

A COMPARATIVE STUDY OF TECHNIQUES FOR HANDLING IMBALANCED DATASET

1AUTHOR'S NAME (<Diksha>), 2CO-AUTHOR'S NAME (<Dr. Payal Gulati>)

¹Author's Student, J.C Bose University of Science and Technology, Faridabad, Haryana

²Co-Author's Mentor, J.C Bose University of Science and Technology, Faridabad, Haryana

Mostly in regression and classification problems, unbalanced data seriously challenges machine learning since under-representation of the minority class produces distorted predictions and lower model performance. This work proposes a comparison of histogram-based data-level sampling methods, namely HistUS (Histogram-Based Undersampling), HistOS (Histogram-Based Oversampling), and their hybrid combination with SMote (Synthetic Minority Over-sampling Technique) and NCL (Neighbourhood Cleaning Rule) in order to solve the imbalanced data stream issue. By use of synthetic data generation and histogram-based techniques, these methods balance the class distribution and thereby solve the shortcomings of traditional approaches such as Chebyshev's inequality. We investigate both synthetic and real-world datasets in order to evaluate HistUS, HistOS, and SMote-NCL against conventional approaches including random oversampling and undersampling. The core focus of the research is model accuracy; integrity, resilience, and robustness in dynamic data environments follow second. Our results reveal that HistUS, HistOS, and the SMote-NCL hybrid strategy perform rather better than standard resampling techniques in managing rare events and prediction accuracy. The findings provide perceptible data to improve predictive maintenance, healthcare diagnostics, and fraud detection model performance. This work aims to provide machine learning professionals with effective tools for quickly managing unbalanced data, hence producing more accurate and dependable models for pragmatic application.