AG Publishing House

Handbook On Fundamentals And Methods Of Machine And Deep Learning(VOLUME-1)

Volume 1
Year: 2024

AG PH Books

AutoML: Streamlining and Automating the Machine Learning Pipeline

Dr. Muthukumar Subramanian[1*]

[1]*Professor, CSE, Hindustan Institute of Technology and Science (HITS), Deemed-to-be University, Chennai.*

Abstract

The goal of the developing discipline of "Automated Machine Learning," or "AutoML," is to automate the process of creating machine learning models. By automating as much of the repetitive, unproductive labor that arises when machine learning is used, autoML was developed to boost productivity and efficiency. The development of machine learning models is made easier by the revolutionary technique known as Automated Machine Learning (AutoML). An overview of current developments in automated machine learning pipelines and AutoML approaches is given in this study. To sum up, automated machine learning pipelines have a great deal of potential to advance the field of machine learning, spur innovation, and influence the direction of artificial intelligence. Through the use of AutoML and multidisciplinary cooperation, we can solve intricate problems, open up new avenues, and build a future where intelligent automation enables people and businesses to prosper in the digital era.

Keywords: Automated Machine Learning, Neural architecture search, Reinforcement learning, Machine learning.

1 Introduction

The increasing demand for advanced machine learning models has created an urgent need for scalable and efficient model development techniques. However, the traditional machine learning pipelines' human inputs pose major barriers to the seamless transition to optimal model creation. This section will describe the challenges inherent in these conventional pipelines and emphasize the impediments resulting from

* ISBN No. - 978-81-974433-9-8

manual tasks throughout crucial stages such as feature engineering, model selection, data preparation, and hyperparameter tuning.

With applications in several study fields, including biomedical informatics—a discipline associated with noisy, complicated, diverse, and often large-scale data—machine learning (ML) has emerged as a key component of contemporary data science. The ability of machine learning (ML) to train models that can be used to find complicated multivariate connections within ever-larger feature spaces (such as integrated multi-omics data and "omics" data) and make predictions is what has sparked a surge in interest in the field. The viability of these initiatives has increased with increased access to strong computational resources. To make it easier to perform bespoke ML studies, a multitude of ML tools, packages, and other resources have been produced. One well-known and approachable example is the scikit-learn library, which was developed using the Python programming language. Packages like scikit-learn concentrate on making certain ML analytic pipeline components (such feature selection, cross validation, and ML modeling) easier to use. Even when working with comparable data, there is a great degree of variation in the "how" these aspects are combined since this is often left up to the practitioner. The majority of guidelines for doing machine learning analysis are available as a community knowledge base of specific "potential pitfalls" and "best practices." The majority of machine learning research focuses on enhancing or customizing certain techniques for a particular data type, task, or application area. Remarkably few publications have addressed the topic of how to properly and successfully put up a complete machine learning pipeline or have offered clear, simple examples of how to do so. It might be intimidating for those who are new to an area of expertise to know where to begin.

2 Literature Review

(Salehin et al., 2024) [1] We give a thorough description and a summary of the data processing needs for AutoML techniques in this semantic review study. We provide more attention to neural architecture search (NAS) since it is one of the most talked-about subtopics in the AutoML space right now. NAS techniques explore through a wide range of potential topologies using machine learning algorithms to identify the optimal one for a particular job. We offer an overview of the results attained by typical NAS algorithms on popular benchmark datasets like as CIFAR-10, CIFAR-100, ImageNet, and others. Furthermore, we explore a number of significant avenues for future study in NAS techniques, such as combined hyperparameter with architecture optimization, one-shot NAS, and one- or two-stage NAS. We spoke about how the particular issue being addressed might affect the size and complexity of the search space in NAS. Finally, we review a number of outstanding issues (SOTA issues) in the state-of-the-art AutoML techniques that guarantee more study in the future.

(Dharani et al., 2024) [2] We go over the foundations of autoML, the elements of automated machine learning pipelines, cutting-edge platforms and frameworks, obstacles and constraints, potential paths forward, and real-world applications. Additionally, we traverse the maze of cutting-edge AutoML frameworks and platforms, such as Auto-Sklearn, H2O AutoML, and Google AutoML, and elucidate their functionalities, advantages, and practical uses. We unveil the democratizing potential of AutoML—

which enables both new and seasoned practitioners to access the power of machine learning without being constrained by the complexity of conventional model development—through a thorough analysis of these platforms. We want to shed light on the present status of AutoML research and its ramifications for machine learning going forward with this thorough study.

(Baratchi et al., 2024) [3] Give a thorough summary of AutoML's history, current state, and prospects for the future. Initially, we present the notion of AutoML, explicitly specify the issues it seeks to resolve, and outline the three fundamental elements of AutoML methodologies: the search space, search strategy, and performance assessment. We next go over hyperparameter optimization (HPO) approaches that are often employed in AutoML systems design. Finally, we provide a summary of neural architecture search, a specific application of AutoML that generates deep learning models automatically. We go over and contrast the current AutoML solutions. Lastly, we provide a list of open problems and potential lines of further investigation. All things considered, we provide a thorough overview to machine learning researchers and practitioners and lay the groundwork for future advancements in AutoML.

(Moharil et al., 2024) [4] Traditionally, pre-trained transformer models are strategically integrated to reduce dependency on the computationally intensive Neural Architecture Search. This novel method simplifies the pipeline creation process by allowing disparate data modalities to be effectively united into high-dimensional embeddings. We use a sophisticated Bayesian Optimization approach, guided by meta-learning, to enable the pipeline synthesis to warm up and increase computational effectiveness. Our approach shows how it may be used to build sophisticated, personalized multimodal pipelines with constrained processing power. The findings add to the current body of work in the area of autoML and provide new avenues for effectively managing intricate multimodal data. This work acknowledges the collaborative and dynamic character of multimodal machine learning and marks a step toward the creation of more effective and adaptable tools in this area.

(Kaur, 2023) [5] By automating intricate procedures that are necessary for creating models, Automated Machine Learning (AutoML) revolutionizes the machine learning pipeline. Its primary objective is to minimize the amount of manual intervention needed to boost production. AutoML systems use evolutionary algorithms, reinforcement learning, and Bayesian optimization to automatically determine which models and hyperparameters are optimum for a particular task. The time and effort required to develop machine learning models is significantly reduced by these tools, which enable users to describe data, establish issue statements, and impose limits. AutoML's enhanced accessibility makes machine learning more accessible and draws in a wider user base.

(Bifarin & Fernández, 2023) [6] In order to improve metabolomics analysis, this work presents a unified pipeline that integrates AutoML with explainable AI (XAI) approaches. In OC patients and those with other gynecological tumors (Non-OC), as well as in distinguishing between RCC and healthy controls, autoML employing autosklearn outperformed standalone ML techniques like SVM and random forest. Using a combination of ensemble methods and algorithms, Autosklearn produced better results (AUC of 0.97 for RCC and 0.85 for OC). According to a worldwide ranking of feature significance produced by Shapley Additive Explanations (SHAP), the most discriminative metabolites for RCC and OC,

respectively, were dibutylamine and ganglioside GM(d34:1). A thorough error analysis was carried out using decision plots to compare the feature relevance for samples that were properly vs wrongly categorized. Our pipeline essentially highlights how important it is to harmonize AutoML with XAI, enabling both easier ML application and better interpretability in metabolomics data research.

(Baudart et al., 2021) [7] The use of automated machine learning (AutoML) may increase the output of data scientists. But data scientists won't be able to use their intuition if machine learning is fully automated. As a result, data scientists often choose incremental automation over complete automation, in which humans make certain decisions and AutoML does the rest. Unfortunately, implementing AutoML gradually requires significant non-compositional code modifications and is difficult even with modern tools. Combinators, a potent idea from functional programming, allow for more compact composing code. A new collection of orthogonal combinators for building pipelines with machine-learning operators is presented in this study. It outlines a translation strategy that maps search spaces for AutoML optimizers to pipelines and related hyperparameter schemas. Building upon this framework, this article introduces Lale, an open-source AutoML library compatible with sklearn, and assesses it via a user research.

(Kißkalt et al., 2020) [8] In industrial application settings, machine learning has often shown superiority over conventional white-box modeling. Nevertheless, because of human elements in the development process, there is little predictability in obtaining a solution near the theoretical optimum. However, all aspects of the machine learning pipeline may be fully automated using automated machine learning (AutoML), including feature extraction, preprocessing, model selection, and hyperparameter tuning. This study illustrates how AutoML may simplify the creation of data-driven industrial applications using a widely used public dataset. Results from previous methods applied to the same dataset are used as a benchmark.

3 Conclusion

In summary, Automated Machine Learning, or AutoML, is a significant development that addresses the challenges associated with traditional model creation operations. AutoML reduces the need for human involvement at crucial stages of machine learning by integrating Bayesian optimization, reinforcement learning, and evolutionary algorithms. The field is advancing as a result of AutoML's development and approach, which mark a paradigm change toward intricate automated processes. The results, when combined with appropriate, reliable references and experiment result compression, aid in a clear understanding of NAS and AutoML. In the next years, autoML is expected to have a big influence on the field of advanced artificial intelligence as well as the AI age. We made an effort to include a variety of well-known machine learning algorithms along with a few more recent or obscure ones that have special benefits (such as rule-based machine learning and genetic programming). Despite enduring challenges like as interpretability and resource needs, AutoML's democratization of machine learning provides more accessibility and efficiency and has the potential to revolutionize several industries and lead to ground-breaking discoveries. spoke about the history of autoML, its essential elements (data

preparation, hyperparameter optimization, and model selection), and its difficulties (scalability, interpretability, etc.). We showcased case studies and use cases from a range of industries to show how AutoML works well for addressing practical issues. We also offered best practices and implementation concerns for choosing, integrating, and implementing AutoML systems.

References

[1] I. Salehin et al., "AutoML : A systematic review on automated machine learning with neural architecture search," J. Inf. Intell., vol. 2, no. 1, pp. 52–81, 2024, doi: 10.1016/j.jiixd.2023.10.002.

[2] R. Dharani et al., "AutoML and Automated Machine Learning Pipelines Dharani," no. 5, pp. 9661–9668, 2024.

[3] M. Baratchi et al., Automated machine learning : past , present and future, vol. 57, no. 5. Springer Netherlands, 2024. doi: 10.1007/s10462-024-10726-1.

[4] A. Moharil, J. Vanschoren, P. Singh, and D. Tamburri, Towards efficient AutoML: a pipeline synthesis approach leveraging pre-trained transformers for multimodal data Ambarish, vol. 113, no. 9. Springer US, 2024. doi: 10.1007/s10994-024-06568-1.

[5] H. Kaur, "AutoML : Streamlining the Machine Learning Pipeline for Efficient Model Development," vol. 4, pp. 1–4, 2023.

[6] O. O. Bifarin and F. M. Fernández, "Automated machine learning and explainable AI (AutoML-XAI) for metabolomics : improving cancer diagnostics .," no. Ml, 2023.

[7] G. Baudart, P. Ram, M. Hirzel, K. Kate, J. Tsay, and A. Shinnar, "Pipeline Combinators for Gradual AutoML," no. NeurIPS, 2021.

[8] D. Kißkalt, A. Mayr, B. Lutz, A. Rögele, and J. Franke, "Streamlining the development of data-driven industrial applications by automated machine learning," Procedia CIRP, vol. 93, no. March, pp. 401–406, 2020, doi: 10.1016/j.procir.2020.04.009.