2022 21st International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

# Application of Data-driven Method for Automatic Machine Learning in Economic Research

1st Wei Wang
*School of Economics and Management*
*China University of Mining and Technology*
*Xuzhou 221116, China*
*School of Internet of Things Engineering*
*Wuxi Taihu University*
*Wuxi 214064, China*
*Email: 463347100@qq.com*

2nd Wenbo Xu
*School of Internet of Things Engineering*
*Jiangnan University*
*Wuxi 214122, China*
*Email: xwb@jiangnan.edu.cn*

3rd Xiang Yao, 3rd Huajun Wang
*School of Internet of Things Engineering*
*Wuxi Taihu University*
*Wuxi 214064, China*
*Email: 53371448@qq.com, 63025389@qq.com*

***Abstract*—At present, the role of machine learning in data analysis is becoming increasingly important, and the digital economy has become the major economic form in the world, as well as the core driving force for China's economic development. Machine learning plays an increasingly significant role in economic research based on big data. To reduce the difficulty of using machine learning and improve the efficiency of machine learning, this paper systematically studies the application of automated machine learning (AutoML) in economic research, focusing on the principles and characteristics of data-driven automated machine learning. Through the experimental comparison of specific automated machine learning methods on the classification of data sets, the optimal applicable method is found. Data-driven automated machine learning can be effectively applied in economic data mining, economic indicator analysis, and policy evaluation.**

***Keywords*-automated machine learning; data-driven; economic research; data analysis**

## I. INTRODUCTION

The world as we know it is continually changing, and one of the fundamental drivers is digital technology. At present, the application of digital technology involves all aspects of human life, and meanwhile the global economy is increasingly digitalized. Digital economy has become the major economic form in the world, as well as the core driving force to promote China's economic and social development. In economic research, the data exhibits the big data characteristics of 5V[1] (Volume, Variety, Velocity, Variability, Veracity). Therefore, the application of machine learning in economic research is also becoming more and more widespread. Machine learning is ability to extract effective information from complex data and construct optimal models through continuous training on data sets. Machine learning provides advanced data-driven techniques to analyze valid information from a large amount of economic data, and then discover the patterns and mechanisms, which are widely applied in economic data mining, economic indicator analysis and policy evaluation.

With the in-depth of the application of big data technology, a large amount of data in economic research has presents unstructured characteristics[2]. Deep learning can automatically learn useful features of the data in model training, making it more suitable for unstructured data analysis. Before the advent of deep learning, the processing of unstructured data usually required manual extraction and design of data features. For small and medium-sized datasets, deep learning can be overfitted, while traditional machine learning can be an effective complement to deep learning.

Machine learning which focuses on selecting the appropriate model training and optimization parameters based on the characteristics of the data set has high technical requirements and usually requires professional technicians to achieve. Automated machine learning techniques can be used to replace part of the work of professional technicians in the machine learning process, reduce the dependence of machine learning on experts in specific applications, and drive the rapid adoption of machine learning techniques. From the perspective of dataset feature similarity, this paper analyzes the construction of a data-driven automated machine learning process, and applies it to dataset analysis in economic research, and then evaluates the performance of typical automatic machine learning methods in comparison.

## II. A BRIEF INTRODUCTION TO MACHINE LEARNING

### A. *Machine Learning Framework*

Machine learning, as an important branch of artificial intelligence (AI), is playing an increasingly important role in the field of information science. It uses algorithms and model training to explore the relevance of data and improve the effectiveness of data analysis. Machine learning uses model training on a dataset to generalize

2473-3636/22/$31.00 ©2022 IEEE
DOI 10.1109/DCABES57229.2022.00019

reasonable trends of change to make the best decisions and predictions[3]. The application of machine learning continues to improve as it is used, and the more data is obtained, the more accurate the predictions will be. A typical machine learning framework consists of four parts: Data Cleaning, Feature Extraction, Feature selection, and Model Selection[4], as shown in Figure 1.

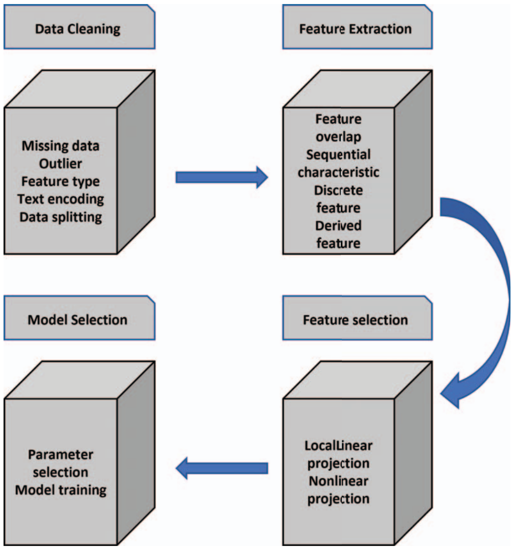

Figure 1. Machine Learning Framework

- Data cleaning: data cleaning is the primary step in the machine learning process to be completed before model training, and its often consumes the most time as well. Usually, the initial data we obtain are often irregular, missing, and incorrect, and cannot be used directly. Therefore, the first necessary step is to standardize the data to ensure the relevance, correctness, and completeness of the data, so as to obtain more satisfactory results in analysis and processing.
- Feature extraction: machine learning based on big data becomes more prevalent, and the datasets increasingly present complex variables and high dimensions, and the computational resources for model training are enormous. Feature extraction is a method to select and combine feature variables to obtain satisfactory processing results by extracting effective variables, reducing the data dimensions, accelerating processing speed, and reducing resource usage. Commonly used methods for feature extraction include principal component analysis (PCA), independent component analysis (ICA), linear discriminant analysis (LDA), etc.
- Feature selection: feature selection is also known as attribute selection, and its goal is to find the optimal subset of features. Feature selection eliminates irrelevant or redundant features and selects the most relevant attributes of the prediction model, thus reducing the number of features, simplifying the model, improving the model accuracy and reducing the running time. In addition, feature selection can also avoid over-fitting of the model. In general, feature selection algorithms include filter methods, wrapper methods, and embedding methods.
- Model selection: model selection is the final step in the machine learning process, where models are selected and evaluated after they have been trained, applying some strategy and algorithm to select an optimal model from the collection of training models. The most important step in model selection is estimating the predictive performance. Typical model selection methods include regularization, cross validation, etc.

### B. *Machine Learning Methods*

There are three primary categories of machine learning methods.

*1) Supervised learning:* Supervised learning is defined as training a model on a labeled dataset and then processing the target dataset through the model to solve a realistic classification or prediction problem. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. Methods that use supervised learning include linear regression, logistic regression, naive bayes, support vector machines, neural networks, and random forests, etc.

*2) Unsupervised learning:* Unsupervised learning performs algorithmic analysis and model training on unlabeled datasets. These algorithms can discover data patterns and classifications based on similarities and differences in the data without human intervention. It also can be used for dimensionality reduction and exploratory data analysis, such as customer classification, image and pattern recognition, etc. The core of unsupervised is self-learning ability, and the method is applied to hierarchical clustering, principal component analysis, k-means clustering, matrix decomposition, etc.

*3) Semi-supervised learning:* Semi-supervised learning belongs to a subfield of weakly supervised learning, which is a compromise between supervised and unsupervised learning. Semi-supervised learning addresses the problem that supervised learning algorithms do not have enough labeled data or labeling enough data is too costly[5]. Methods that use semi-supervised learning include self-training, generative and discriminative model, low-density separation, co-training, etc.

## III. DATA-DRIVEN METHOD FOR AUTOMATIC MACHINE LEARNING

### A. *Automatic Machine Learning(AutoML)*

Automated machine learning (AutoML) is one of the cutting-edge technologies of machine learning at present, the goal of automatic machine learning is to automate model building, where the user only needs to enter data sets and tasks, and the machine automatically optimizes the model via a data-driven method to solve the task related to the dataset[6].

Automated machine learning automates a series of steps in machine learning, including data preprocessing, feature engineering, model selection, and hyperparameter optimization. A commonly used automated machine learning system is shown in Figure 2.

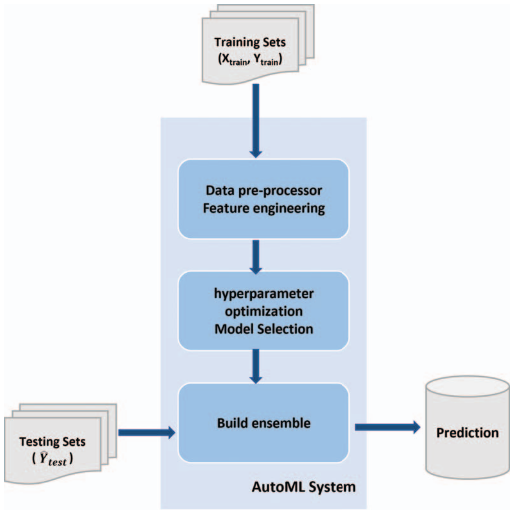

Figure 2. Automatic Machine Learning System

### B. CASH Problem

Automated machine learning (AutoML) faces two important problems, one is that no machine learning method performs best on all datasets, and the other is that some machine learning methods rely heavily on hyperparameter optimization. Then, the core of AutoML is the Combined Algorithm Selection and Hyperparameter optimization (CASH) problem which is to find the joint algorithm and hyperparameter setting that minimizes this loss[7], as shown in Equation 1.

$$A^{\star}, \lambda_{\star} \in \operatorname*{arg\,min}_{A^{(j)}\in\mathcal{A},\lambda\in\Lambda^{(j)}} \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}(A_{\lambda}^{(j)}, D_t^{(i)}, D_v^{(i)}) \qquad (1)$$

- $\mathcal{A} = \left\{A^{(1)}, ..., A^{(R)}\right\}$ is a set of algorithms $A^{(j)}$, and the hyperparameters of each algorithms have domain $\Lambda^{(j)}$.
- $D_t = \{(x_1, y_1), ..., (x_n, y_n)\}$ is a training set which is split into K cross-validation folds $\{D_v^{(1)}, ..., D_v^{(K)}\}$ and $\{D_t^{(1)}, ..., D_t^{(K)}\}$.
- $D_t^{(i)} = D_t/D_v^{(i)} for\ i = 1, ..., K$.
- $\mathcal{L}(A_{\lambda}^{(j)}, D_t^{(i)}, D_v^{(i)})$ denote the loss that algorithm $A^{(j)}$ achieves on $D_v^{(i)}$ when trained on $D_t^{(i)}$ with hyperparameters $\lambda$.

### C. AlphaD3M

AlphaD3M is an automatic machine learning framework using a sequence model and belongs to the application of meta-reinforcement learning. AlphaD3M's operations performed over pipeline primitives and are completely explanatory[8].

DARPA's Data-Driven Model Discovery (D3M) program aims to develop the infrastructure to automatically discover models to solve any task specified by the user on a dataset. the D3M system is built with modules based on a set of computational primitives that synthesize different pipelines and set appropriate hyperparameters to solve various analysis problems on the dataset, while it provides a user-friendly interface to interact with the user.

AlphaD3M uses a collection of primitives developed under the D3M initiative, as well as primitives available in open source libraries such as scikit-learn, to construct pipelines for machine learning tasks, that can be applied to different data types and provide standard performance metrics. AlphaD3M simplifies the process of creating predictive models. Users can interact with the system through a simple development environment and derive models using a small amount of code.

The goal of AlphaD3M is to perform machine learning within a large space, and pre- and post-processing primitives and parameters together constituting the pipeline for solving the task for a given dataset. It use a neural network along with a Monte-Carlo tree search in an iterative process, as shown in Figure 3.

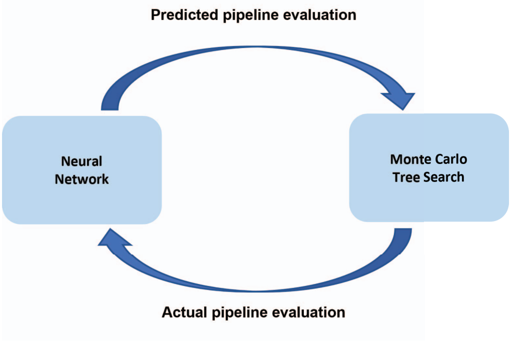

Figure 3. AlphaD3M iterative improvement

*1) Neural Network:* AlphaD3M optimization algorithm is shown in Equation 2.

$$L(\Gamma) = D\log T + (u - f)^2 + \alpha\|\Gamma\|_2 + \beta\|D\|_1 \qquad (2)$$

- $\Gamma$ as a network parameter is optimized by $D$.
- $D$ as a predicted model match real world model $T$.
- $u$ is the predicted and $f$ is the real world evaluation.
- $u$ match $f$ by minimizing the cross entropy loss between $D$ and $T$.
- mean squared error between $u$ and $f$.

*2) Monte Carlo Tree Search (MCTS):* MCTS is very efficient in searching for solutions in high dimensional search because MCTS balances exploration and exploitation in the search process, and the AlphaD3M algorithm uses the neural network to predict the results by running multiple simulations, and based on the prediction results, the MTCS method is used to search for the pipeline sequence $Q$ with higher evaluation scores. The pipeline sequence $Q$ improve the network strategy and thus the prediction result $S$ by using the update rule, as shown in Equation 3.

$$U(s,a) = Q(s,a) + cP(s,a)\frac{\sqrt{N(s)}}{1+N(s,a)} \tag{3}$$

- $Q(s,a)$ is the expected reward value.
- $N(s,a)$ is the number of times.
- $P(s,a)$ is the estimate of the neural network.

## IV. EXPERIMENTS AND ANALYSIS

### A. Experiment Design

The experimental data comes from open source datasets such as OpenML, UCI Repository, etc. According to the access ranking, release time, data characteristics, data size, etc. this paper selects the datasets that related to economic research in the open source datasets to test the performance of AutoML for data analysis in economic research. The data-driven AlphaD3M was used to perform classification tasks on the dataset and the results were compared and analyzed for performance. In this paper, a confusion matrix is used to evaluate the performance of the classification model. P (precision) and R (recall) are derived from the confusion matrix to calculate the $\mathrm{F}_1 - score$ score, as shown in Equation 4. For multiclassification problems, the average value of $\mathrm{F}_1 - score$ is calculated to obtain the value of $\mathrm{F}_1 - macro$.

$$F_{1-Score} = \frac{2 \bullet P \bullet R}{P+R} \tag{4}$$

### B. Results Analysis

In this experiment, the data set is trained in terms of operation time limit, number of data features, and number of samples. The $\mathrm{F}_1 - macro$ score is used as the evaluation basis for the model. The experimental results of AlphaD3M are shown in Figure 4.

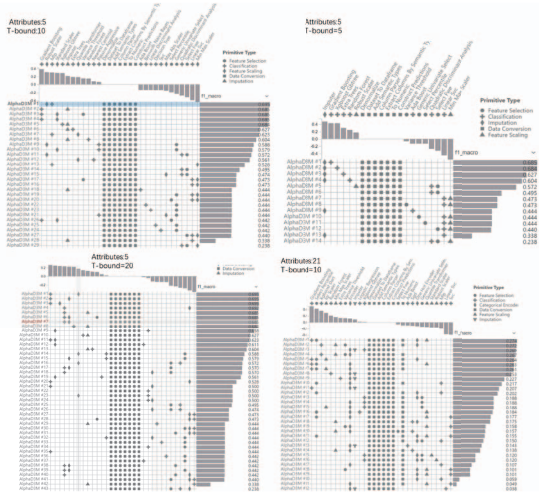

Figure 4. AlphaD3M experimental results

Figure 4 shows that the larger the Time_bound value, the more pipelines are generated and the higher the $\mathrm{F}_1 - macro$ value is, but when the Time_bound value is greater than 10, the $\mathrm{F}_1 - macro$ score does not increase significantly with the number of pipelines, and usually, the value of Time_bound not lower than 10.

Table I
TEST RESULTS TABLE

| F1_macro | Pipelines | Time_bound | Attributes |
|---|---|---|---|
| 0.7126 | 43 | 20 | 5 |
| 0.7126 | 29 | 10 | 5 |
| 0.5950 | 14 | 5 | 5 |
| 0.1403 | 32 | 10 | 21 |

As shown in Table1, the low F1_macro score because of overly complex data features indicates that AlphaD3M is not strong enough for complex datasets, but whether this is related to the number of samples and features requires further experimental verification. Usually, we can perform dimensionality reduction for complex data sets before using AlphaD3M.

## V. CONCLUSION

Data-driven automated machine learning can greatly reduce the difficulty of using machine learning and improve the efficiency of data analysis and can be used as an effective method for economic data mining, economic indicator analysis and policy evaluation. In the experiments we found that it also suffers from performance shortcomings, and in the follow-up study, we will further test the effect of different parameter optimization on its performance, and how to improve its efficiency through deep learning.

## REFERENCES

[1] F. Jiang and W. Zhang, “Application of Machine Learning Methods in Economic Research,” *Statistics & Decision*, vol. 38, no. 4, pp. 43–49, 2022.

[2] J. Xu, Y. Zhu, and C. Xing, “Application of Machine Learning in Financial Asset Pricing: A Review,” *Computer Science*, vol. 49, no. 6, pp. 276–286, 2022.

[3] L. Zhou, Y. Song, W. Ji, and H. Wei, “Machine learning for combustion,” *Energy and AI*, vol. 7, p. 100128, Jan. 2022.

[4] M. Zhang and Y. Zhao, “Application of Machine Learning in Human Resource Management,” *Human Resources Development of China*, vol. 39, no. 1, pp. 71–83, 2022.

[5] D. Rui, Y. Ma, and L. Ye, “Application of Machine Learning Methods in Wastewater Treatment Systems,” *Environmental Engineering*, vol. 40, no. 6, pp. 145–153, 2022.

[6] G. Chen, J. Wang, Q. Li, Y. Yuan, and J. Cao, “Data-driven Method for Automatic Machine Learning Pipeline Generation,” *Journal of Guangxi Normal University(Natural Science Edition)*, vol. 40, no. 3, pp. 185–193, 2022.

[7] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, “Efficient and Robust Automated Machine Learning,” p. 9.

[8] I. Drori, Y. Krishnamurthy, R. Rampin, R. d. P. Lourenco, J. P. Ono, K. Cho, C. Silva, and J. Freire, “AlphaD3M: Machine Learning Pipeline Synthesis,” p. 8.