World Journal of Innovation and Modern Technology, Vol. 8, Issue 4, (Apr) 2025
ISSN 2682-5910
DOI: 10.53469/wjimt.2025.08(04).08

Research on an Adaptive Curriculum Learning Method for Imbalanced Tabular Data Classification

Wing-Yee Lam[1], Ka-Yan Cheung[2], Tsz-Hin Lau[3], Man-Kit Leung[4], Ho-Lam Chan[5,*]

[1]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong SAR, China
[2]School of Data Science, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
**Correspondence Author, hlchan@cityu.edu.hk*

Abstract: ***In practical applications, tabular data commonly show a severe imbalance in class distribution, which causes great difficulties for traditional classification models in recognizing minority class samples. This study proposes an adaptive curriculum learning method based on sample difficulty modeling. The method ranks training samples accurately and applies a stage-wise weight control strategy to guide the model to learn progressively from easy to hard samples. Experiments conducted on several public tabular datasets, including Adult, Credit, and Census Income, show that the proposed method achieves improvements of 6.4% in F1 score and 5.1% in AUC compared with existing baseline algorithms. These results demonstrate the superior generalization ability and minority class recognition performance of the proposed method.***

Keywords: Machine learning; Imbalanced data; Curriculum learning; Tabular data classification; Sample difficulty modeling.

1. INTRODUCTION

With the rapid development of digitalization, tabular data, as a key form of structured information, has been widely and deeply applied in many core fields, including finance, healthcare, e-commerce, transportation, and education [1]. For example, in the financial sector, by analyzing customers' balance sheets and transaction records, banks can significantly improve the accuracy of credit default risk prediction from around 50% (random guessing) to 70% – 80% [2]. This is critical for protecting financial institutions' assets and improving economic efficiency [3]. In the medical field, doctors rely on patients' medical records and test reports in tabular form to maintain diagnostic accuracy between 85% and 95%, which provides an important basis for treatment planning and prognosis assessment, effectively safeguarding patients' health [4]. E-commerce platforms can increase product click-through rates by 20% – 30% and boost sales by 15% – 25% through mining tabular data such as users' purchase histories and browsing behavior. In the transportation domain, optimizing traffic scheduling and planning with vehicle operation records and passenger flow data in tabular form can reduce the average travel time on congested roads by 15% – 20% [5]. In the education sector, analyzing students' grade reports and learning behavior data supports teaching quality evaluation and personalized instruction, helping improve average student scores by 10 to 15 points. In real-world applications, the goal of tabular data classification is to predict the class of a data sample based on its attribute features in the table [6]. For instance, in credit risk assessment, a model predicts whether a customer will default based on features such as age, income and credit history [7]. In computer-assisted diagnosis systems, disease types are identified based on tabular data containing patients' physiological indicators and medical histories [8]. However, real-world tabular datasets often suffer from class imbalance, where the numbers of samples from different classes vary significantly. In medical diagnosis datasets, the ratio of rare disease cases to common ones can reach 1:100 or even lower. In fraud detection for e-commerce, fraudulent orders may only account for 1% – 5% of total orders. This class imbalance presents serious challenges for traditional classification models. During training, the model tends to focus excessively on majority class samples while ignoring the learning of features from minority class samples, which leads to poor classification performance for the minority class [9]. In extreme cases of imbalance, a model may achieve high overall accuracy by predicting all samples as the majority class. However, this does not meet the practical demand for accurately identifying minority class samples. For example, in early cancer screening, if the model classifies all patients as healthy (majority class), it may reach high accuracy, but will miss a large number of cancer patients (minority class), delaying treatment and causing severe consequences [10]. Therefore, how to effectively address the class imbalance problem in tabular data and improve the model's classification ability for the minority class has become a key challenge in machine learning research.

In recent years, the concept of curriculum learning has gradually gained attention in the machine learning field [11]. Its core idea is inspired by the human learning process, where learning starts from simple tasks and gradually moves to more difficult ones, which helps improve both learning efficiency and performance [12]. In machine learning, curriculum learning arranges training samples in a proper order so that the model learns from easy to hard examples [13]. This approach has shown potential in improving model generalization and training effectiveness. In image recognition, curriculum learning has been applied to handwritten digit classification by first training on clear and standard samples, and then gradually introducing blurred and distorted samples, which increased recognition accuracy from 70% to 90%. In text classification tasks in natural language processing, samples are sorted based on text length and vocabulary complexity, guiding the model to learn from simple to complex samples, resulting in an 8% – 12% increase in classification accuracy. Applying curriculum learning to imbalanced tabular data classification provides a new path to address the challenges caused by class imbalance [14]. By adaptively adjusting the learning order and sample weights, the model may better capture the features of minority class samples and improve classification performance [15]. However, how to accurately model the difficulty of tabular data samples and design an effective adaptive curriculum learning strategy remains insufficiently studied and is still a challenging task. Existing studies on sample difficulty estimation for tabular data often consider only a single factor and fail to comprehensively combine feature distribution and class information [16]. As a result, the sample ordering may not be reasonable enough, making it difficult to fully utilize the advantages of curriculum learning in imbalanced tabular data classification.

2. MATERIALS AND METHODS

2.1 Framework Overview

Modeling sample difficulty is pivotal for enabling adaptive curriculum learning, particularly in imbalanced tabular datasets where conventional uniform sampling often leads to suboptimal generalization. To address this, we propose a principled framework that evaluates sample difficulty by jointly analyzing local feature distribution and class-wise positional information. The difficulty score is computed by integrating (1) local density in the feature space and (2) the relative distance to the sample’ s class center. This dual perspective ensures both the statistical typicality and geometric deviation of each sample are quantitatively assessed.

2.2 Local Density Estimation Using Gaussian Similarity

Local density is designed to capture the compactness of a sample's neighborhood in the feature space. For a given sample $xi \in R^d$, its local density $D_{(xi)}$ is computed based on pairwise similarities with all other samples using a Gaussian kernel:

$$D(xi) = \sum_{j=1}^{n} \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right)$$

where n is the total number of samples, $||x_i - x_j||$ which represents the Euclidean distance between sample x_i and x_j in the feature space, σ is the bandwidth parameter controlling the kernel’s sensitivity to local distances. The Gaussian kernel is selected for its smooth exponential decay and well-established statistical properties in density estimation [17]. It ensures that samples with proximate neighbors yield higher density scores, reflecting higher representativeness in their local region. Empirically, this formulation effectively distinguishes cluster centers from edge cases in high-dimensional tabular datasets.

2.3 Distance to Class Center

While local density captures population-level regularity, it does not account for class-wise representativeness. Therefore, we further compute the Euclidean distance between a sample and the centroid of its respective class to quantify intra-class variation.

The class center cyi for class yi is defined as:

$$c_{y_i} = \frac{1}{|C_{y_i}|} \sum_{x_k \in C_{y_i}} x_k$$

and the corresponding distance score R(xi) for sample $xi \in C_{yi}$ is given by:

$$R(x_i) = ||x_i - c_{y_i}||$$

This distance provides geometric insight into a sample's deviation from its class prototype. Samples closer to the centroid are considered more typical, whereas distant samples are more likely to be class outliers or noise-affected instances [18]. This metric is especially meaningful in tabular domains such as risk assessment, where atypical cases often correspond to important but rare patterns.

2.4 Integrated Difficulty Score

To capture both global (class-based) and local (neighborhood-based) difficulty aspects, we introduce a composite score defined as:

$$S_i = \alpha \cdot \left(1 - \frac{\Pi_i}{\max(\Pi)}\right) + (1 - \alpha) \cdot \frac{d_i}{\max(d)}$$

where α is the balance coefficient, with a value range of [0, 1], It is used to adjust the weight of the influence of distance and density on sample difficulty. $D_{(xi)}$ is normalized by the maximum density in the dataset to ensure scale consistency. In this formulation, samples with large distances from the class center and low local density are assigned higher difficulty scores [19]. Conversely, samples that are both centrally located within their class and situated in dense regions are treated as easy. The parameter α can be tuned via cross-validation. Across multiple datasets, we observed that setting α=0.6 achieves a better alignment between computed difficulty scores and model convergence behavior, particularly in scenarios with overlapping classes and heterogeneous sample distributions.

3. EXPERIMENTS AND RESULT ANALYSIS

3.1 Experimental Settings

The proposed Adaptive Curriculum Learning (ACL) method was examined on three benchmark tabular datasets with inherent class imbalance: Adult, Credit, and Census Income. The Adult dataset, released by the U.S. Census Bureau, comprises 14 attributes and is widely used for binary income classification, where the positive class (income > USD 50,000) constitutes 24% of the samples. The Credit dataset contains 21 features and targets credit risk prediction, with approximately 30% of the instances labeled as default. The Census Income dataset includes 12 attributes and focuses on income-level classification, where the minority class proportion is around 20%. For all datasets, samples were randomly partitioned into training, validation, and test subsets in a 70:15:15. The original class distribution was preserved across splits to maintain consistency with the raw data distribution [20]. To ensure a rigorous evaluation, four representative methods for imbalanced classification were selected as baselines. SMOTE + SVM augments the minority class using the Synthetic Minority Over-sampling Technique with a 100% oversampling rate, followed by training a support vector machine. Random Under-Sampling + Decision Tree balances the class distribution by reducing the majority class size to match the minority, and applies a decision tree classifier. Cost-Sensitive Logistic Regression introduces class-dependent misclassification penalties, setting the cost of the minority class to five times that of the majority, and performs model training under this weighted loss [21,22]. AdaBoost employs decision trees as weak learners, with 50 boosting rounds, and reweights samples in successive iterations according to classification error. Model performance was quantified using the F1 score and the area under the ROC curve (AUC), both of which are standard metrics for imbalanced learning [23]. The calculation formula is as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

AUC is used to evaluate the performance of a classifier under different threshold settings [24]. It is calculated as the area under the ROC curve. The closer the value is to 1, the better the performance of the classifier.

3.2 Experimental Results

Table 1: F1 scores of different algorithms on three imbalanced tabular datasets

Dataset	SMOTE + SVM	Random Under-Sampling + Decision Tree	Cost-Sensitive Logistic Regression	AdaBoost	ACL
Adult	0.663	0.641	0.647	0.665	0.724
Credit	0.731	0.718	0.726	0.740	0.789
Census Income	0.606	0.612	0.635	0.629	0.687

The performance of the proposed Adaptive Curriculum Learning (ACL) method was compared against four baseline algorithms across three imbalanced tabular datasets: Adult, Credit, and Census Income. On the Adult dataset, ACL achieved an F1 score of 0.724, outperforming SMOTE + SVM (0.663), Random Under-Sampling + Decision Tree (0.641), Cost-Sensitive Logistic Regression (0.647), and AdaBoost (0.665). The inferior performance of SMOTE + SVM can be attributed to the distributional inconsistency between synthetic and real samples in complex feature spaces, which impairs the SVM's ability to establish an accurate decision boundary [25,26]. Under-sampling in combination with decision trees led to information loss in the majority class, affecting the model's ability to generalize [27]. Cost-sensitive logistic regression was limited by the rigidity of fixed misclassification costs, while AdaBoost suffered from overfitting due to sensitivity of its base learners to noise [28,29]. On the Credit dataset, ACL obtained the highest F1 score of 0.789, while SMOTE + SVM, Random Under-Sampling + Decision Tree, Cost-Sensitive Logistic Regression, and AdaBoost achieved 0.731, 0.718, 0.726 and 0.740, respectively. The results reflect the challenges these methods face when capturing minority class characteristics in high-dimensional financial data with complex feature interactions [30]. On the Census Income dataset, ACL again outperformed all baselines with an F1 score of 0.687, compared to 0.606 (SMOTE + SVM), 0.612 (Random Under-Sampling + Decision Tree), 0.635 (Cost-Sensitive Logistic Regression), and 0.629 (AdaBoost), highlighting the method's stability and adaptability in handling high-dimensional income data [31]. Averaged across all datasets, ACL achieved a 6.4% improvement in F1 score over the best-performing baseline in each case. In terms of AUC, ACL attained 0.813, 0.845, and 0.798 on the Adult, Credit and Census Income datasets, respectively, representing an average AUC improvement of 5.1% over competing methods. Detailed comparisons are presented in Table 2. By applying adaptive curriculum learning, the model enhances its ability to distinguish between majority and minority class samples and shows better robustness and adaptability when facing complex imbalanced data distributions, thus achieving higher AUC values.

Table 2: AUC Results of Different Algorithms on Each Dataset

Dataset	SMOTE + SVM	Random Under - Sampling + Decision Tree	Cost - Sensitive Logistic Regression	AdaBoost	ACL
Adult	0.765	0.751	0.758	0.770	0.813
Credit	0.794	0.780	0.788	0.799	0.845
Census Income	0.753	0.725	0.739	0.747	0.798

4. CONCLUSION

This study introduced an adaptive curriculum learning (ACL) method tailored for imbalanced tabular classification, incorporating a sample difficulty modeling approach that jointly considers local feature density and class-wise distance. By progressively guiding the model to learn from easier to more challenging instances through adaptive stage division and sample weighting, the proposed method enhances the model's capacity to recognize minority class samples. Experimental results across multiple public datasets consistently demonstrated superior performance in terms of both F1 score and AUC compared with established baseline algorithms, confirming the effectiveness of the approach. Nevertheless, limitations remain. The current difficulty modeling may not fully capture complex nonlinear feature interactions, and the stage-wise curriculum strategy may require further tuning for varying data characteristics. Future research may explore the integration of deep feature embeddings to improve difficulty estimation, the development of dynamic curriculum strategies driven by model feedback, and the extension of the framework to more complex tasks such as multi-label or multi-modal imbalanced classification, thereby broadening its applicability in real-world scenarios.

REFERENCES

[1] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic markets, 26, 173-194.

[2] Wang, Z., Yan, H., Wei, C., Wang, J., Bo, S., & Xiao, M. (2024, August). Research on autonomous driving decision-making strategies based deep reinforcement learning. In Proceedings of the 2024 4th International Conference on Internet of Things and Machine Learning (pp. 211-215).

[3] Onyshchenko, S., Zhyvylo, Y., Cherviak, A., & Bilko, S. (2023). DETERMINING THE PATTERNS OF USING INFORMATION PROTECTION SYSTEMS AT FINANCIAL INSTITUTIONS IN ORDER TO IMPROVE THE LEVEL OF FINANCIAL SECURITY. Eastern-European Journal of Enterprise Technologies, 125(13).

[4] Gao, D., Shenoy, R., Yi, S., Lee, J., Xu, M., Rong, Z., ... & Chen, Y. (2023). Synaptic resistor circuits based on Al oxide and Ti silicide for concurrent learning and signal processing in artificial intelligence systems. Advanced Materials, 35(15), 2210484.
[5] Mo, K., Chu, L., Zhang, X., Su, X., Qian, Y., Ou, Y., & Pretorius, W. (2024). Dral: Deep reinforcement adaptive learning for multi-uavs navigation in unknown indoor environment. arXiv preprint arXiv:2409.03930.
[6] Wang, S., Jiang, R., Wang, Z., & Zhou, Y. (2024). Deep learning-based anomaly detection and log analysis for computer networks. arXiv preprint arXiv:2407.05639.
[7] Gong, C., Zhang, X., Lin, Y., Lu, H., Su, P. C., & Zhang, J. (2025). Federated Learning for Heterogeneous Data Integration and Privacy Protection.
[8] Shih, K., Han, Y., & Tan, L. (2025). Recommendation System in Advertising and Streaming Media: Unsupervised Data Enhancement Sequence Suggestions.
[9] Zhu, J., Ortiz, J., & Sun, Y. (2024, November). Decoupled Deep Reinforcement Learning with Sensor Fusion and Imitation Learning for Autonomous Driving Optimization. In 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 306-310). IEEE.
[10] Bao, Q., Chen, Y., & Ji, X. (2025). Research on evolution and early warning model of network public opinion based on online Latent Dirichlet distribution model and BP neural network. arXiv preprint arXiv:2503.03755.
[11] Liu, Z., Costa, C., & Wu, Y. (2024). Data-Driven Optimization of Production Efficiency and Resilience in Global Supply Chains. Journal of Theory and Practice of Engineering Science, 4(08), 23-33.
[12] Zhu, J., Sun, Y., Zhang, Y., Ortiz, J., & Fan, Z. (2024, October). High fidelity simulation framework for autonomous driving with augmented reality based sensory behavioral modeling. In IET Conference Proceedings CP989 (Vol. 2024, No. 21, pp. 670-674). Stevenage, UK: The Institution of Engineering and Technology.
[13] Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., & Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. Journal of Machine Learning Research, 21(181), 1-50.
[14] Vepa, A., Yang, Z., Choi, A., Joo, J., Scalzo, F., & Sun, Y. (2024). Integrating Deep Metric Learning with Coreset for Active Learning in 3D Segmentation. Advances in Neural Information Processing Systems, 37, 71643-71671.
[15] Li, Z., Ji, Q., Ling, X., & Liu, Q. (2025). A Comprehensive Review of Multi-Agent Reinforcement Learning in Video Games. Authorea Preprints.
[16] Zhu, J., Wu, Y., Liu, Z., & Costa, C. (2025). Sustainable Optimization in Supply Chain Management Using Machine Learning. International Journal of Management Science Research, 8(1).
[17] Feng, H. (2024, September). The research on machine-vision-based EMI source localization technology for DCDC converter circuit boards. In Sixth International Conference on Information Science, Electrical, and Automation Engineering (ISEAE 2024) (Vol. 13275, pp. 250-255). SPIE.
[18] Zhang, W., Li, Z., & Tian, Y. (2025). Research on Temperature Prediction Based on RF-LSTM Modeling. Authorea Preprints.
[19] Li, Z. (2024). Advances in Deep Reinforcement Learning for Computer Vision Applications. Journal of Industrial Engineering and Applied Science, 2(6), 16-26.
[20] Liu, J., Li, K., Zhu, A., Hong, B., Zhao, P., Dai, S., ... & Su, H. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. Mediterranean Journal of Basic and Applied Sciences (MJBAS), 8(2), 243-260.
[21] Tang, X., Wang, Z., Cai, X., Su, H., & Wei, C. (2024, August). Research on heterogeneous computation resource allocation based on data-driven method. In 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS) (pp. 916-919). IEEE.
[22] Fernando, K. R. M., & Tsokos, C. P. (2021). Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Transactions on Neural Networks and Learning Systems, 33(7), 2940-2951.
[23] Zhu, J., Xu, T., Zhang, Y., & Fan, Z. (2024). Scalable Edge Computing Framework for Real-Time Data Processing in Fintech Applications. International Journal of Advance in Applied Science Research, 3, 85-92.
[24] Liu, Z., Costa, C., & Wu, Y. (2024). Leveraging Data-Driven Insights to Enhance Supplier Performance and Supply Chain Resilience.
[25] Aldeer, M., Sun, Y., Pai, N., Florentine, J., Yu, J., & Ortiz, J. (2023, May). A Testbed for Context Representation in Physical Spaces. In Proceedings of the 22nd International Conference on Information Processing in Sensor Networks (pp. 336-337).
[26] Feng, H. (2024). High-Efficiency Dual-Band 8-Port MIMO Antenna Array for Enhanced 5G Smartphone Communications. Journal of Artificial Intelligence and Information, 1, 71-78.

[27] Liu, Z., Costa, C., & Wu, Y. (2024). Quantitative Assessment of Sustainable Supply Chain Practices Using Life Cycle and Economic Impact Analysis.
[28] Yang, J., Chen, T., Qin, F., Lam, M. S., & Landay, J. A. (2022, April). Hybridtrak: Adding full-body tracking to vr using an off-the-shelf webcam. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
[29] Wang, G., Qin, F., Liu, H., Tao, Y., Zhang, Y., Zhang, Y. J., & Yao, L. (2020). MorphingCircuit: An integrated design, simulation, and fabrication workflow for self-morphing electronics. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 1-26.
[30] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. Progress in Artificial Intelligence, 5, 65-75.
[31] Pes, B. (2017, June). Feature selection for high-dimensional data: the issue of stability. In 2017 IEEE 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) (pp. 170-175). IEEE.