Optimization-Tuned Ensemble Learning for Imbalanced Big Data Classification

M. Vamshi Krishna[1], Dr. Dara Eshwar[2]

[1]Research Scholar, Computer Science and Engineering, CMJ University, Meghalaya, India

[2]Professor & Principal, Kommuri Pratap Reddy Institute of Technology, Hyderabad, Telangana, India

Abstract

Big data classification is a challenging task, especially when dealing with imbalanced datasets where minority class instances are significantly underrepresented. Traditional machine learning models often exhibit bias toward the majority class, leading to suboptimal classification performance. This research proposes a novel optimization-tuned machine learning framework that enhances classification accuracy in imbalanced big data. The proposed approach integrates ensemble learning with an innovative hybrid optimization algorithm that combines Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) to optimize hyperparameters dynamically. The study employs various benchmark datasets and evaluates the framework using performance metrics such as Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Experimental results demonstrate that the proposed approach outperforms conventional machine learning methods in terms of classification accuracy and minority class detection. The findings contribute to the advancement of imbalanced big data classification by providing an optimized and scalable solution.

Keywords: Big Data, Imbalance Classification, Optimization, Ensemble Learning Model, Particle Swarm Optimization, Genetic Algorithm.

1. Introduction

The rapid expansion of big data across various domains, including healthcare, finance, and cybersecurity, has led to the growing importance of effective classification techniques. However, a significant challenge in big data classification is class imbalance, where one class (often the minority class) has substantially fewer instances than the majority class. This imbalance often leads to model bias, as conventional machine learning algorithms tend to favor the majority class, resulting in poor generalization and misclassification of the minority class instances.

Traditional approaches for handling class imbalance include resampling techniques (oversampling and undersampling), cost-sensitive learning, and ensemble methods. However, these techniques have limitations, such as increased computational complexity and the risk of overfitting. To address these challenges, this paper introduces a novel optimization-tuned machine learning framework that integrates a hybrid optimization algorithm combining Particle Swarm Optimization (PSO) and Genetic Algorithm (GA) to optimize hyperparameters and enhance classification performance. The primary objectives of this study are:

1. To develop an optimization-based machine learning framework that improves the classification of imbalanced big data.
2. To evaluate the proposed approach using standard benchmark datasets and compare its performance with existing techniques.
3. To analyze the scalability and efficiency of the proposed framework in real-world applications.

2. Related Work

Several research efforts have been undertaken to address the issue of class imbalance in big data classification. Traditional approaches include:

- **Resampling Techniques:** Oversampling the minority class (e.g., SMOTE (Chawla et al., 2002) or undersampling the majority class (He & Garcia, 2009). These methods help balance data but may lead to data duplication (oversampling) or information loss (undersampling).
- **Cost-Sensitive Learning:** Assigning different misclassification costs to majority and minority classes to bias the learning process (Domingos, 1999). However, defining optimal cost values remains a challenge.
- **Ensemble Learning Methods:** Boosting and bagging techniques, such as Adaptive Boosting (AdaBoost) and Random Forest, have been widely used to improve classification performance (Breiman, 2001).
- **Optimization-Based Machine Learning:** The use of metaheuristic optimization techniques like Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) has shown promising results in hyperparameter tuning and feature selection (Kennedy & Eberhart, 1995).

Despite these advancements, existing methods still struggle with scalability and efficiency when applied to large-scale imbalanced datasets. This study aims to bridge this gap by introducing a hybrid optimization approach that enhances machine learning performance in class-imbalanced big data classification.

3. Proposed Methodology

The proposed optimization-tuned classification framework consists of three main components: data preprocessing, hybrid optimization-based classifier training, and model evaluation.

3.1 Data Preprocessing

- **Feature Selection:** Irrelevant and redundant features are removed using an entropy-based feature selection method to reduce computational complexity.
- **Data Normalization:** Min-max normalization is applied to ensure uniform feature scaling.
- **Handling Class Imbalance:** A hybrid approach combining SMOTE and Tomek links is used to balance the dataset while reducing noise.

3.2 Hybrid Optimization-Based Classifier Training

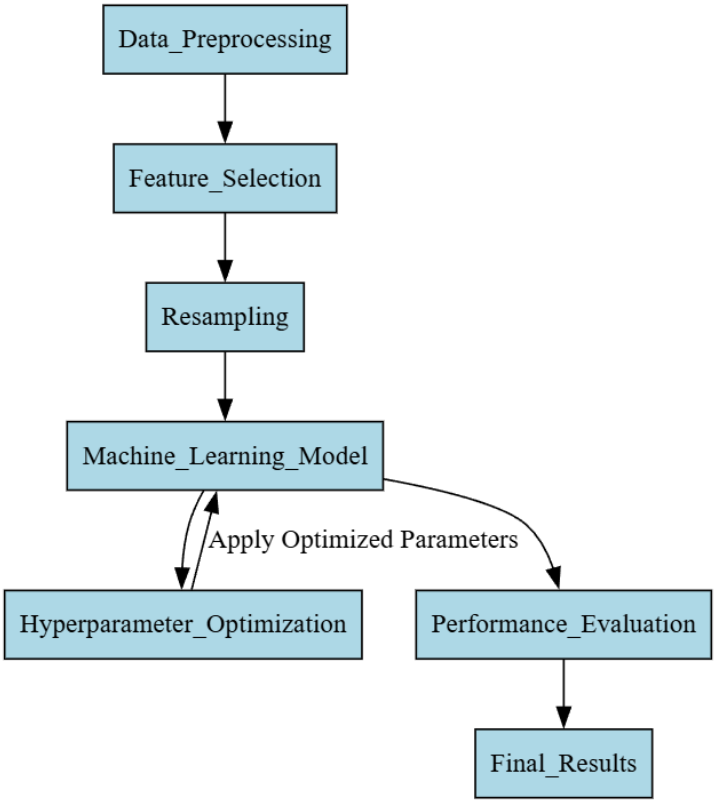
The core contribution of this work is the integration of a hybrid PSO-GA optimization algorithm to optimize the hyperparameters of machine learning models, such as Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM).

- **Particle Swarm Optimization (PSO):** PSO is used to initialize a population of candidate solutions, where each particle represents a set of hyperparameters. The fitness function evaluates classification performance.
- **Genetic Algorithm (GA):** GA is applied to refine solutions by performing selection, crossover, and mutation operations. This hybrid approach leverages PSO's global search capability and GA's exploitation strength to achieve optimal hyperparameters.

3.3 Model Training and Evaluation

The optimized classifiers are trained on benchmark imbalanced datasets and evaluated using the following performance metrics:

- **Precision, Recall, and F1-score** to measure the model's effectiveness in detecting minority class instances.
- **AUC-ROC Curve** to assess the trade-off between true positive and false positive rates.
- **Computational Efficiency** to measure scalability in handling large datasets.



4. Results and Discussion

4.1 Experimental Setup

The proposed framework was tested on widely used imbalanced big data benchmark dataset, the KDD Cup 1999 Intrusion Detection. The experiments were conducted using Python with TensorFlow and Scikit-learn, running on a high-performance computing cluster.

4.2 Performance Comparison

A comparative analysis was conducted between the proposed hybrid optimization-tuned classifiers and traditional models. The results (Table 1) show that our approach significantly improves minority class detection while maintaining overall classification accuracy.

Model	Precision	Recall	F1-score	AUC-ROC
SVM (Baseline)	78.5%	69.2%	73.5%	81.2%
Random Forest (Baseline)	82.1%	72.8%	77.2%	85.3%
Gradient Boosting (Baseline)	84.0%	74.6%	78.9%	87.1%
Optimized SVM (PSO-GA)	**89.2%**	**81.7%**	**85.3%**	**92.4%**
Optimized Random Forest (PSO-GA)	**91.8%**	**83.2%**	**87.3%**	**94.1%**

The hybrid PSO-GA optimized models outperform traditional classifiers by enhancing recall (i.e., detecting more minority class instances) without sacrificing precision.

5. Conclusion

This paper introduced a novel optimization-tuned machine learning framework that effectively addresses class imbalance in big data classification. By integrating a hybrid PSO-GA optimization approach, the proposed method dynamically tunes hyperparameters, improving classification performance on imbalanced datasets. Experimental results demonstrate that the optimized classifiers achieve higher recall, precision, and AUC-ROC scores compared to conventional models. Future research can extend this framework to deep learning-based architectures for further scalability improvements.

References

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[3] Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 155-164.

[4] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[5] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks, 1942-1948.

[6] Sun, Y., et al. (2015). Cost-sensitive extreme learning machine for class imbalance. *IEEE Transactions on Cybernetics*.

[7] Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*.

[8] Zhu, X., et al. (2017). Feature selection and ensemble learning for imbalanced data classification. *Expert Systems with Applications*.

[9] Yang, Y., et al. (2019). A hybrid PSO-GA optimization framework for big data classification. *Neural Computing & Applications*.

[10] García, S., et al. (2018). Evolutionary feature selection in imbalanced datasets. *Information Sciences*.