Exploring SMOTE & Other Areas

Brian Cervantes Alvarez
April 23, 2024
Advisor Meeting

Introduction

Background Information

Briefly explain the importance of predictive modeling in education and the challenges associated with imbalanced data.

Problem Statement

Define the specific problem that S-SMOTE addresses and the relevance of comparing its performance across different machine learning models.

Objectives of the Paper

Outline the goals of the comparison study, focusing on evaluating the effectiveness and applicability of S-SMOTE.

Structure of the Paper

Provide a roadmap of the paper's layout and content.

Literature Review

Overview of SMOTE and Its Variants

Discuss the development and various adaptations of the SMOTE algorithm.

Previous Work on Machine Learning in Educational Data

Summarize key studies that have used SVM, GBM, and NN for predictive purposes in education.

Gaps in Current Research

Highlight the lack of comprehensive comparisons between these models using S-SMOTE.

Methodology

Data Description

Describe the dataset(s) used, including the source, variables, and any preprocessing steps.

Description of S-SMOTE Algorithm

Explain the mechanics of S-SMOTE and its intended benefits.

Machine Learning Models

Support Vector Machines

Gradient Boosting Machines

Neural Networks

Evaluation Metrics

Define the metrics for assessing model performance (e.g., accuracy, TPR, TNR).

Experimental Design

Experiment Setup

Outline the experimental framework and the parameters set for each model.

Training and Testing Procedure

Describe how the data will be split, the training process, and the testing phases.

Comparative Analysis Plan

Detail the methods for comparing the results across the different models.

Results

Model Performance

Present the performance results of each model using S-SMOTE.

Comparative Analysis

Discuss the strengths and weaknesses of each model in handling imbalanced data with S-SMOTE.

Discussion

Interpretation of Results

Analyze the implications of the comparative results.

Practical Implications

Discuss how these findings can be applied in real-world educational settings.

Limitations and Assumptions

Acknowledge any limitations in the study and potential biases in the models.

Conclusion and Future Work

Summary of Findings

Recap the key outcomes and their significance.

Recommendations for Future Research

Suggest areas for further investigation, possibly with other machine learning strategies or in different contexts.

Final Thoughts

Conclude with the broader impact of this research on the field of educational data analytics.

References

Cited Works

List all scholarly sources and references used throughout the paper in an appropriate format.

Appendices

Additional Tables and Figures

Include any supplementary material that supports the analysis but is too detailed for the main text.

Questions for Discussion

1. Data Simulation Techniques

Problem: How can we improve our data simulation techniques to better capture the complexities of real-world data scenarios, including distribution shifts and interaction effects?

Path forward: Incorporate generative adversarial networks (GANs) to model complex data distributions and interactions more effectively.

Source: Goodfellow, I., et al. (2014). Generative adversarial nets. *Neural Information Processing Systems*.

2. Handling of Categorical Variables

Problem: What are the best practices for managing categorical variables, particularly in terms of encoding and handling categories that appear in the test dataset but not in the training dataset?

Path forward: Use embedding layers in neural networks or apply target encoding, which can handle unseen categories by sharing information between categories.

Source: Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.

3. Missing Data Imputation

Question: Could you recommend more advanced methods for missing data imputation that take into account the underlying distribution and relationships in the data, beyond simple techniques like `na.roughfix`?

Path forward: Implement MICE or augment this strategy?

Source: Azur, M. J., et al. (2011). Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*.

4. Model Evaluation Metrics

Problem: What additional metrics should we consider to comprehensively evaluate our models, particularly to diagnose and address issues of model bias, variance, and other performance aspects? **Path forward**: Incorporate model-agnostic metrics such as SHAP (SHapley Additive exPlanations) values to understand model decisions and potential biases in model predictions. **Source**: Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

5. Optimization and Computational Efficiency

Problem: What strategies can we employ to enhance the computational efficiency of our data processing and modeling pipelines, including code optimization and the use of parallel processing? **Path forward**: Leverage distributed computing frameworks like Apache Spark for handling large datasets and utilize GPU acceleration for model training.

Source: Zaharia, M., et al. (2010). Spark: Cluster computing with working sets. *HotCloud*.