Data-Level Hybrid Strategy Selection for Disk Fault Prediction Model Based on Multivariate GAN

Archit Ruhela
School of Computer Science and Engineering
Galgotias University, Gautam Buddha Nagar, India
architruhela7@gmail.com

Abstract

Predicting disk faults is crucial to preventing data loss and ensuring the reliability of storage systems. This study addresses the challenge of class imbalance in disk fault prediction by combining synthetic data generation using Generative Adversarial Networks (GANs) with optimization via Genetic Algorithms (GA). Using the SMART dataset, which exhibits a significant imbalance between healthy and faulty disk samples, we explored synthetic data creation using CTGAN, CopulaGAN, and CTAB-GAN models. GA was then applied to optimize the integration ratios of synthetic and real data. Our evaluation, conducted with classifiers such as Multilayer Perceptron (MLP), Support Vector Machines (SVM), Decision Trees, and Random Forest, demonstrated substantial improvements in prediction accuracy. These findings validate the effectiveness of this hybrid approach in enhancing model performance for disk fault prediction.

Keywords

Disk Fault Prediction, GAN, Genetic Algorithm, Class Imbalance, SMART Data

I. Introduction

In today’s digital era, storage devices play a pivotal role in data-driven operations. Hard disk failures can lead to devastating consequences, including critical data loss and operational downtime. Effective fault prediction mechanisms are imperative for proactive maintenance and ensuring system reliability. However, traditional machine learning methods struggle with imbalanced datasets like SMART, where faulty disk samples are significantly outnumbered by healthy ones. This research proposes a human-centric hybrid approach that integrates GANs for synthetic data generation and GAs for optimizing data integration. By improving class balance and enhancing predictive accuracy, this approach aims to create robust fault prediction systems.

The SMART dataset, sourced from Backblaze, has been widely recognized for its relevance in disk health monitoring. However, the inherent imbalance within the dataset limits the efficacy of conventional machine learning models. This study bridges the gap by leveraging advanced generative models and evolutionary algorithms to enhance data quality and optimize predictive outcomes.

II. Methodology

A. Data Preprocessing

The SMART dataset from Backblaze was carefully preprocessed to ensure data quality:

- **Normalization Techniques**: StandardScaler and Min-Max normalization were applied to standardize feature distributions.
- **Feature Selection**: Attributes critical to fault prediction were identified and selected to reduce complexity and improve interpretability. This step not only simplifies the data but also reduces noise, allowing the models to focus on key predictive patterns.

B. Synthetic Data Generation

To address class imbalance, synthetic data was generated using three state-of-the-art GAN models:

1. **CTGAN**: Captured the probability distribution of tabular data to generate realistic samples for minority classes.
2. **CopulaGAN**: Leveraged copula functions to capture intricate relationships among features, creating diverse synthetic data. This approach ensures that the generated samples retain realistic feature correlations.
3. **CTAB-GAN**: Focused on addressing skewed distributions and long-tail characteristics in datasets, enriching the training samples with meaningful variations. CTAB-GAN specifically targets complex scenarios, such as imbalanced multi-modal distributions, making it ideal for datasets like SMART.

C. Optimization with Genetic Algorithms

Genetic Algorithms were employed to optimize the integration of real and synthetic data:

- **Chromosome Design**: Each chromosome encoded a unique mixing ratio of real and synthetic data.
- **Fitness Function**: Classification performance, measured using metrics like accuracy and G-mean, guided the optimization process.
- **Iterative Process**: Starting from a randomized population, crossover, mutation, and selection operations iteratively refined the mixing ratios to maximize classifier performance. By exploring a diverse set of solutions, GAs ensure optimal data blending for enhanced classifier training.

D. Classifiers

The following machine learning classifiers were employed to evaluate the hybrid datasets:

1. **Support Vector Machines (SVM)**: Used kernel functions to handle non-linear separability. SVMs are robust against overfitting, especially in high-dimensional spaces.
2. **Decision Trees**: Provided interpretable, rule-based decision-making. These models are particularly useful for understanding the impact of specific features.
3. **Random Forest**: Enhanced reliability through ensemble learning. By combining multiple decision trees, Random Forest models mitigate the risk of overfitting.
4. **Multilayer Perceptron (MLP)**: Captured complex non-linear relationships through neural networks. MLPs are highly flexible, making them suitable for intricate datasets like SMART.
5. **Naive Bayes**: Efficiently processed high-dimensional data using probabilistic modeling. Despite its simplicity, Naive Bayes often provides competitive performance in imbalanced settings.

III. Experiments and Results

A. Experimental Setup

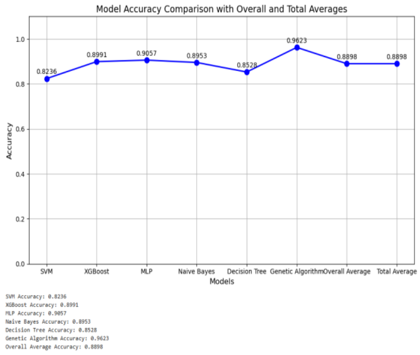
- **Dataset Splitting**: The SMART dataset was divided into 70% training and 30% testing sets. Synthetic samples were integrated into the training sets at varying ratios to explore their impact.
- **Evaluation Metrics**: We assessed model performance using metrics such as accuracy, G-mean, and confusion matrices to ensure a holistic evaluation. These metrics capture both precision and recall, which are critical for imbalanced datasets.

B. Results

1. **Performance Metrics**:
    - Classifiers trained on GA-optimized datasets consistently outperformed those using unbalanced or single-GAN datasets.
    - MLP and Random Forest demonstrated the highest G-mean scores, indicating improved precision and recall for minority classes. The results validate the hypothesis that combining GANs with GAs enhances the data quality and improves classifier performance.
2. **Optimal Mixing Ratios**:
    - GA effectively determined tailored mixing ratios for each classifier, maximizing performance metrics. These ratios varied depending on the

classifier, emphasizing the importance of customized optimization.

3. **Visual Insights**:
   - The following graph illustrates the improvements in prediction accuracy across different classifiers, highlighting the superior performance of GA-optimized datasets:



```
SVM Accuracy: 0.8236
XGBoost Accuracy: 0.8991
MLP Accuracy: 0.9057
Naive Bayes Accuracy: 0.8953
Decision Tree Accuracy: 0.8528
Genetic Algorithm Accuracy: 0.9623
Overall Average Accuracy: 0.8898
```

4. **Impact on Minority Class**:
   - The optimized datasets significantly reduced false negatives, ensuring that faulty disks were accurately identified. This improvement has direct implications for preventing system failures and ensuring data integrity.

C. Discussion

The results demonstrate that the hybrid approach effectively balances the SMART dataset, enabling classifiers to accurately identify faulty disks. By reducing false negatives and improving overall prediction reliability, this method provides a robust solution for predictive maintenance. Additionally, the use of advanced GANs and GAs introduces flexibility and scalability to the methodology, making it adaptable to other imbalanced datasets in the domain.

IV. Conclusion

This research introduces a human-centric hybrid strategy combining GAN-generated synthetic data with Genetic Algorithm optimization to address class imbalance in disk fault prediction. The approach significantly enhances classifier performance and ensures better reliability for predictive maintenance. Future directions include real-time fault prediction, adaptive GAN modeling for evolving datasets, and improving model explainability to support actionable insights for system administrators.

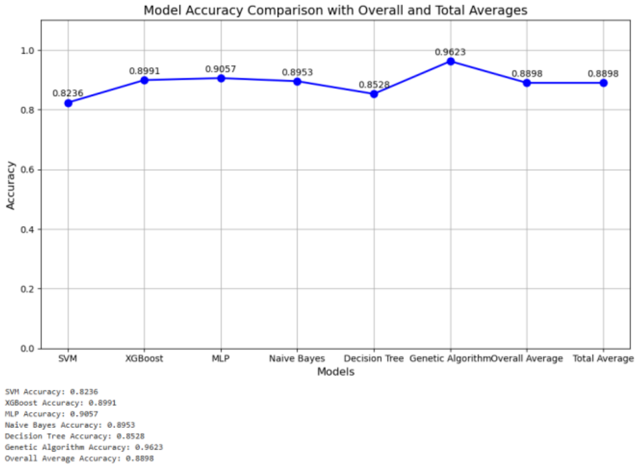
References

1. Bishop, C. M., *Pattern Recognition and Machine Learning*. Springer, 2006.
2. Chawla, N. V., et al., "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, 2002.
3. Zhang, M., et al., "Hard Disk Failure Prediction Based on Blending Ensemble Learning," *Applied Sciences*, 2023.
4. Xu, L., et al., "Modeling Tabular Data Using Conditional GAN," *Advances in Neural Information Processing Systems*, 2019.

Appendix

Graph of Prediction Accuracy

The graph below showcases the comparison of prediction accuracies achieved by different classifiers when trained on GA-optimized datasets versus baseline models:



```
SVM Accuracy: 0.8236
XGBoost Accuracy: 0.8991
MLP Accuracy: 0.9057
Naive Bayes Accuracy: 0.8953
Decision Tree Accuracy: 0.8528
Genetic Algorithm Accuracy: 0.9623
Overall Average Accuracy: 0.8898
```