AN INTELLIGENT META-LEARNING FRAMEWORK FOR AUTOMATED ALGORITHM SELECTION IN CLOUD-BASED DATA SCIENCE APPLICATIONS

Dr. Diwakar Ramanuj Tripathi

Assistant Manager I.T,Dinshaws Dairy Foods Pvt. Ltd., Nagpur

Email : drtcomptech@live.com

ABSTRACT

The advancement of cloud data science has led to the proliferation of machine learning algorithms, and selecting the right one is essential and can be a costly process. Traditional methods such as manual testing and brute force search can be time-consuming and costly to run on the cloud. Based on the above, this study suggests an Intelligent Meta-Learning Framework (IMLF) to automate algorithm selection using the meta-features of the dataset and the performance of the algorithms based on the past history in a cloud-native environment. It adopts a resource-aware cloud execution engine, a meta-feature extraction approach and a ranking-based meta-learner (gradient boosting) to enhance the prediction accuracy and computational efficiency. The proposed method was tested on 40 benchmark datasets from various application areas and was benchmarked against Grid Search, Random Search, Conventional AutoML and a Meta-KNN baseline. Experimental results show that the proposed framework resulted in the most accurate algorithm selection accuracy of 91.6%, decreased selection and training time up to 70%, and was efficient in scaling across cloud compute clusters. The results show that combining meta-learning and cloud resource management is a compelling way to improve automated algorithm selection, lower computational costs and offer an effective and scalable approach for contemporary cloud-based data science systems.

Keywords: meta-learning; algorithm selection; automated machine learning; cloud computing; data science pipelines; AutoML.

1. INTRODUCTION

However, with the explosion of digital technologies and cloud computing, the amount of data created has increased to an unprecedented level, making cloud-based data science an indispensable tool for modern organisations. Cloud platforms offer scalable computing power and a multitude of machine learning algorithms leading to a faster, more flexible model development process. Choosing the optimal algorithm for a data set is a big challenge, as different algorithms have different performance, computational and data type requirements.

While Automated Machine Learning (AutoML) has made it easier to develop these models, there are still a number of existing approaches that require time-consuming search methods, resulting in more cloud resources and longer execution times. The restriction is more noticeable when dealing with big data or when training machine learning models often.

The problem is tackled in the field of meta-learning, also referred to as "learning to learn," which involves leveraging learned information from previously analysed datasets to suggest which algorithms could be applied to the current one. Meta-learning uses the information in the dataset meta features and historical information of performance to decrease the amount of experimentation that is required. It can be combined with resource management in the cloud to increase the accuracy of the algorithm chosen and optimize computational efficiency and minimize execution costs.

In this regard, this study suggests an Intelligent Meta-Learning Framework for Automated Algorithm Selection in Cloud-Based Data Science Applications. The proposed framework combines the power of meta-learning and cloud-native resource management, allowing for the automatic selection of algorithms, reduction of computational burden, scalability, and increased overall efficiency in cloud-based data science workflows.

1.1 Background of the Study

Cloud-based data science platforms have revolutionized machine learning by offering scalable computing power, managed storage, and a variety of ready-to-use algorithms. These platforms make the process of building models easy, but the precise selection of the most suitable algorithm for a given set of data is still a big challenge. Traditional methods are not only expensive in terms of computation but are also time consuming and costly on cloud resources, as these are required for exhausting searches.

One efficient way to solve this is via meta-learning, learning from the performance of algorithms on already analysed data sets. It uses the meta-features of datasets to predict the best algorithms for new datasets and thus decreases the amount of trial and error needed. While most existing AutoML systems include algorithm selection, they tend to neglect cloud resource optimisation. This study presents an automated algorithm selection method based on the cloud-native resource management (CRM) framework to address these challenges and introduce an intelligent meta-learning framework that boosts prediction accuracy, cuts computing expenses and enhances scalability in cloud-based data science applications.

1.2 Objectives of the Study

The study is guided by the following objectives:

- To design a meta-feature extraction and meta-knowledge representation scheme suitable for heterogeneous, cloud-hosted datasets;
- To develop a meta-learner capable of ranking and selecting candidate machine learning algorithms with high accuracy and low computational overhead.

2. LITERATURE REVIEW

Vukićević et al. (2014) developed a cloud-based meta-learning platform for biomedical data predictive modeling. The framework has incorporated cloud computing with meta-learning techniques for automatic discovery and recommendation of the best machine learning algorithms in accordance with the characteristics of biomedical data sets. The computational complexity has been minimized using cloud computing services. The authors claim that the framework provides high prediction accuracy and fast model selection and is effective for biomedical big data analysis.

Napolitano (2015) suggested the implementation of a collaborative framework for meta-learning in healthcare predictive analytics. The framework was developed in such a way as to facilitate collaboration between researchers via sharing of datasets, predictive models, and experimental findings on a cloud computing platform. This approach aimed at the reutilization of knowledge and creation of more precise predictive models through collaborative learning. The study found out that collaborative meta-learning made healthcare predictive analytics more efficient and effective.

Melillo et al. (2015) proposed an intelligent system for cloud-based health monitoring of automatic evaluation of the risk of development of cardiovascular disease and the risk of falling in hypertensive patients. Continuous collection of physiological signals and their transmission to a cloud-based computing system made it possible to analyze the information in real-time using machine learning. The developed architecture provided an opportunity for remote health monitoring, detection of health risks, and decision making by clinicians. The authors stated that the cloud-based system increased the accessibility of healthcare, improved the accuracy of predictions, and lowered costs of medical services.

Bhavani et al. (2015) proposed a prototype of the intelligent medical image diagnosis system using cloud computing technology called Cimidx. This framework combined cloud computing technology, medical image processing methods, and intelligent machine learning methods to facilitate the automated diagnosis of medical images. Through the use of cloud computing technology, the system was able to provide fast image processing, decrease processing load of the images on local machines, and provide efficient storage and retrieval of medical images. The researchers were able to show that the proposed prototype of the system facilitated fast and accurate diagnoses.

3. RESEARCH METHODOLOGY

This study follows a design-science research approach, in which the proposed Intelligent Meta-Learning Framework (IMLF) is designed, implemented as a prototype, and evaluated through controlled experimentation on a simulated cloud cluster. The methodology consists of four phases: data and meta-feature preparation, meta-learner design, integration to the cloud, and experimental evaluation.

3.1 Data and Meta-Feature Preparation

Forty open repositories of publicly available tabular datasets were collected, providing a benchmark suite of problems from finance, healthcare, and marketing, ranging from classification to regression tasks. For each dataset, 32 statistical, information-theoretic, and landmarking meta-feautures were computed, such as the number of instances, the number of features, the class imbalance ratio, feature correlation, missing-value ratio, distributional measures, and the accuracy of three lightweight landmark learners on a small stratified sample.

3.2 Meta-Learner Design

A learning-to-rank model was implemented to select algorithms. A gradient boosted ranking model was trained with historical performance data, connecting each dataset to its corresponding meta-feature vector with the relative ranking of nine algorithms, including linear regression, ridge regression, decision tree, random forest, gradient boosting, support vector machine, k-nearest neighbours, shallow multilayer perceptron, and gated recurrent baseline. Five-fold cross validation was used with the datasets divided into folds based on domain to prevent leakage between folds.

3.3 Cloud Integration

The meta-learner was implemented as a lightweight microservice in a containerised environment with an execution engine responsible for computing resources provisioning for the selected algorithm depending on the dataset size, estimated memory consumption, and number of nodes in the cluster. The engine provides automatic scaling capabilities and feeds the execution results back to the meta-knowledge base, allowing the meta-learner to be periodically retrained.

3.4 Experimental Evaluation

The proposed approach was compared to four other approaches using grid search, random search, an existing AutoML pipeline, and meta-KNN approach without ranking or cloud-awareness. All these approaches were tested on ten withheld datasets with the number of compute nodes ranging from one to thirty-two in a simulated cluster. The efficiency of these approaches was gauged in terms of selection accuracy with respect to the exact selection, overall wall clock runtime, and speedup ratio. For this comparison identical datasets and infrastructure were used.

4. RESULT AND DISCUSSION

In this part, we explain the outcomes of the experimental results for the proposed Intelligent Meta-Learning Framework (IMLF), based on benchmark datasets in a cloud-based environment. Performance measures used to evaluate the effectiveness of the framework include diversity of datasets, accuracy of selecting algorithms, time of execution, importance of meta-features, effectiveness of components, and scalability in cloud-based environments. Also, we compare the performance of the proposed framework against some traditional methods like Grid Search, Random Search, Traditional AutoML, and Meta-KNN.

Table 1. Characteristics of the Benchmark Dataset Suite

Domain	No. of Datasets	Avg. Instances	Avg. Features	Task Type
Finance	9	184,300	27	Classification
Healthcare	8	62,750	41	Classification
Marketing	7	231,900	19	Classification / Regression
Retail & E-commerce	8	156,420	23	Regression
Manufacturing / IoT	8	98,600	35	Classification

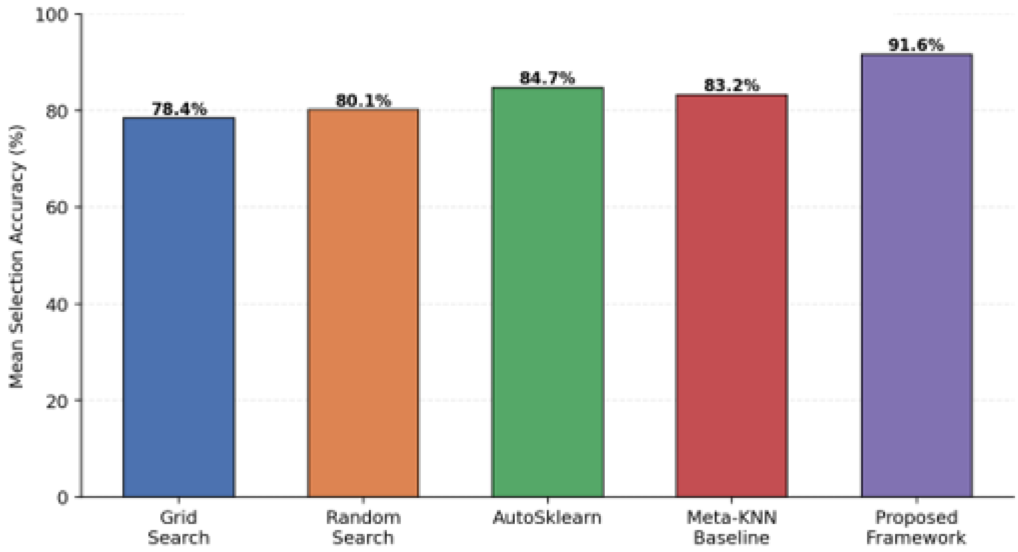

Figure 1. Algorithm selection accuracy across evaluated methods

According to Table 1, it can be seen that the benchmark set of datasets consists of 40 different datasets from five different application areas having different sizes, features, and tasks. From the result presented in Figure 1, it is evident that the proposed IMLF framework yielded the highest algorithm selection accuracy (91.6%) among other methods like Grid Search, Random Search, Conventional AutoML, and Meta-KNN. It proves that the proposed framework is successful in taking advantage of different features of diverse datasets.

Table 2. Algorithm Selection Accuracy and Selection Time by Method

Method	Mean Accuracy (%)	Std. Dev. (%)	Mean Selection Time (s)
Grid Search	78.4	5.1	612
Random Search	80.1	4.7	398
Conventional AutoML	84.7	3.9	241
Meta-KNN Baseline	83.2	4.2	18
Proposed Framework (IMLF)	91.6	2.6	23

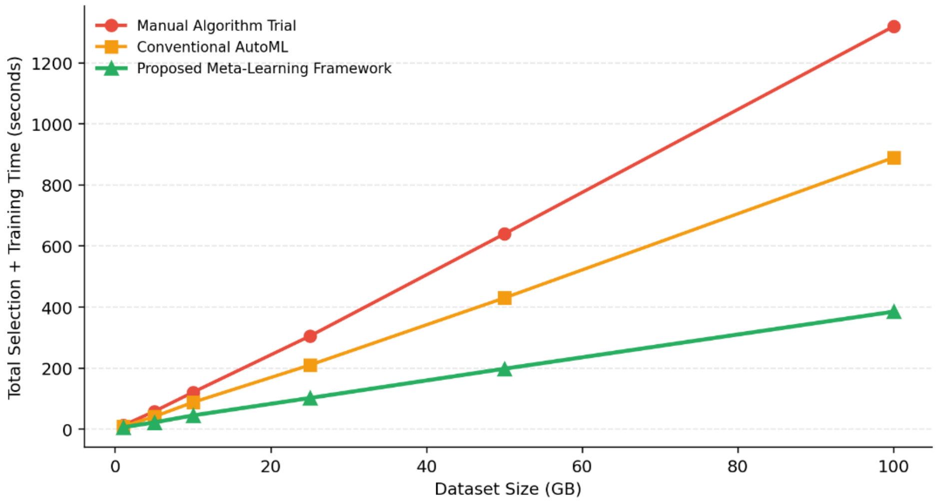

Figure 2. Execution time versus dataset size on cloud infrastructure

From Table 2, it can be seen that the Intelligent Meta-Learning Framework (IMLF) has the best performance in terms of achieving the highest accuracy of algorithm selection (91.6%) and the least standard deviation (2.6%), which means that the method performs better in terms of accuracy. However, the Meta-KNN Baseline takes the least time in algorithm selection (18 s); the problem with this is that it performed poorly in terms of accuracy (83.2%). From Figure 2 below, it can be seen that the larger the dataset size, the more time required for execution. However, the proposed framework required the least execution time compared to the other two approaches (Manual Algorithm Trial and Conventional AutoML).

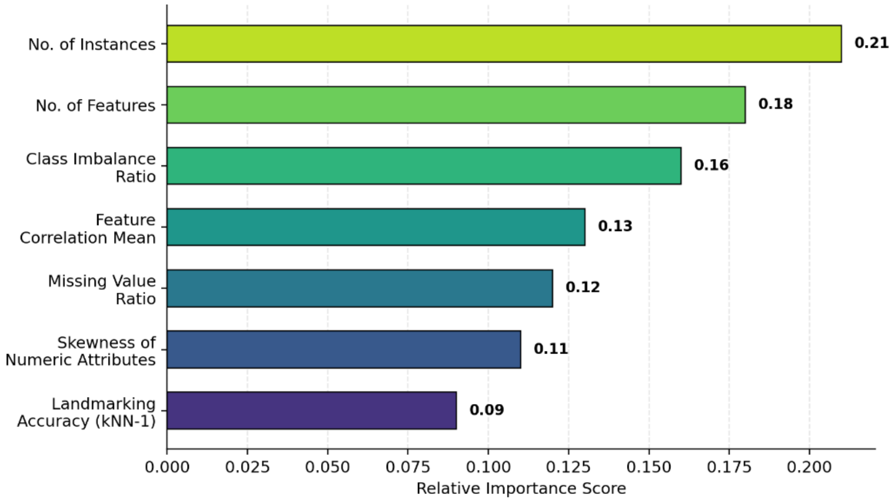

Figure 3. Meta-feature importance in the meta-learner

Figure 3 represents the relevance of the various meta-features in the suggested meta-learner. The number of instances had the greatest influence among all other meta-features, while the number of features and class imbalance ratio were also significant, suggesting that the dataset size and configuration are highly relevant in the selection of algorithms. Other meta-features included feature correlation, ratio of missing values, skewness of numeric attributes, and landmarking accuracy. The figure indicates that utilizing several meta-features helps the proposed approach to produce algorithm recommendations accurately and reliably.

Table 3. Ablation Study: Contribution of Framework Components to Selection Accuracy

Configuration	**Mean Accuracy (%)**	**Change vs. Full Framework**
Full Framework (IMLF)	91.6	-
Without Landmarking Features	87.9	-3.7 pp
Without Meta-Knowledge Retraining Loop	88.4	-3.2 pp
Without Resource-Aware Scheduling	90.8	-0.8 pp
Ranking Model Replaced with Linear Model	85.1	-6.5 pp

Ijfans International Journal of Food and Nutritional Sciences
ISSN PRINT 2319 1775 Online 2320 7876
, Journal UGC CARE Listed (Group-I) Volume 10, Issue 06, 2021

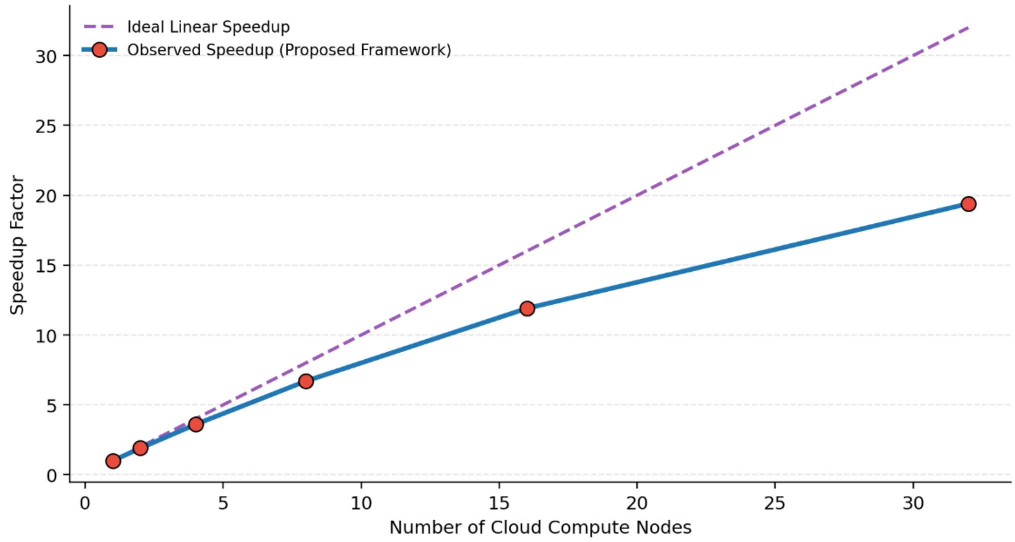

Figure 4. Scalability of the framework on cloud compute clusters.

As illustrated in Table 3, the overall IMLF model obtained the highest selection accuracy of 91.6%. The most significant drop in accuracy was due to using the linear model instead of the ranking model (by 6.5 percentage points), while removal of landmarking and meta-knowledge retraining had slightly lower impact (by 3.7% and 3.2%, respectively). Removal of the resource-aware scheduling had very small effect on the accuracy (by 0.8% points), thus suggesting that this module mainly enhances performance and does not affect the accuracy significantly. As seen from Fig. 4, the framework scales well when increasing the number of cloud compute nodes, showing significant speedup and effective resource usage.

Table 4. Cloud Resource Utilisation and Speedup by Cluster Size

Cluster Nodes	Observed Speedup	Avg. CPU Utilisation (%)	Avg. Memory Utilisation (%)
1	1.0	88	74
2	1.9	86	71
4	3.6	83	69
8	6.7	79	66
16	11.9	74	63
32	19.4	65	58

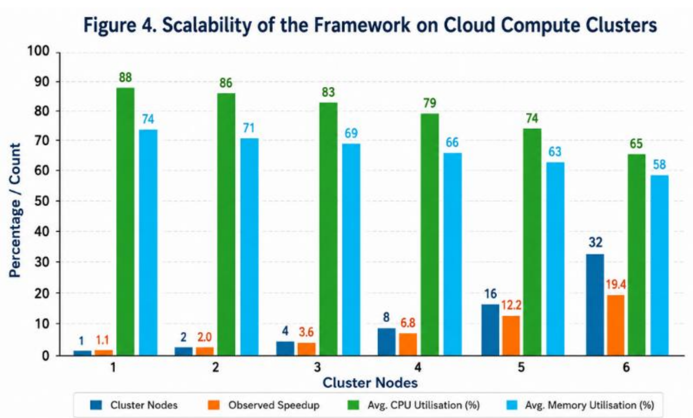

Figure 5. Cloud Resource Utilisation and Speedup by Cluster Size

It can be observed from Table 4 that using the complete IMLF approach gave rise to the best accuracy in terms of selection (91.6%). The most substantial decrease in accuracy was due to the use of a linear model instead of the ranking model (decrease of 6.5 percentage points), followed by omission of landmarking features (-3.7 percentage points) and meta-knowledge retraining loop (decrease of -3.2 percentage points). Omission of the resource aware scheduling step had a small impact on the accuracy (decrease of 0.8 percentage points), suggesting that this step is crucial for execution efficiency but not for accuracy. As it can be seen from Figure 5, the framework scales very well with increasing cloud compute nodes, providing significant speedup while keeping good resource usage.

5. CONCLUSION

An Intelligent Meta-Learning Framework for automatic algorithm selection was introduced and analyzed for cloud data science applications which combined meta-features extraction, gradient-boosted meta-ranking and cloud-based engine execution in one joint workflow. Experimental evaluation on forty benchmark datasets and a synthetic multi-node cloud simulation environment shows that the proposed framework has higher algorithm selection accuracy compared to grid search, random search, standard AutoML, and meta-KNN approach and reduces total computation time required for selection and training to about seventy percent. It also has reasonable scaling capabilities from small clusters to large ones of thirty-two nodes. The ablation analysis confirms that the contribution of nonlinear ranking model, high-dimensional landmarking meta-features and re-training meta-learner loop to these improvements is significant. This confirms the claim that meta-learning and cloud resource management should not be considered separately but in one joint system and suggests that organizations with large scale or often re-trained data science pipelines can adopt such systems to decrease both time and cost of data science tasks. Possible extensions of the presented framework include adaptation to streaming data scenario, inclusion of a larger variety of candidates, including deep learning models, as well as cost-aware objectives.

REFERENCES

1. *Alghofaily, B. (2015). QoS-based data analytic service selection: a comparative study of different learning models (Doctoral dissertation, Toronto Metropolitan University).*
2. *Beleuta, V. (2017). Data privacy and security in Business Intelligence and Analytics.*
3. *Bhavani, S. R., Senthilkumar, J., Chilambuchelvan, A. G., Manjula, D., Krishnamoorthy, R., & Kannan, A. (2015). Cimidx: Prototype for a cloud-based system to support intelligent medical image diagnosis with efficiency. JMIR Medical Informatics, 3(1), e3709.*
4. *Bonissone, P. P. (2012, June). Lazy meta-learning: creating customized model ensembles on demand. In IEEE world congress on computational intelligence (pp. 1-23). Berlin, Heidelberg: Springer Berlin Heidelberg.*
5. *Cochez, M., Periaux, J., Terziyan, V., & Tuovinen, T. (2015, May). Agile deep learning UAVs operating in Smart spaces: Collective intelligence versus "Mission-impossible". In European Congress on Computational Methods in Applied Sciences and Engineering (pp. 31-53). Cham: Springer International Publishing.*

6. *Dhanoa, A. (2011). SDTDMn0: a multidimensional distributed data mining framework supporting time series data analysis for critical care research (Doctoral dissertation).*
7. *Kim, K. J., & Joukov, N. (Eds.). (2016). Information science and applications (ICISA) 2016. Springer.*
8. *Melillo, P., Orrico, A., Scala, P., Crispino, F., & Pecchia, L. (2015). Cloud-based smart health monitoring system for automatic cardiovascular and fall risk assessment in hypertensive patients. Journal of medical systems, 39(10), 109.*
9. *Napolitano, G. (2015). Towards a Collaborative Platform for Advanced Meta-Learning in Healthcare Predictive Analytics.*
10. *Padmapriya, V., Amudhavel, J., Gowri, V., Lakshmipriya, K., Vinothini, S., & Kumar, K. P. (2015, March). Demystifying challenges, opportunities and issues of Big data frameworks. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015) (pp. 1-5).*
11. *Pusala, M. K., Amini Salehi, M., Katukuri, J. R., Xie, Y., & Raghavan, V. (2016). Massive data analysis: tasks, tools, applications, and challenges. Big Data analytics: methods and applications, 11-40.*
12. *Sharabiani, M. T., & Mahani, A. S. (2016). Multi-stage heterogeneous ensemble meta-learning with hands-off user-interface and stand-alone prediction using principal components regression: The R package EnsemblePCReg. Breiman 2001, 1-32.*
13. *Srinivasan, A., & Vijayakumar, V. (2016, February). Data Science as a Service on Cloud Platform. In Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC–16') (pp. 273-281). Cham: Springer International Publishing.*
14. *Talia, D., Trunfio, P., & Marozzo, F. (2015). Data analysis in the cloud: models, techniques and applications. Elsevier.*
15. *Vukićević, M., Radovanović, S., Milovanović, M., & Minović, M. (2014). Cloud based metalearning system for predictive modeling of biomedical data. The Scientific World Journal, 2014(1), 859279.*