Scholar Publishing Group
Machine Learning Theory and Practice
https://doi.org/10.38007/ML.2025.050112
ISSN 2790-0983 Vol. 5, Issue 1: 117-124

Optimization of Machine Learning Models and Application Supported by Data Engineering

Xindi Wei

Pepperdine Graziadio Business School, Master of Science in Business Analytics, Malibu, California, 90263, USA

Keywords: data engineering; Machine learning model; Feature engineering; Hyperparameter tuning

Abstract: With the rapid progress of intelligent technology, machine learning is increasingly widely used in all walks of life, and the role of data engineering is increasingly prominent, becoming the core link that determines the efficiency and practicability of the model. The quality, purification, storage and control of data are directly related to the quality and speed of model training. The optimization of techniques, such as feature extraction, hyperparameter fine-tuning, regularization processing, etc., continues to promote the leap in model performance. Data engineering is not only the basis of model training, but also plays an indispensable role in the subsequent steps of model deployment, monitoring and iterative upgrading. This paper introduces the practical application effect of machine learning model enhanced by means of feature engineering, distributed computing, big data environment, etc., in order to promote the popularization and deepening development of intelligent technology.

1. Introduction

In the era of widespread application of machine learning technology, data engineering plays an indispensable role in improving algorithm performance. The accuracy of data, cleaning process, processing method and storage method directly determine the quality and effectiveness of algorithm training. Techniques such as feature extraction, preprocessing, and distributed computing continue to upgrade the performance of machine learning algorithms. With the emergence of massive data, efficient processing and management of these massive data has become a major challenge. Relying on the help of data engineering, machine learning can effectively solve complex problems in many industries and promote the development of industrial intelligence.

2. The central role of data engineering in machine learning

2.1. Impact of data quality on machine learning models

The performance of the model depends first on the quality of the data used. Good data can input more precise information into the model, helping it to grasp the underlying patterns in the data. Omissions, contradictions or interference in the data will weaken the training effect of the model, and then affect its performance. For example, when performing a classification task, if the label data is wrong, it will lead to model bias and affect the accuracy of the prediction, while in regression analysis, abnormal data points may cause the model to overfit abnormal cases and lose its grasp of the general trend. Improving data quality is the fundamental way to enhance model performance. In the process of data collection, sorting, and subsequent processing and analysis, data quality must be strictly monitored to ensure the accuracy and consistency of data, so as to provide solid support for model training and prediction.

2.2. Role of data storage and management in model optimization

In the field of machine learning, data storage and management play a crucial role, especially when dealing with huge data sets. Since conventional independent storage devices cannot cope with the needs of large-scale data processing, it is particularly important to adopt distributed storage systems such as HDFS and NoSQL databases, which can save and manage a large amount of data efficiently. A proper storage architecture can speed up the reading and processing of data and mitigate the performance limitations caused by storage access delays. Data management also covers aspects such as data versioning, data quality monitoring, and data lifecycle management, which ensure data consistency and availability throughout the whole process of machine learning. Optimizing data storage and management efficiency can promote parallel data processing capabilities in distributed computing environments, thereby increasing the speed of model training and inference.

3. Optimization method of machine learning model based on data engineering

3.1. Feature engineering and data preprocessing optimization

Feature engineering and data preprocessing play a crucial role in optimizing the performance of machine learning algorithms. By filtering, transforming or generating new features on the original information, the training effect of the model can be effectively enhanced. Feature engineering covers several steps, such as selecting key features, constructing new features and extracting information, so as to improve the learning efficiency of the model and the accuracy of prediction. Data preprocessing involves eliminating data impurities, supplementing incomplete data, achieving unified scale and standardization of data, etc., to ensure the consistency and quality of data input to the model.

The accuracy of the model is closely related to the quality of the feature engineering, and the optimized features can show the essential regularity of the data. When dealing with some difficult tasks, proper feature screening can eliminate unnecessary and irrelevant features, thus simplifying the structure of the model and enhancing the speed of calculation. Through the effective extraction of features, the model can grasp the deep information contained in the data more deeply, so as to enhance the accuracy of its prediction. The accuracy of the model will increase with the improvement level of feature engineering, so there is a positive correlation between the effectiveness of feature engineering and data preprocessing and the accuracy of the model. This relationship can be expressed by the following formula.

$$Model\text{ Accuracy} = \sum_{i=1}^{n} f\left(Feature_i\right) \tag{1}$$

In formula (1), $Model\ Accuracy$ Represents the accuracy of the model, $f\left(Feature_i\right)$ Is the contribution of each optimization feature to the model effect.With the gradual improvement of feature optimization, the accuracy of the model is gradually improved.

3.2. Hyperparameter tuning and model training optimization

In the construction of machine learning model, the selection of appropriate hyperparameters plays a decisive role in the effectiveness and performance of the model. Adjusting these parameters can greatly improve the efficiency and accuracy of model training. Different combinations of common hyperparameters such as learning rate, batch size, tree depth, etc., will bring different effects on model training. The common strategies for hyperparameter optimization include grid search, random search and Bayesian optimization, etc. Selecting the appropriate optimization algorithm is also crucial to the training effectiveness of the model. For example, gradient descent and its variants, such as the Adam optimizer, optimize training outcomes by optimizing model parameters to reduce the loss function. In a specific scenario, the application of early stop strategy can avoid overfitting phenomenon and enhance the generalization ability of the model. The optimization of hyperparameters is closely related to the training time, and the correlation between the learning rate and the training time can be expressed by the following formula.

$$Training\ Tine = \frac{\mathrm{C}}{\mathrm{Learning\ Rate}} + D \qquad (2)$$

In formula (2), C Is a constant term related to other factors in the training process, $Learning\ Rate$ Is the learning rate, which controls the size of the step at each update, D Is a constant term independent of hyperparameters.With the increase of learning rate, the training time is often shortened, but the higher learning rate may also lead to the instability of training, so we must find a moderate point between efficiency and stability. In addition, the optimization of hyperparameters plays a decisive role in the model effect, which can be further explained by the following formula.

$$Model\ Performance = \sum_{i=1}^{n} Hyperparameters_i \qquad (3)$$

In formula (3), $Model\ Performance$ Represents the final performance of the model, $Hyperparameters_i$ For each hyperparameter optimization result, n Is the number of hyperparameters. With the optimization of hyperparameters, the performance of the model can be continuously improved. By precisely adjusting the training parameters and optimizing the training strategy, the efficiency of the machine learning model can be effectively enhanced to better match the complex data characteristics in the actual scene.

3.3. Techniques for regularization and prevention of overfitting

By introducing additional disciplinary factors to the loss function, regularization techniques aim to control the complexity of the model and effectively prevent the model from overfitting on the training set. Overfitting will cause the model to learn too much about the random fluctuations in the training data, which will affect its prediction accuracy on the unknown data set. Common regularization methods include L1 regularization (Lasso) and L2 regularization (Ridge), which enhance the generalization ability of the model through their own unique parameter punishment mechanisms. L1 regularization can filter features by zeroing some model parameters, while L2 regularization limits model complexity by penalizing larger parameter values. The loss function of L2 regularization can be expressed by the following formula.

$$J(\theta)=L(\theta)+\lambda\sum_{i=1}^{n}\theta_i^2 \tag{4}$$

In formula (4), $J(\theta)$Is the total loss function, representing the overall loss of the model. $L(\theta)$Is the original loss function, usually the error function of the model, λ Is the regularization coefficient, controls the intensity of regularization, θ_i For the parameters of the model, n Is the number of model parameters. Regularization coefficient λ The bigger the,The stronger the regularization effect, the lower the complexity of the model, thus avoiding overfitting.

In addition, regularization helps to maintain the balance between training errors and validation errors, effectively reducing the possibility of overfitting. If the error of the model on the training data set is significantly smaller than that on the validation data set, it indicates that the model may have overfitting phenomenon. In this case, the problem can be alleviated by strengthening regularization measures. The risk of overfitting can be quantified by the following formula.

$$Overfirring\ Risk=\frac{Training\ Error-Validation\ Error}{Training\ Error} \tag{5}$$

In formula (5), $Training\ Error$ Represents training error, $Validation\ Error$ Represents verification error, $Overfirring\ Risk$ Represents the risk of overfitting.If the training error is much smaller than the verification error, the model has a hidden danger of overfitting. Adopting proper regularization means can help to prevent this problem and guarantee the generalization performance of the model. The regularization and anti-overfitting strategies are helpful to improve the model's ability to cope with unknown data sets, strengthen its robustness and reliability, and ensure its better performance in practice.

3.4. Large-scale data processing and distributed computing optimization

With the increase in the amount of data, the traditional single-machine processing capacity is no longer enough to meet the high efficiency requirements of machine learning model training, while distributed computing technology, by subdividing the overall task into several small tasks for parallel computing, greatly improves the training speed, and can effectively manage large-scale data sets. Distributed computing frameworks, such as Hadoop and Spark, excel at processing and storing large-scale data.

The charm of distributed computing is not only in its computing speed, but also through the mutual assistance of resources and task allocation between nodes, which effectively overcome the difficulties of big data processing. It can also balance the computing load, optimize the efficiency of computing resources, and ensure the efficiency and scalability of training work. The logical relationship between distributed computing and training time can be described by the following formula.

$$T_{\text{total}}=\frac{T_{\text{single}}}{N} \tag{6}$$

In formula (6), $\mathrm{T}_{\text{total}}$ Is the total training time of distributed computing, $\mathrm{T}_{\text{single}}$ Training time calculated for a single machine, N Indicates the number of compute nodes.With the increase of the number of nodes, the training time is greatly reduced, thus improving the efficiency of model training.

4. Application practice of machine learning supported by data engineering

4.1. Application of machine learning models in the industry

Machine learning models have been deeply applied in many fields, which has promoted the digitization process of the industry and greatly promoted the improvement of operational efficiency. In the field of finance and economics, algorithm learning technology has been applied to risk assessment, cheating detection, market trend analysis and other links. Using past data and user behavior information, algorithms can quickly find possible risk points or business opportunities to assist financial institutions to make more accurate decisions.

In the retail industry, machine learning models are widely used to predict consumer trends, manage inventory, and customize recommendations. In the medical and health field, machine learning technology provides doctors with assistance in image diagnosis, disease prediction and the development of personalized medical solutions, especially under the promotion of deep learning, automated image diagnosis has greatly improved the speed and accuracy of disease detection. Table 1 below lists the classic applications and results of machine learning in different industries, demonstrating the utility and benefits of this technology across industries.

Table 1 Overview of machine learning applications in the industry

profession	Application field of machine learning	Application effect
Financial industry	Credit scoring, fraud detection, market forecasting	Improve decision accuracy and reduce risks
Retail industry	Customer behavior prediction, inventory management, recommendation system	Increase sales and optimize inventory management
Medical industry	Image recognition, disease prediction, personalized treatment	Improve diagnostic accuracy and reduce misdiagnosis rate
Manufacturing industry	Production process optimization and equipment maintenance prediction	Improve production efficiency and reduce maintenance costs
Energy industry	Energy management and equipment monitoring	Reduce power consumption and improve device utilization

Success stories in these industries show that machine learning has become a core driver of innovation across all sectors.

4.2. Model deployment and management in big data environment

In the big data environment, the deployment and management of machine learning models encounter many difficulties, and the rapid expansion of data puts forward the requirements of high data processing capacity for the delivery platform. Distributed computing frameworks, such as Hadoop and Spark, have become mainstream solutions for processing huge amounts of data, which can realize distributed storage and parallel processing of data, greatly speeding up the speed of data processing. The delivery of models must be efficient and flexible in order to quickly adapt to the changing needs of the big data environment.

In large-scale data processing scenarios, the role of supervision and management is indispensable to ensure the continuous and stable operation of models. After the model is put into use, it is necessary to continuously observe the data flow dynamics and algorithm adaptability in order to make timely adjustment and improvement. For example, impurities in the data stream or shifts in the data distribution may affect the predictive effect of the model, so it is particularly important to establish an automated model updating and retraining mechanism. Table 2 below lists some common strategies and tools for deploying and managing machine learning algorithms when working with big data.

Table 2 Model deployment and management strategies in a big data environment

Strategies/Tools	Description	advantage
Distributed computing platform	Use Hadoop, Spark and other platforms for distributed data processing	Improve data processing efficiency and support large-scale computing
Containerized deployment	Deploy models using container technologies such as Docker and Kubernetes	Supports rapid deployment, update, and expansion of models
Real-time monitoring	Model monitoring with tools such as Prometheus and Grafana	Track model performance in real time to identify potential problems
Automatic retraining	Automate model updates with tools such as Airflow	Improve long-term performance by automatically updating models based on new data

These deployment and management strategies significantly improve the ability to cope with big data scenarios, and ensure that the model maintains its superior performance as the business and data environment continues to evolve.

4.3. Real-time data stream processing and online learning applications

Real-time data stream processing and online learning technologies are widely used in many industries, especially those where the data is updated quickly and the need for timeliness is high. This technology ensures that data is processed as soon as it is generated, and is particularly suitable for dynamic monitoring of financial markets and personalized recommendation systems for e-commerce. Online learning techniques optimize the model by updating the data in real time, avoiding complete retraining of the model, showing a high degree of flexibility and the ability to quickly adapt to changes in the data. The integration of these two technologies can greatly improve the model's response speed and adaptation level to information, especially when dealing with massive data, the system can receive data immediately and give feedback quickly. Table 3 below lists the specific application areas and advantages of real-time data stream processing and online learning technologies.

Table 3 Application scenarios and advantages

Application scenario	Technical means	advantage
Financial transaction monitoring	Real-time data flow processing	Respond to market changes instantly, reduce risk and improve forecasting accuracy
E-commerce recommendation system	Online learning	Improve personalized recommendation accuracy and increase conversion rate
Social media analysis	Online learning and real-time data flow processing	Adjust recommendations in real time to enhance user engagement and engagement

Through the combination of these technologies, the system can effectively improve the response speed and ensure the stable performance of machine learning models in the dynamic data flow.

4.4. Implementation of model monitoring and continuous optimization

During the continuous application of machine learning systems, model monitoring and continuous optimization are particularly critical. Through the continuous tracking of the model performance, the performance reduction or deviation can be quickly identified, and these problems can be corrected or retrained in time. To reduce the reliance on humans, many companies have implemented automated model retraining processes that adjust parameters based on the latest data or model errors to improve model accuracy and reliability. Table 4 below shows the strategies commonly used in model monitoring and optimization and their effectiveness.

Table 4 Model monitoring and optimization strategies

Strategies/Tools	Description	advantage
Monitoring system	Monitor model performance in real time	Identify problems quickly to ensure long-term model stability
Automatic update	Automatic retraining and parameter optimization	Improve optimization efficiency and reduce manual intervention
Log analysis	Analyze model run logs and data changes	Identify potential problems and make adjustments in a timely manner

By implementing these monitoring and optimization strategies, it is possible to adapt machine learning models to changes in the environment, thus ensuring their long-lasting high performance and reliability.

5. Conclusion

As an important cornerstone in the field of machine learning, data engineering greatly improves the optimization degree of algorithms and the operation efficiency of systems through advanced data processing and management skills. With the pace of technological progress, machine learning technology is gradually penetrating into many industries, especially under the integration of real-time analysis of massive data and online learning, which is expected to promote many industries to a new stage of intelligent decision-making. The effective monitoring and continuous optimization of the model is the key to ensure the lasting stability and efficient operation of the machine learning system. Therefore, deepening the deployment and management of machine learning models supported by data engineering will bring unlimited innovation potential and economic benefits to various industries.

References

[1] *Palma G R , Thornberry C ,Seán Commins,et al.Understanding Learning from EEG Data: Combining Machine Learning and Feature Engineering Based on Hidden Markov Models and Mixed Models[J].Neuroinformatics, 2024, 22(4):487-497.*

[2] *Purbey R , Parijat H , Agarwal D ,et al.Machine learning and data mining assisted petroleum reservoir engineering: a comprehensive review[J].International Journal of Oil, Gas and Coal Technology: IJOGCT, 2022(4):30.*

[3] *Anaraki F , Hariri-Ardebili M , Becker S ,et al.Call for Special Issue Papers:Big Scientific Data and Machine Learning in Science and Engineering.[J].Big data, 2021, 9(5):404-405.*

[4] *Habib M , Okayli M .Evaluating the Sensitivity of Machine Learning Models to Data Preprocessing Technique in Concrete Compressive Strength Estimation[J].Arabian journal for science and engineering, 2024(10):49.*

[5] *Klamrowski M M , Klein R , Mccudden C ,et al.Derivation and Validation of a Machine Learning Model for the Prevention of Unplanned Dialysis[J].Clinical Journal of the American Society of Nephrology, 2024, 19(9):1098-1108.*

[6] *K. Zhang, "Optimization and Performance Analysis of Personalized Sequence Recommendation Algorithm Based on Knowledge Graph and Long Short Term Memory Network," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-6, doi: 10.1109/IACIS65746.2025.11211298.*

[7] *Y. Zhao, "Design and Financial Risk Control Application of Credit Scoring Card Model Based on XGBoost and CatBoost," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-5, doi: 10.1109/ICICNCT66124.2025.11233033.*

[8] *B. Li, "Research on the Spatial Durbin Model Based on Big Data and Machine Learning for Predicting and Evaluating the Carbon Reduction Potential of Clean Energy," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-5, doi: 10.1109/ICICNCT66124.2025.11232698.*

[9] *Q. Xu, "Implementation of Intelligent Chatbot Model for Social Media Based on the Combination of Retrieval and Generation," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-7, doi: 10.1109/IACIS65746.2025.11210989.*

[10] *Y. Zou, "Research on the Construction and Optimization Algorithm of Cybersecurity Knowledge Graphs Combining Open Information Extraction with Graph Convolutional Networks," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-5, doi: 10.1109/IACIS65746.2025.11211353.*

[11] *M. Zhang, "Research on Joint Optimization Algorithm for Image Enhancement and Denoising Based on the Combination of Deep Learning and Variational Models," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-5, doi: 10.1109/ICICNCT66124.2025.11232800.*

[12] *W. Han, "Using Spark Streaming Technology to Drive the Real-Time Construction and Improvement of the Credit Rating System for Financial Customers," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India, 2025, pp. 1-6, doi: 10.1109/ICICNCT66124.2025.11232932.*

[13] *J. Huang, "Research on Multi-Model Fusion Machine Learning Demand Intelligent Forecasting System in Cloud Computing Environment," 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-7, doi: 10.1109/IACIS65746.2025.11210946.*

[14] *J. Huang, "Performance Evaluation Index System and Engineering Best Practice of Production-Level Time Series Machine Learning System," 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), Bidar, India,*

[15] *X. Liu, "Research on User Preference Modeling and Dynamic Evolution Based on Multimodal Sequence Data," 2025 2nd International Conference on Intelligent Algorithms for*

Computational Intelligence Systems (IACIS), Hassan, India, 2025, pp. 1-7, doi: 10.1109/IACIS65746.2025.11211273.