоритмы построения признакового пространства и модифицировать их для повышения производительности и точности обученных моделей.

Ключевые слова: автоматизированное машинное обучение; анализ данных; байесовская оптимизация; интеллектуальные алгоритмы оптимизации; метаобучение.

ВВЕДЕНИЕ

В последнее время машинное обучение становится все более актуальным: автомати-

Работа поддержана грантом РФФИ (проект № 20-07-01065), грантом Президента Российской Федерации по государственной поддержке ведущих научных школ (НШ-2502.2020.9), а также стипендии Президента Российской Федерации молодым ученым и аспирантам (СП-3652.2021.5).

ческое распознование речи, самостоятельное управление автомобилями или профилактическое обслуживание сетей основаны на методах и алгоритмах интеллектуального анализа данных.

Для построения необходимого для решения подобных задач конвейера машинного обучения необходима высококвалифицированная команда специалистов не только со знаниями предметной области и опытом работы, но и с навыками в сфере анализа данных. В рамках совместной работы таких специалистов возможно построить эффективный конвейер машинного обучения, включающий специализированную предварительную обработку данных, разработку значимых функций на основе предметной области и тонко настроенные модели, обеспечивающие высокую предсказательную силу. Как правило, процесс построения подобного конвейера представляет собой очень сложную задачу, решаемую многократно, методом проб и ошибок. В связи с этим, создание эффективных конвейеров машинного обучения требует длительного времени, и на практике довольно часто используют неоптимальную конфигурацию конвейера по умолчанию.

Автоматическое машинное обучение направлено на совершенствование существующих способов создания приложений анализа данных с помощью автоматизации. AutoML позволяет решать такую задачу, как оптимизация гиперпараметров, что приводит к повышению эффективности. Специалисты в предметной области получают возможность самостоятельно создавать конвейеры машинного обучения, не прибегая к изучению алгоритмов анализа данных.

Заметим, что на данный момент существует множество коммерческих решений AutoML, позволяющих решать задачи предварительной обработки данных, выбора соответствующих алгоритмов машинного обучения и оптимизации гиперпараметров. В большинстве случаев, исследуемые подходы объединяют в себе высокопараметрическую структуру машинного обучения с методом байесовской оптимизации для создания экземпляров для заданного набора данных.

В связи с этим, в рамках данной статьи проанализированы существующие современные платформы анализа данных, а также методы и алгоритмы AutoML, способные расширить.разработанные подходы и значительно улучшить их эффективность и надежность.

ОБЗОР ИССЛЕДОВАНИЙ ПОДХОДОВ AUTOML

Проблемой построения автоматизированной системы машинного обучения, позволяющей выполнять последовательную оптимизацию параметров, активно занимаются ученые со всего мира.

Большинство из существующих подходов автоматизации помещают проблему AutoML в жесткие рамки байесовской задачи оптимизации с фиксированным числом переменных решения [1, 2]. Как правило, существует переменная для алгоритма предварительной обработки, переменная для алгоритма обучения и переменная для каждого параметра каждого алгоритма. Хотя этот способ формализации проблемы AutoML приводит к пространству решения фиксированной размерности, он сопровождается значительной потерей структурной информации для поиска [3]. В рамках данной работы, большой интерес представляют гибко настраиваемые конвейеры машинного обучения, использующие интеллектуальные алгоритмы оптимизации структуры и параметров и решающие проблему масштабируемости. В рамках публикации [4] рассмотрена задача поиска эффективных методов Байесовской оптимизации решения проблемы AutoML. На основе анализа производительности конвейеров на аналогичных наборах данных и построении ансамблей из оцениваемых моделей во время оптимизации, авторы исследования разработали систему AUTO-SKLEARN. Новая система анализа данных продемонстрировала повышение производительности и эффективности на широком диапазоне наборов данных: 140 наборов данных бинарной и многоклассовой классификации из репозитория OpenML, не относящихся к классу больших

данных. В исследовании [5] авторы дополнили систему с помощью модифицированных алгоритмов метаобучения и методов обработки итеративных алгоритмов, а также разработали новую стратегию распределения вычислений.

Авторы исследования [6] разработали инструмент оптимизации конвейеров на основе деревьев (TPOT) с открытым исходным кодом и продемонстрировали его эффективность на серии смоделированных и реальных наборов данных. Основным достоинством предложенной системы AutoML является разработка достаточно сложных конвейеров путем интеграции метода оптимизации Парето, формирующего компактные конвейеры без ущерба для точности классификации. Наиболее полная оценка возможностей TPOT представлена в работе [7] и содержит экспериментальные исследования оптимизации конвейеров машинного обучения методами генетического программирования для автоматического построения компьютерных программ. Метод байесовской оптимизации для автоматического поиска в объединенном пространстве алгоритмов обучения WEKA и их соответствующих настроек гиперпараметров рассмотрен в работе [8] и получил название Auto-WEKA. Предложенная система AutoML отличается простотой построения и интерпретации результатов интеллектуального анализа данных. Определение наилучшей точности прогнозов моделей с помощью Auto-WEKA представлено в исследовании [9]. Результаты оценки средней относительной ошибки прогнозирования ремонтопригодности программного обеспечения показывают достоверность и эффективность применения Auto-WEKA в задачах AutoML.

Исследователи H. Jin, Q. Song и X. Hu в статье [10] предложили новую структуру Auto-Keras, позволяющую использовать байесовскую оптимизацию для управления морфизмом сети для эффективного поиска нейронной архитектуры и автоматической настройки глубоких нейронных сетей. При этом, платформа разрабатывает ядро нейронной сети и алгоритм оптимизации функции сбора данных с древовидной структурой для исследования пространства поиска. Экспериментальные исследования показали высокую производительность и эффективность разработанной системы.

В исследовании [11] используется эволюционный подход RECIPE на основе грамматики для развития проектирования конвейера анализа данных. Однако на практике данный подход до сих пор оценивался только на довольно небольших наборах данных, что, конечно, не исключает его полезности. Кроме того, использование суррогатных функций для ускорения оценки возможных решений позволит повысить его эффективность.

Сведение проблемы AutoML к задаче поиска графов посредством планирования иерархической сети задач (ИСЗ) использовалось при создании ML-Plan в публикации [12] коллегами из Падерборнского университета Германии. При этом алгоритм поиска наилучшего конвейера в графе основан на прямой декомпозиции задачи планирования ИСЗ. Подобные стандартные решатели ИСЗ, такие как SHOP2 [13] требуют разложить стоимость решения (плана) и оповестить о необходимых затратах, в то время как ML-Plan преодолевает это ограничение.Первоначально, использовать планирование ИСЗ для интеллектуального анализа данных в системе Meta-Miner было предложено в работах [14, 15].

Однако, вместо оценки конвейеров во время поиска, применялась стратегия «восхождения к вершине», где решения принимаются на основе прошлого опыта. Подобная конфигурация позволяет Meta-Miner выполнять поиск быстро за счет отсутствия точной оценки возвращаемого решения.

Основным недостатком предложенного подхода является возможность рассмотрения только небольшого подмножества параметров из-за комбинаторного перебора. Основные достоинства и недостатки существующих систем AutoML представлены в табл. 1 и показывают необходимость построения новой многофункциональной системы обучения.

Обзор проведенных исследований показал, что построение гибко настраиваемых конвейеров AutoML, использующих интеллектуальные алгоритмы оптимизации структуры и параметров, является сложной задачей и требует разработки новых нетривиальных подходов. Таким образом, в рамках данной статьи исследован и формализован двухэтапный процесс оптимизации для построения конвейеров данных и настройки алгоритмов машинного обучения. Кроме того, изучено влияние конвейеров данных для оценки важности предварительной обработки данных.

Таблица 1

Сравнительный анализ систем AutoML

Table 1

Comparative analysis of AutoML systems

	Алгоритм оптимизации	*Предобработка данных*	*Построение признаков*	*Выбор модели*	*Оптимизация гиперпараметров*	*Ансамблевое обучение*	*Мета-обучение*
Auto-WEKA	Байесовская оптимизация (SMAC)	+		+	+		
Auto-Sklearn	Совместная байесовская оптимизация и Bandit Search (BOHB)	+		+	+	+	+
TPOT	Эволюционный алгоритм	+	+	+	+		
TuPAQ	Bandit Search			+	+		
ATM	Совместная байесовская оптимизация и Bandit Search		+		+		+
ML-Plan	Иерархические сети задач	+		+	+		
Autostacker	Эволюционный алгоритм			+	+	+	
AlphaD3M	Обучение с подкреплением / Поиск по дереву Монте-Карло	+		+	+		

ПРОЦЕСС ДВУХЭТАПНОЙ ОПТИМИЗАЦИИ AUTOML

Задачу построения системы AutoML следует рассматривать с этапа предварительной обработки данных, а не с этапа выбора и настройки алгоритма. Связано это с тем, что на практике довольно сложно встретить готовые к анализу наборы данных и они должны быть преобразованы с помощью тщательно подобранной последовательности операций предварительной обработки.

В связи с этим, создание высококачественной модели машинного обучения для развертывания в производстве – это сложная задача, требующая много времени и вычислений. Опишем процесс автоматизации машинного обучения, представленный на рис. 1, с помощью двух этапов:

1. Нахождение такой правильной последовательности преобразований данных, чтобы набор данных можно было обрабатывать алгоритмом машинного обучения.

2. Выбор правильного алгоритма машинного обучения и его гиперпараметров такой, чтобы модель обеспечивала хорошее обобщение относительно заданной метрики производительности. Существует множество причин, требующих предварительной обработки исходных данных: большое количество независимых переменных, несбалансированный набор данных, пропущенные значения, выбросы, шум, специфические ограничения области алгоритмов и т.д. При этом, обработка данных оказывает огромное влияние на производительность модели. Конвейер данных зависит как от источника данных, так и от алгоритма, так что не существует универсального конвейера, который мог бы работать для каждого источника данных и каждого алгоритма. Как правило, конвейер данных определяется методом проб и ошибок, опираясь на опыт специалистов по обработке данных и экспертные знания о них. В связи с этим, данный шаг может занимать большое количество времени.

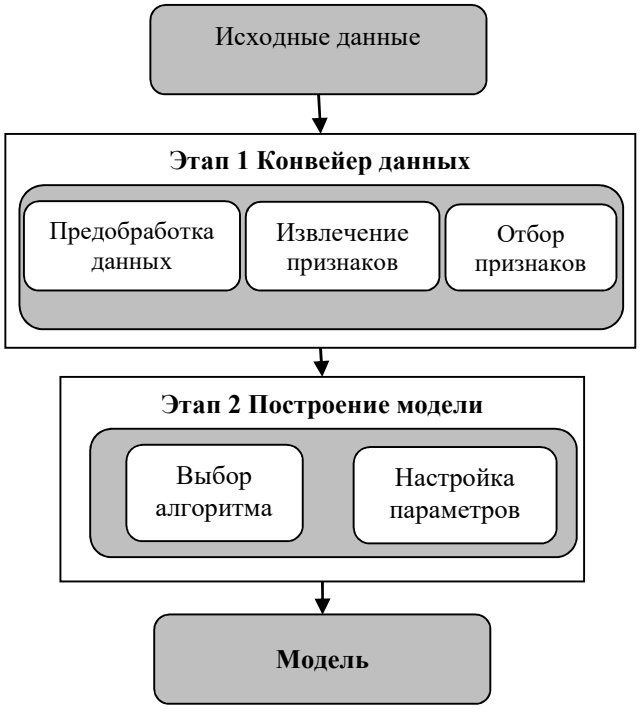

Рис. 1. Процесс двухэтапной оптимизации AutoML

Fig. 1. AutoML two-stepoptimization process

Обобщая полученные выше заключения, можно утверждать, что задача построения системы AutoML состоит в решении следующей задачи оптимизации черного ящика. На исходном наборе данных построить оптимальную конфигурацию λ* такую, что

$$\lambda^* \in \arg\max_{\lambda \in \Lambda} F(\lambda), \qquad (1)$$

где Λ – пространство конфигураций машинного обучения, а $F(\lambda)$ – производительность модели, обученной на наборе данных с использованием конфигурации λ.

Большинство существующих систем AutoML решают задачу (1) путем агрегирования операторов конвейера данных, набора алгоритмов и соответствующего им конфигурационного пространства в единое гигантское пространство поиска. Для того, чтобы расширить применяемый подход байесовской оптимизации в рамках реализации описанного процесса включим этап метаобучения, а также этап автоматизированного построения ансамбля для повышения эффективности получаемых результатов.

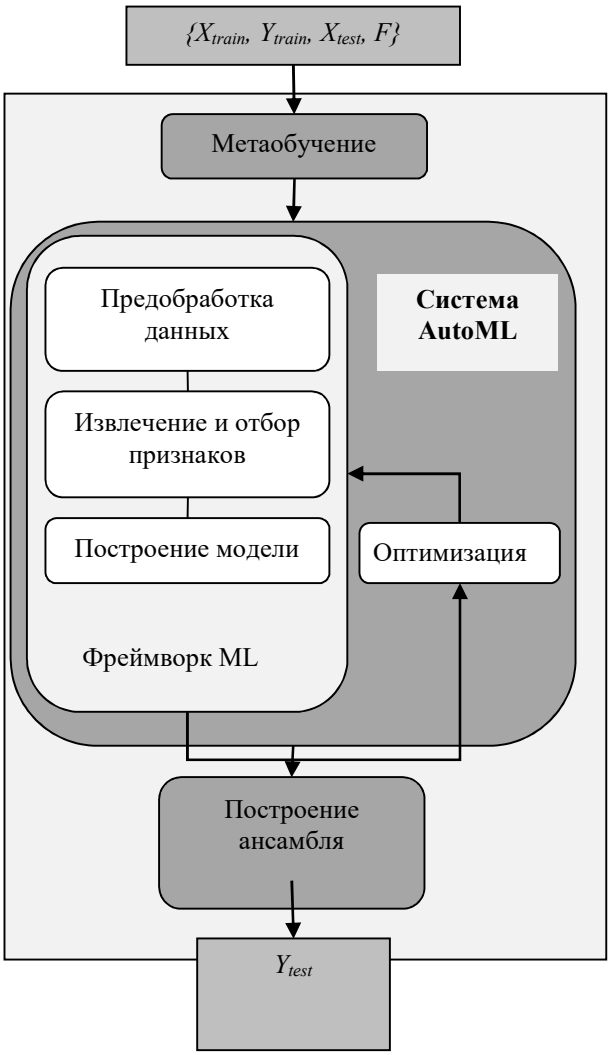

Рис. 2. Общий рабочий процесс AutoML

Fig. 2. General AutoML workflow

На рис. 2 показан общий рабочий процесс AutoML. Отметим, что представленная структура автоматизации машинного обучения в дальнейшем будет сравнена по производительности и точности с существующими гибкими фреймворками машинного обучения. Входными параметрами для процесса AutoML являются исходные данные $D = \{X_{train}, Y_{train}, X_{test}, Y_{test}\}$, разделенные на тестовый и обучающий набор, где X – признаковое пространство, Y – целевой признак. В рамках данной работы формально представим поиск решения такой задачи, как построение и конфигурирование конвейера данных, а также выбор алгоритма и его конфигурирование.

ФОРМАЛИЗАЦИЯ МАТЕМАТИЧЕСКОЙ ПОСТАНОВКИ ЗАДАЧИ AUTOML

Задача машинного обучения состоит в нахождении или построении аппроксимации неизвестной функции $f : X \to Y$, которая строит отображение признакового пространства в область целей.

Алгоритмом обучения A назовем отображение набора обучающих точек $D = \{d_i\}_{i=1}^{n}$ с $d_i = (x_i, y_i) \in X \times Y$ на множество Y^X. При этом алгоритм обучения A параметризуется некоторыми гиперпараметрами $\lambda \in \Lambda$, которые изменяют способ обучения алгоритма A_λ. Каждый гиперпараметр λ_i принадлежит пространству Λ_i и Λ является подмножеством векторного произведения каждой области, т.е. $\Lambda \subset \Lambda_1 \times ... \times \Lambda_n$. В общем случае Λ может быть более структурированным (дерево условий, направленный ациклический граф и т.д.).

Задача оптимизации выбора комбинированного алгоритма и настройки гиперпараметров.

Для заданного набора алгоритмов $A = \{A^{(1)}, ..., A^{(m)}\}$ с ассоциированными гиперпараметрическими пространствами $\Lambda^{(1)}, ..., \Lambda^{(m)}$ задача выбора комбинированного алгоритма и оптимизации гиперпараметров (CASH) определяется следующим образом:

$$A^*_{\lambda^*} \in \arg\max_{A^{(j)} \in A} \max_{\lambda \in \Lambda^{(j)}} \frac{1}{k} \sum_{i=1}^{k} F(A^{(j)}_\lambda, D^{(i)}_{train}, D^{(i)}_{test}), \quad (2)$$

где F – функция потерь (например, частота ошибок), полученная на тестовом наборе D_{test} моделью, обученной алгоритмом A с параметром $\lambda \in \Lambda$ на обучающем множестве D_{train}.

В предположении, что метаобучение и ансамблирование моделей являются независимыми, построенный процесс оптимизации дает следующие преимущества:

1. Уменьшив пространство поиска, повышается скорость общего процесса оптимизации.

2. Возможность статистически оценить, является ли конвейер данных специфическим для алгоритма или, скорее, универсальным для набора данных, обеспечивающий метаобучение на более низком уровне детализации.

Задача выбора комбинированного алгоритма и оптимизации гиперпараметров является важнейшим элементом систем AutoML, от методов и алгоритмов решения которой зависит производительность и эффективность конечных моделей обучения. В связи с этим, следующим этапом данного исследования является разработка и модификация подходов к решению задачи CASH.

ЗАКЛЮЧЕНИЕ

Таким образом, в рамках данной работы формализован процесс оптимизации построения конвейеров данных и настройки алгоритмов машинного обучения. Кроме того, изучено влияние конвейеров данных для оценки важности предварительной обработки данных, а также сформулирована постановка задачи оптимизации выбора комбинированного алгоритма и настройки гиперпараметров. В дальнейшем планируется подобрать наиболее эффективные алгоритмы выбора и построения признакового пространства, а также модифицировать их для повышения производительности и точности обученных моделей. Кроме того, предполагается разработка и модификация методов выбора алгоритмов обучения и методов гиперпараметрической оптимизации.

СПИСОК ЛИТЕРАТУРЫ

1. **Auto-WEKA:** Combined selection and hyperparameter optimization of classification algorithms / C. Thornton, *et al.* // Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. Pp. 847-855. **[** C. Thornton et al, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms", in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 847-855, 2013. **]**

2. **Nguyen P., Hilario M., Kalousis A.** Using meta-mining to support data mining workflow planning and optimization // Journal of Artificial Intelligence Research. 2014. Vol. 51. Pp. 605-644. **[** P. Nguyen, M. Hilario, A. Kalousis ,"Using meta-mining to support data mining workflow planning and optimization," in *Journal of Artificial Intelligence Research*, vol. 51, pp. 605-644, 2014. **]**

3. **Komer B., Bergstra J., Eliasmith C.** Hyperopt-sklearn: Automatic hyperparameter configuration for scikit-learn // Proc. of the 13th Python in Science Conference (SciPy 2014). USA. 2014. Pp. 32-37. **[** B. Komer, J. Bergstra, C. Eliasmith, "Hyperopt-sklearn: Automatic hyperparameter configuration for scikit-learn", in *Proc. of the 13th Python in Science Conference*, pp. 32-37, 2014. **]**

4. **Efficient** and robust automated machine learning / M. Feurer, *et al.* // Advances in Neural Information Processing Systems. 2015. Vol. 28. Pp. 2962-2970. **[** M. Feurer, *et al*., "Efficient and robust automated machine learning", in *Advances in Neural Information Processing Systems*, vol. 28, pp. 2962-2970, 2015. **]**

5. **Auto-Sklearn** 2.0: The Next Generation / M. Feurer, *et al.* // ArXiv preprint arXiv:2007.04074. 2020. Pp. 1-18. **[** M. Feurer, *et al.,* "Auto-Sklearn 2.0: The Next Generation", in *ArXiv preprint arXiv:2007.04074*, pp. 1-18, 2020. **]**

6. **Evaluation** of a tree-based pipeline optimization tool for automating data science / R. S. Olson, *et al.* // Proc. of the Genetic and Evolutionary Computation Conference (GECCO). USA. 2016. Pp. 485-492. **[** R. S. Olson, *et al., "*Evaluation of a tree-based pipeline optimization tool for automating data science", in *Proc. of the Genetic and Evolutionary Computation Conference*, pp. 485-495, 2016. **]**

7. **Olson R. S., Moore J. H.** TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning // JMLR: Workshop and Conference Proceedings. 2016. Vol. 64. Pp. 66-74. **[** R. S. Olson, J. H. Moore, "TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning", in *JMLR: Workshop and Conference Proceedings*, vol. 64, pp. 66-74, 2016. **]**

8. **Auto-weka** 2.0: Automatic model selection and hyperparameter optimization in weka / L. Kotthoff, *et. al*. // The Journal of Machine Learning Research. 2017. Vol. 18 (1). Pp. 826-830. **[** L. Kotthoff, et al., "Auto-weka 2.0: Automatic model selection and hyperparameter optimization in weka", in *The Journal of Machine Learning Research*., vol. 18 (1), pp. 826-830, 2017. **]**

9. **Alsolai H., Roper M.** Determining the Best Prediction Accuracy of Software Maintainability Models Using Auto-WEKA // Advances in Data Science, Cyber Security and IT Applications. Springer. 2019. Vol. 1098. Pp. 60-70. **[** H. Alsolai, M. Roper, "Determining the Best Prediction Accuracy of Software Maintainability Models Using Auto-WEKA", in *Advances in Data Science, Cyber Security and IT Applications*, vol. 1098, pp. 60-70, 2019. **]**

10. **Jin H., Song Q., Hu X.** Auto-keras: An efficient neural architecture search system // arXiv:1806.10282. 2018. Pp. 1-11. **[** H. Jin, Q. Song, X. Hu, "**.** Auto-keras: An efficient neural architecture search system", in *arXiv preprint arXiv:1806.10282,* pp. 1-11, 2018. **]**

11. **RECIPE:** A Grammar-Based Framework for Automatically Evolving Classification Pipelines / G. Alex, *et al*. // Proc. of the 20th European Conference on Genetic Programming (EuroGP'17). Amsterdam. 2017. Pp. 446-461. **[** G. Alex, *et al.,* "RECIPE: A Grammar-Based Framework for Automatically Evolving Classification Pipelines", in *Proc. of the 20th European Conference on Genetic Programming,* pp. 446-461, 2017. **]**

12. **Mohr F., Wever M., Hüllermeier E.** ML-Plan: Automated machine learning via hierarchical planning // Machine Learning. 2018. Vol. 107. Pp. 1495-1515. **[** F. Mohr, M. Wever, E. Hüllermeier, "ML-Plan: Automated machine learning via hierarchical planning", in *Machine Learning*, vol. 107, pp. 1495-1515, 2018. **]**

13. **SHOP2:** An HTN planning system / D. S. Nau, *et al*. // Journal of Artificial Intelligence Research (JAIR). 2003. Vol. 20. Pp. 1-26. **[** D. S. Nau, *et al*., "SHOP2: An HTN planning system", in *Journal of Artificial Intelligence Research*vol, vol. 20, pp. 1-26, 2003. **]**

14. **Nguyen P., Kalousis A., Hilario M.** A meta-mining infrastructure to support KD workflow optimization // Proc. of the PlanSoKD-11 Workshop at ECML/PKDD. Greece. 2011. Pp. 1-10. **[** P. Nguyen, A. Kalousis, M. Hilario, "A meta-mining infrastructure to support KD workflow optimization", in *Proc. of the PlanSoKD-11 Workshop at ECML/PKDD,* pp. 1-10, 2011. **]**

15. **Nguyen P., Kalousis A., Hilario M.** Experimental evaluation of the e-lico meta-miner // Proc. of the 5th planning to learn workshop WS28 at ECAI.France, Montpellier. 2012. Pp. 18-19. **[** P. Nguyen, A. Kalousis, M. Hilario, "Experimental evaluation of the e-lico meta-miner", in *Proc. of the 5th planning to learn workshop WS28 at ECAI*, pp. 18-19, 2012. **]**

ОБ АВТОРАХ

БОЛОДУРИНА Ирина Павловна, зав. каф. прикладной математики. Д-р техн. наук по упр. в соц. и эк. сист. (ЮУрГУ, 2004). Иссл. в обл. теории оптимального управления, математического моделирования.

ПАРФЕНОВ Денис Игоревич, доц. каф прикладной математики. Канд. техн. наук по сист. и сетям (ПГУТИ, 2014) Иссл. в обл. математического моделирования, архитектуры высоконагруженных вычислительных систем и систем обеспечения безопасности в сетях связи.

ШУХМАН Александр Евгеньевич, зав. каф. геометрии и компьютерных наук. Канд. пед. наук по теории и методике обучения информатике в высшей школе (МПГУ, 2000). Иссл. в обл. облачных вычислений, эволюционных алгоритмов оптимизации.

ЗАБРОДИНА Любовь Сергеевна, преп. каф. прикладной математики. Магистр прикладной математики и информатики (ОГУ, 2020). Иссл. в обл. интеллектуальных методов оптимизации и машинного обучения.

METADATA

Title: Automated machine learning: overview of the capabilities of modern data analysis platforms.

Authors: I. P. Bolodurina[1], D. I. Parfenov[2], A. E. Shukhman[3], L. S. Zabrodina[4]

Affiliation: Orenburg State University (OSU), Russia.

Email:[1]prmat@mail.osu.ru, [2]parfenovdi@mail.ru,

[3]shukhman@gmail.com, [4]zabrodina97@inbox.ru

Language: Russian.

Source: SIIT (scientific journal of Ufa State Aviation Technical University), vol. 3, no. 1 (5), pp. 50-57. ISSN 2686-7044 (Online), ISSN 2658-5014 (Print).

Abstract: Automatic machine learning (AutoML) methods play a crucial role in working with a huge amount of data and are used in almost every possible field. The use of AutoML tools in applied data analysis tasks allows you to extract new knowledge from the source information, reveal relations-hips within the data, and solve problems of classifi-

cation, clustering, regression, forecasting, etc. In this paper, we review the existing approaches and systems for building flexible AutoML pipelines that use intelligent algorithms for optimizing the structure and parameters. To expand the applied Bayesian optimization approaches within the framework of the implementation of the automatic machine learning process, the meta-learning stage is included, as well as the automated ensemble construction to improve the efficiency of the results obtained. The stages of the process of optimizing data pipelines and configuring machine learning algorithms are formalized, and the problem of optimizing the choice of a combined algorithm and configuring parameters (CASH) is formulated. The CASH problem is the most important element of AutoML systems, the performance and efficiency of the final learning models depend on the methods and algorithms for solving it. In this regard, the next stage of this study is the development and modification of approaches to solving this problem, and it is also planned to select the most effective algorithms for constructing the feature space and modify them to increase the productivity and accuracy of the trained models.

Key words: automated machine learning; data analysis; Bayesian optimization; intelligent optimization algorithms; meta-learning; ensembles of models.

About authors:

BOLODURINA, Irina Pavlovna, head of applied mathematics department. Dr. of Tech. Sci. (SUSU, 2004).

PARFENOV, Denis Igorevich, associate professor of applied mathematics department. Cand. of Tech. Sci. (PSUTI, 2014).

SHUKHMAN, Alexander Evgenievich, head of the department of geometry and computer science. Cand. of Pedagogical Sciences (MSPU, 2000).

ZABRODINA, Lyubov Sergeevna, lecturer of the Dept. of Applied Mathematics. Master of Applied Mathematics and Computer Science (OSU, 2020).