WIH3001 DATA SCIENCE PROJECT

[PROJECT PROPOSAL]

SEMESTER 1

SESSION 2023/2024

TITLE OF PROJECT:

HYPERPARAMETER OPTIMIZATION USING HALF-GRID SEARCH, GENETIC ALGORITHMS, SIMULATED ANNEALING AND COMPARISON WITH GRID SEARCH

PREPARED BY:

WONG YAN JIAN

(U2102753)

SUPERVISED BY:

ASSOCIATE PROF. TS. DR. SRI DEVI A/P RAVANA

Hyperparameter Optimization using Half-Grid Search, Genetic Algorithms, Simulated Annealing and Comparison with Grid Search

Wong Yan Jian
u2102753@siswa.um.edu.my[a]

[a] *Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Wilayah Persekutuan Kuala Lumpur, Malaysia*

Keywords: Hyperparameter Optimization, Machine Learning, Half-Grid Search, Genetic Algorithm, Simulated Annealing

1. Introduction

The number of data generated has been growing from time to time since the beginning of the Internet age, whether from online transactions, emails, social media interactions, sensors and other sources (Sagiroglu and Sinance, 2013). In order to tap into the value that could be gained from these data, machine learning (ML) is the most intelligent way to do so. Various types of ML methods, such as supervised, unsupervised, semi-supervised, deep learning and reinforcement learning have been used to analyze and search the hidden patterns, trends and outliers in the datasets. By using these techniques, we can obtain valuable insights from the data to solve business problems, support data-driven decision-making and promote the industry from machine automation to knowledge automation (Ge et al., 2017). Therefore, ML techniques have been applied in many real-world application domains, including predictive analytics and intelligent decision-making, Internet of Things (IoT) and smart cities, E-commerce and product recommendations, as well as other business problems (Sarker, 2021). However, there are many challenges that have to be tackled when building an ML model with a high level of accuracy and precision. One such challenge is hyperparameter tuning.

Hyperparameters are a set of settings that are used to configure an ML model, such as the learning rate for training a deep learning network or the regularization strength (L1 or L2 penalties) in linear models. In addition, they define the algorithm used for minimizing the loss function, like the number of trees in random forests as well as the activation function and optimizer types in neural networks (Yang and Shami, 2020). Hyperparameters must be carefully chosen as they could affect the performance of the model to various degrees (Hoque and Aljamaan, 2021). Thus, a range of possibilities have to be explored in order to build the desirable and usable ML model. The process of finding an optimal set of hyperparameter configurations for the model is known as hyperparameter tuning. It is a part of the key process, especially in designing tree-based ML models and deep neural networks as these types of models have many hyper-parameters (Hutter et al, 2019).

Traditionally, hyperparameter selection has been guided by heuristic rules or experimental evaluation of various combinations within a predefined grid (Clasen and De Moor, 2015). Conventional automated approaches such as Grid Search have been prevalent for hyperparameter optimization (HPO). However, these methods often prove to be exhaustive and inefficient, particularly when navigating vast HPO search spaces, leading to the curse of dimensionality (Cooper et al., 2021). To mitigate these challenges, advanced algorithmic strategies like Half-Grid Search, Genetic Algorithms (GA) and Simulated Annealing (SA) have emerged as promising alternatives, offering more sophisticated solutions to the complexities of HPO.

This project aims to compare various hyperparameter optimization techniques, specifically Half-Grid Search, GA and SA, against Grid Search. The comparison will focus on performance and runtime characteristics, utilizing real-world problems for a comprehensive evaluation.

2. Problem Statement

One of the standard approaches to fine-tune the hyperparameters is by using Grid Search, an algorithm that conducts an exhaustive search across a specified subset of the hyperparameter space of the training algorithm (Petro and Pavlo, 2019). Although Grid Search is capable of evaluating all possible combinations to select the best-performing one, it is highly time-consuming and computationally expensive. The issue becomes worse when there are numerous hyperparameters to test across a broad range of values, leading to increased processing time as it iterates through every possible hyperparameter combination (Soleymani, 2023). This phenomenon is often referred to as the curse of dimensionality. Consequently, a variety of advanced optimization techniques have been developed to tackle this problem. Each technique employs different strategies aimed at improving efficiency and reducing the resources required to identify optimal hyperparameters.

Novel methods such as evolution strategies like GA, hyperband like Half-Grid Search and adaptive explore-exploit trade-offs method like SA can be used in place of Grid Search when solving optimization problems (Bischl et al., 2023). These methods have their own strategies that can utilise exploration vs exploitation trade-offs when finding the suitable set of hyperparameters. It started with exploring a vast hyperparameter search space to seek potential solutions, and then slowly reducing the search space around the solutions to find the better one (Rocca, 2021). Even though there are research papers that are focused on using these techniques to improve HPO performance for a certain ML model, there is still a lack of research on comparing Half-Grid Search, GA, and SA with Grid Search in HPO.

Therefore, this project aims to fill the gap in the research of HPO by addressing the following problems when comparing Grid Search against Half-Grid Search, GA and SA:

1. **Computational Inefficiency in HPO**: Traditional grid search is computationally expensive and often impractical for high-dimensional search spaces.
2. **Inadequate Exploration and Exploitation in Hyperparameter Search**: Grid search may not effectively balance exploration and exploitation, potentially missing optimal hyperparameter configurations.
3. **Scalability Challenges in HPO**: As the complexity and dimensionality of models and problems increase, the scalability of Grid Search method becomes a significant concern.

3. Objectives

In order to explore and compare novel optimization methods, specifically Half-Grid Search, GA and SA, for hyperparameter tuning in ML models against traditional Grid Search method, the following objectives have to be achieved:

1. **Measure** and **compare** the average computation time taken by each method to find the optimal set of hyperparameters.
2. **Assess** the quality of hyperparameters found by each method by evaluating the performance of the resulting models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
3. **Quantify** the diversity of hyperparameters evaluated by each method to assess their exploratory capabilities.
4. **Evaluate** the performance of each method as the dimensionality of the hyperparameter space increases.

By doing so, the study can uncover insights into the relative advantages and limitations of each method, which provides a nuanced understanding of their applicability and performance in searching for suitable hyperparameters.

4. Data Science Methodology

Cross Industry Standard Process for Data Mining (CRISP-DM) is a conventional process model that serves as a foundation for a data science project. It provides step-by-step guidance on how to conduct the project from start to finish, which can systematize the process flow and ensure a high-quality result from the project (Wirth and Hipp, 2000). The model consists of six iterative phases, which are Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment.

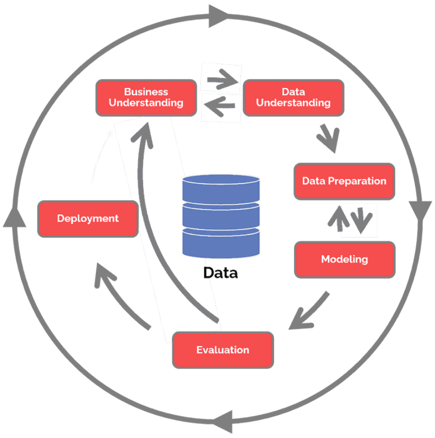

Figure 1: An overview of CRISP-DM model (Hotz, 2023)

4.1. Business Understanding

Understanding the business context is the critical first step in any project. It involves a thorough comprehension of the project's goals and requirements, as well as establishing clear success criteria (Wirth and Hipp, 2000). This foundational stage is facilitated by defining the problem statement and objectives.

The focus of this project is to conduct a comparative analysis of Half-Grid Search, GA and SA against traditional Grid Search within the domain of HPO. By addressing the known challenges associated with Grid Search in HPO, this project aims to enhance efficiency and effectiveness in model tuning. Success will be measured by the project's ability to meet its objectives and demonstrate the comparative advantages of these HPO techniques through various metrics.

4.2. Data Understanding

Data understanding is a process that consists of data collection, exploration and verification to ensure comprehensive knowledge of the dataset and to confirm that there are no issues with its quality. (Schröer et al, 2021). Drawing from the hyperparameter tuning experiments conducted by Elgeldawi (2021) and Bhandare and Hajiarbabi (2023), which optimized the hyperparameters of classifiers such as the Random Forest Classifier, Light Gradient Boosting Machine (LGBM) Classifier and Support Vector Machine Classifier (SVC), it is clearly seen that these models are well-suited for classification problems. A similar approach will be adopted in this project.

For this project, a credit card dataset (Goyal, 2020) from Kaggle will be used to build various ML models to predict customer churn. This dataset comprises data from 10,000 customers with 18 features. Exploratory Data Analysis (EDA), including both univariate and bivariate analysis, will be conducted to gain a thorough understanding of the data. This step is crucial for assessing data quality and informing subsequent data cleaning procedures. Additionally, valuable features can be extracted by analyzing the relationships between features. This can be used to enhance the predictve performance of the ML models.

4.3. Data Preparation

Data preparation is often a time-consuming phase, as it involves multiple steps to ready the raw data for processing and analysis. These steps include standardizing the data format, handling missing data and removing outliers (Bandgar, 2021). Aside from that, since some data are categorical, it is necessary to perform mapping and one-hot encoding to ensure that the ML models can interpret the data correctly. Another challenge is imbalanced data, which can compromise the quality of predictions. Therefore, techniques such as the Synthetic Minority Oversampling Technique (SMOTE) will be employed to address this issue. Feature scaling will also be conducted to enhance training speed. Finally, the data will be divided into training and testing sets for the purpose of modelling.

4.4. Modelling

The credit card dataset will be processed using various classifiers, including ensemble methods such as the Random Forest Classifier, LGBM Classifier and XGBoost, as well as SVM and Logistic Regression. These ML architectures vary in the number of hyperparameters that require tuning, making them ideal candidates for an HPO experiment designed to test the capabilities of Half-Grid Search, GA, SA, and Grid Search.

The experiment will begin with ML models that have fewer hyperparameters and a narrower range of values, and will gradually progress to models with a larger number of hyperparameters and a broader search space. Python code will be developed to implement the different HPO techniques in the search for the optimal hyperparameters for the classifiers.

At each epoch, K-fold cross-validation will serve as the primary method for evaluating the performance of the classifier after optimization by the HPO techniques (Guo et al., 2019). Depending on the strategies employed by the HPO methods, some suboptimal hyperparameters will be retained to facilitate the exploration of improved solutions, while more promising hyperparameters will be preserved to refine the search for the optimal solution.

4.5. Evaluation

To ensure a fair evaluation of the different HPO techniques, each round of the experiment will be conducted with the same seed and identical data splits for training and testing. The same computational resources will be allocated, and each experiment will be repeated multiple times to account for variability in performance. A baseline model that uses default hyperparameters will be included to serve as a reference point for the enhancements provided by each HPO technique.

The efficacy of Half-Grid Search, GA, SA and Grid Search in HPO will be assessed based on their computation time and performance metrics such as accuracy, precision, recall, F1-score and ROC-AUC of the resulting models. Additionally, the number of unique sets of hyperparameters that provide optimal performance identified by each method will be evaluated after several iterations. The findings will be organized in a tabular format and added with graphical representations for clear understanding and analysis.

4.6. Deployment

The outcomes and methodologies of the experiments will be documented in a scholarly article. This will serve as a comprehensive guide for researchers who want to do similar experiments. It provides them with the necessary framework to replicate and refine the experiments for comparing various HPO techniques.

An open-source framework, such as Streamlit, will serve as the primary platform for presenting the findings from the comparative analysis of the HPO methods. This presentation will consist of the problem statement, objectives, dataset background, experimental design and the results of the experiments. The well-performed classifiers that utilize hyperparameters identified by each HPO technique will be made available for public use. Users will have the capability to input data and employ these models for classification purposes, with the extra feature of displaying a confidence score for each prediction.

In addition, the source code for executing the HPO on the classifiers will be made publicly available on GitHub. This will enable developers worldwide to access and integrate these methods into their projects as needed. A link to the GitHub repository will also be provided on the Streamlit website for easy access.

References

Bandgar, S. (2021, May 28). *Data Preparation in Data Science*. Medium. https://medium.com/analytics-vidhya/data-preparation-in-data-science-16f9311760

Bhandare, Y. S., & Hajiarbabi, M. (2023). *Hyperparameter Tuning with Simulated Annealing and Genetic Algorithm*. Social Science Research Network. https://doi.org/10.2139/ssrn.4432719

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A., Deng, D., & Lindauer, M. (2023). Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges. *WIREs Data Mining and Knowledge Discovery, 13*(2), e1484. https://doi.org/10.1002/widm.1484

Claesen, M., & De Moor, Bart. (2015). *Hyperparameter Search in Machine Learning*. ArXiv.org. https://arxiv.org/abs/1502.02127

Cooper, A. F., Lu, Y., Forde, J. Z., & De Sa, C. (2021). *Hyperparameter Optimization Is Deceiving Us, and How to Stop It.* ArXiv.org. https://doi.org/10.48550/arXiv.2102.03034

Du, X., Xu, H., & Zhu, F. (2021). Understanding the Effect of Hyperparameter Optimization on Machine Learning Models for Structure Design Problems. *Computer-Aided Design*, *135*, 103013. https://doi.org/10.1016/j.cad.2021.103013

Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. *Informatics, 8*(4), 79. https://doi.org/10.3390/informatics8040079

Hoque, K. E., & Aljamaan, H. (2021). Impact of Hyperparameter Tuning on Machine Learning Models in Stock Price Forecasting. *IEEE Access, 9*, 163815-163830. https://doi.org/10.1109/access.2021.3134138

Hotz, N. (2023). CRISP-DM Diagram [Online image]. Data Science Process Alliance. https://www.datascience-pm.com/wp-content/uploads/2021/02/CRISP-DM.png

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated Machine Learning: Methods, Systems, Challenges. The Springer Series on Challenges in Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access, 5*, 20590–20616. https://doi.org/10.1109/access.2017.2756872

Guo, B., Hu, J., Wu, W., Peng, Q., & Wu, F. (2019). The Tabu_Genetic Algorithm: A Novel Method for Hyper-Parameter

Optimization of Learning Algorithms. *Electronics, 8*(5), 579. https://doi.org/10.3390/electronics8050579

Petro, L., & Pavlo, L. (2019). *Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS*. ArXiv.org. https://doi.org/10.48550/arXiv.1912.06059

Rocca, J. (2021, May 23). *The Exploration-exploitation Trade-off: Intuitions and Strategies*. Towards Data Science. https://towardsdatascience.com/the-exploration-exploitation-dilemma-f5622fbe1e82

Sagiroglu, S., & Sinanc, D. (2013). Big Data: A Review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/cts.2013.6567202

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Soleymani, A. (2023, February 13). *Grid Search and Random Search are Outdated. This Approach Outperforms Both*. Medium. https://medium.com/@ali.soleymani.co/stop-using-grid-search-or-random-search-for-hyperparameter-tuning-c2468a2ff887

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science, 181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Goyal, S. (2020). *Credit Card Customers* [Data set]. https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers

Wirth, R. & Hipp, J. (2000, April 11-13). *CRISP-DM: Towards a Standard Process Model for Data Mining* [Paper presentation]. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester. https://cs.unibo.it/ danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Yang, L., & Shami, A. (2020). On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice. *Neurocomputing, 415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061