Impact of Imbalance Class Distribution on Model Performance

Ishkhanuhi Hakobyan

Department of Mathematics and Mechanics
Yerevan State University, Yerevan, Armenia

Abstract: This study investigates the effects of imbalanced class distribution on machine learning model performance. The research explores the issue by using datasets of varying class imbalance levels, training a variety of models on these datasets, and assessing their initial performance. Techniques such as oversampling, undersampling, and Synthetic Minority Over-sampling Technique (SMOTE) were then applied to address the imbalance. Comparative analysis between the models' performance on balanced versus imbalanced datasets unveiled the influence of class imbalance on model accuracy. This research provides valuable insights for developing robust machine learning models in the presence of class imbalances and lays the foundation for future exploration of more effective balancing techniques.

Index Terms: Class Imbalance, Data Imbalance, Oversampling, Undersampling, Classification.

1. Introduction

In the field of machine learning, class imbalance is a pervasive and complex issue that can significantly influence the effectiveness and accuracy of predictive models. It arises when the classes in a given dataset are not represented equally, with one or more classes—termed majority classes—having a larger number of instances than others—termed minority classes. This situation is common in various domains, including fraud detection, medical diagnosis, and text classification, where the event of interest often forms the minority class.

The class imbalance problem poses a significant challenge to the performance of machine learning models. This is primarily due to the tendency of these models to skew towards the majority class during training, resulting in a bias. Consequently, the model's ability to correctly predict the minority class, which often represents the event of most interest, is compromised. In many real-world scenarios, such as fraud detection or cancer diagnosis, the cost associated with misclassification of the minority class can be high, rendering the performance of a biased model unacceptable.

Given its prevalence and potential implications, understanding and addressing class imbalance is crucial to the development of robust and reliable machine learning models. This research aims to investigate the impact of imbalanced class distribution on model performance, providing insights into the effects of class imbalance and techniques for effectively handling it.

Often, traditional performance metrics like accuracy can present a misleading picture of the model's effectiveness when dealing with imbalanced datasets. A model that simply predicts the majority class for all instances will achieve high accuracy due to the larger number of majority class instances, despite completely failing to identify the minority class. This emphasizes the need for more sophisticated metrics and evaluation techniques, such as the Area Under the Receiver Operating Characteristic (AUC-ROC), F1 score, and confusion matrix, which consider the model's performance across all classes.

While various techniques have been proposed to handle class imbalance, including oversampling the minority class, undersampling the majority class, or implementing cost-sensitive learning, there is no one-size-fits-all solution. The efficacy of these techniques can vary depending on the specific characteristics of the dataset and the chosen machine learning model.

The complexity of the class imbalance problem underscores the importance of this research. Through an extensive examination of the impact of imbalanced class distribution on different machine learning models and a comparative analysis of various class balancing techniques, this study aims to contribute valuable insights to the ongoing discourse in this field.

The rest of the paper is structured as follows: a review of the related literature, a detailed description of the methodology, an analysis and comparison of the models' performance on imbalanced and balanced datasets, presentation of the results, and a conclusion summarizing the findings and implications of the study. This research hopes to guide practitioners in making informed decisions when encountering class imbalance, ultimately leading to the development of more robust and accurate predictive models.

1.1. Literature Review

The investigation of sampling methodologies to manage class imbalances has garnered extensive attention, yielding improved classification outcomes. For instance, research conducted by [1] deployed the ADASYN oversampling approach to equilibrate classes within a hypertension dataset. The findings revealed a substantial performance enhancement in each classification model when the oversampling method was applied. Another study by [2] utilized both ADASYN and SMOTE methods to address class imbalances in a diabetes mellitus dataset, followed by classification using the Support Vector Machine (SVM) algorithm. The application of the oversampling method showed an increase in classification performance, with accuracy rates of 87.3% and 85.4% for ADASYN + SVM and SMOTE + SVM methods, respectively. This was in stark contrast to the lower 83% accuracy obtained without oversampling.

Further research combined the Synthetic Minority Oversampling Technique (SMOTE) for oversampling and Edited Nearest Neighbor (ENN) for undersampling to equilibrate Land Use and Land Cover (LULC) classifications, demonstrating that SMOTE-ENN improved the performance of Random Forest and Casboost models [3].

A comparative study by Imran [4] examined SMOTE and ROS (Random Over Sampling) methods, concluding that both methods could enhance the performance of the classification algorithm. Conversely, studies by Rashu [5] and Thammasiri [6] using Random Under Sampling (RUS), an undersampling method, reported decreased performance of the classification algorithm. However, research by Kubat [7] using One-Sided Selection (OSS), another undersampling method, found that applying the OSS method improved the performance of the classification algorithm.

Further exploration into handling class imbalances was conducted by Noorhalim [8] and Zhihao [9] using the SMOTE method. Both studies confirmed that applying class imbalance techniques to datasets can enhance the performance of several classification algorithms. Moreover, a study by Sajid Ahmed [**?**] investigated class imbalance handling in datasets using ensemble resampling. The methods tested included SMOTE-Bagging, RUS-Bagging, ADASYN-Bagging, and RYSIN-Bagging. The findings indicated that all four methods successfully improved the performance of the utilized classification algorithm.

1.2. Problem Statement and Objectives

The predominant issue to be addressed is the impact of class imbalance on the performance of machine learning models. The objectives of this research include:

- To review the existing literature on the impact of class imbalance on machine learning models.
- To discuss the various techniques that have been used to handle class imbalance and evaluate their effectiveness.
- To understand the applicability of these techniques in different contexts and with different

types of data.
- To identify areas where further research is required to enhance the handling of class imbalance in machine learning models.
- To conduct experiments using real-world datasets to assess the performance of different techniques in managing class imbalance.
- To propose recommendations for practitioners on selecting the most suitable technique, based on the specific characteristics of the dataset and the task at hand.
- To provide insights into future trends and challenges in handling class imbalance in machine learning, including the use of more advanced techniques like deep learning and transfer learning.

2. Methodology

2.1. Data Collection

The datasets utilized in this research are obtained from publicly available sources. These datasets were specifically chosen as they embody varying degrees of class imbalance, providing a robust platform for our study. The datasets include:

- **Credit Card Fraud Detection Dataset:** This dataset is highly skewed, consisting of transactions made by credit cards in September 2013 by European cardholders. The dataset presents transactions that occurred over two days, where we have 492 frauds out of 284,807 transactions. The extreme class imbalance in this dataset provides an interesting challenge for machine learning models.
- **Breast Cancer Wisconsin (Diagnostic) Dataset:** This dataset includes features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The dataset has 569 instances, 357 benign and 212 malignant. The moderate class imbalance in this dataset allows for a contrast to the more extreme Credit Card Fraud Detection dataset.

2.2. Data Analysis (Breast Cancer)

After conducting visualizations, calculating correlations, and performing feature selections, we gained valuable insights into the dataset.

2.2.1. Correlation Analysis

To assess the correlation between certain features, we utilized various techniques. For instance, the box plot analysis revealed similarities between the `concavity_worst` and `concave_point_worst` variables. To delve deeper into their relationship, we employed a joint plot. The resulting Pearson correlation coefficient (Pearsonr) of 0.86 indicated a strong positive correlation between the two variables. However, it is important to note that at this stage, feature selection has not taken place, and we are only exploring the correlations between variables.

2.2.2. Feature Selection

After examining the heatmap, several features were found to be correlated. For example, `radius_mean`, `perimeter_mean`, and `area_mean` demonstrated strong correlations. To simplify the analysis, we opted to utilize only `area_mean` as a representative feature. The selection of `area_mean` was based on personal judgment, taking into account the insights gained from swarm plots. However, it is crucial to conduct further experimentation to accurately differentiate among the correlated features.

Furthermore, the analysis revealed other correlated features. For instance, `compactness_mean`, `concavity_mean`, and `concave_points_mean` exhibited correlations, leading us to select `concavity_mean` as the representative feature. Similarly, `radius_se`, `perimeter_se`, and `area_se` demonstrated correlations, prompting the use of `area_se`. Likewise, `radius_worst`, `perimeter_worst`,

and area_worst exhibited correlations, and we chose area_worst as the representative feature.

Moreover, the features compactness_worst, concavity_worst, and concave_points_worst were found to be correlated, and we selected concavity_worst as the representative feature. Similarly, compactness_se, concavity_se, and concave_points_se demonstrated correlations, leading us to choose concavity_se as the representative feature. Additionally, texture_mean and texture_worst were correlated, and we opted to use texture_mean. Lastly, area_worst and area_mean exhibited correlations, and we selected area_mean for this analysis.

3. Modelling

3.1. Breast Cancer

To assess the performance of different classification models, we employed logistic regression, random forest, and SVM algorithms on the preprocessed dataset.

Logistic Regression: Logistic regression demonstrated good performance in classifying the breast cancer dataset. It provided reliable predictions and yielded promising results in terms of accuracy, precision, recall, and F1-score.

Random Forest: Random forest, known for its ability to handle complex relationships and avoid overfitting, exhibited excellent performance. It effectively captured the interactions among features and produced accurate predictions. The ensemble nature of random forest contributed to its robustness and stability.

SVM: The SVM algorithm, with its ability to handle both linear and non-linear relationships, delivered impressive results. It effectively learned complex decision boundaries and achieved high accuracy, precision, recall, and F1-score.

Overall, all three models—logistic regression, random forest, and SVM—performed well in classifying the breast cancer dataset. Their strong performance indicates their suitability for this task and suggests that they can provide reliable predictions for future instances. However, further evaluation and comparison are necessary to determine the optimal model for the specific requirements and constraints of the problem at hand.

3.2. Fraud Classification

To assess the performance of different classification models, logistic regression and random forest algorithms were employed on the preprocessed dataset for fraud classification.

Logistic Regression: Logistic regression demonstrated good performance in fraud classification. It provided reliable predictions and yielded promising results in terms of accuracy, precision, recall, and F1-score. The logistic regression model effectively captured the relationships between features and produced accurate predictions for identifying fraudulent transactions.

Random Forest: Random forest, known for its ability to handle complex relationships and avoid overfitting, exhibited excellent performance in fraud classification. It effectively captured the interactions among features and produced accurate predictions. The ensemble nature of the random forest contributed to its robustness and stability in detecting fraudulent activities.

Overall, both logistic regression and random forest models performed well in fraud classification. Their strong performance indicates their suitability for this task and suggests that they can provide reliable predictions for identifying fraudulent transactions. However, further evaluation and comparison are necessary to determine the optimal model for the specific requirements and constraints of the problem at hand.

4. Balancing Techniques

To address the class imbalance in the breast cancer dataset, we employed various balancing techniques. These techniques aimed to mitigate the challenges posed by the unequal distribution of classes.

Oversampling with SMOTE: We applied the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class. SMOTE generated synthetic examples by interpolating between neighboring instances in the feature space. This approach helped to balance the class distribution and increase the representation of the minority class in the training data.

Undersampling with RandomUnderSampler: We utilized the RandomUnderSampler to randomly remove instances from the majority class, thereby reducing its dominance in the dataset. This technique allowed us to create a more balanced representation of both classes in the training data.

By applying these balancing techniques, we aimed to create more equitable training datasets and improve the model's ability to learn from the minority class instances.

5. Comparison and Analysis

The classification results for both the breast cancer and fraud detection tasks are presented in the following tables. These tables provide a comprehensive overview of the performance metrics and allow for an in-depth analysis and comparison of the classification models' effectiveness.

TABLE I
COMPARISON OF CLASSIFICATION RESULTS FOR DIFFERENT SAMPLING TECHNIQUES AND MODELS FOR BREAST CANCER CLASSIFICATION

Model	Sampling Technique	Accuracy	Precision	Recall	F1-Score
Random Forest	Unbalanced	0.95	0.92	0.94	0.93
	Oversampled	0.97	0.97	0.95	0.96
	Undersampled	0.95	0.94	0.94	0.94
SVM	Unbalanced	0.89	1.0	0.70	0.82
	Oversampled	0.90	0.90	0.83	0.86
	Undersampled	0.89	0.87	0.83	0.85
Logistic Regression	Unbalanced	0.95	0.92	0.94	0.93
	Oversampled	0.95	0.90	0.97	0.93
	Undersampled	0.95	0.90	0.98	0.94

TABLE II
COMPARISON OF CLASSIFICATION RESULTS FOR DIFFERENT SAMPLING TECHNIQUES AND MODELS FOR CREDIT FRAUD CLASSIFICATION

Model	Sampling Technique	Accuracy	Precision	Recall	F1-Score
Random Forest	Unbalanced	1.0	0.83	0.59	0.69
	Oversampled	1.0	0.87	0.78	0.83
	Undersampled	0.98	0.06	0.94	0.11
Logistic Regression	Unbalanced	1.0	0.83	0.59	0.69
	Oversampled	0.98	0.07	0.86	0.13
	Undersampled	0.96	0.03	0.93	0.94

6. Results

6.1. Breast Cancer

The use of these balancing techniques had a significant impact on the model's performance. It helped to address the issue of class imbalance and improve the model's ability to accurately classify instances from both classes. The oversampling technique with SMOTE increased the representation of the minority class, allowing the model to better learn its characteristics. Similarly, the undersampling technique helped to reduce the dominance of the majority class, enabling the model to give more attention to the minority class.

Overall, both oversampling and undersampling techniques played a crucial role in enhancing the performance of the classification models. They provided a more balanced and representative training dataset, which ultimately resulted in improved accuracy, precision, recall, and F1-score for both classes.

It is worth noting that the choice of the balancing technique may depend on the specific characteristics of the dataset and the problem at hand. Therefore, further experimentation and analysis are necessary to identify the most suitable balancing technique for a given scenario.

6.2. Credit Fraud

Based on the results of the fraud detection task, as shown in Table, several models with different sampling techniques were evaluated for their performance. The metrics considered for comparison include accuracy, precision, recall, and F1-score.

In terms of accuracy, both the Random Forest and Logistic Regression models achieved high accuracy scores across all sampling techniques, with scores of 1.0 for unbalanced data in both cases. This indicates that the models correctly classified a large proportion of the transactions, regardless of the sampling technique applied.

When examining precision, which measures the proportion of correctly identified fraudulent cases out of all instances classified as fraudulent, the Random Forest model demonstrated consistently higher precision scores compared to Logistic Regression across all sampling techniques. The oversampled Random Forest model achieved a precision score of 0.87, indicating a high proportion of accurately identified fraudulent cases.

The recall metric, which measures the proportion of actual fraudulent cases correctly identified by the model, varied across the models and sampling techniques. The oversampled Random Forest model achieved the highest recall score of 0.78, indicating its ability to capture a significant portion of the fraudulent transactions. However, the undersampled Logistic Regression model achieved the highest recall score of 0.94, indicating its ability to effectively identify a large proportion of fraudulent cases, although with a lower precision.

The F1-score, which combines precision and recall into a single metric, reflects the balance between correctly identifying fraudulent cases and minimizing false positives and false negatives. The oversampled Random Forest model achieved the highest F1-score of 0.83, indicating a good balance between precision and recall.

Overall, the Random Forest model with oversampling consistently demonstrated favorable performance across multiple metrics, including accuracy, precision, recall, and F1-score. It exhibited a strong ability to identify fraudulent cases accurately while maintaining a low rate of false positives. However, it is important to consider the specific requirements and trade-offs associated with precision and recall, as they may vary depending on the context and resources available for fraud investigation.

The results suggest that oversampling techniques can effectively improve the performance of the models in handling class imbalance for fraud detection tasks. However, further experimentation and analysis may be necessary to identify the most suitable model and sampling technique based on the specific requirements and constraints of the task at hand.

References

[1] N. Chamidah, M. M. Santoni, and N. Matondang, ”Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network (ANN)”, *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 1, no. 3, pp. 635–641, 2017.

[2] N. G. Ramadhan, ”Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus”, *Scientific Journal of Informatics*, vol. 8, no. 2, pp. 276–282, 2021.

[3] H. L. Ngo, H. D. Nguyen, P. Loubiere, T. V. Tran, G. erban, M. Zelenakova, P. Brecan, and D. Laffly, ”The Composition of Time-Series Images and Using The Technique SMOTE ENN for Balancing Datasets in Land Use/Cover Mapping”, *Acta Montanistica Slovaca*, vol. 27, no. 2, pp. 342–359, 2022.

[4] M. Imran, M. Afroze, S. K. Sanampudi, A. Abdul, and M. Qyser, ”Data Mining of Imbalanced Dataset in Educational

Data Using Weka Tool", *International Journal of Engineering Science and Computing*, vol. 6, no. 6, pp. 7666–7669, 2016.

[5] R. I. Rashu, N. Haq, and R. M. Rahman, "Data Mining Approaches to Predict Final Grade by Overcoming Class Imbalance Problem", in *2014 17th International Conference on Computer and Information Technology, ICCIT 2014*, 2014, pp. 14–19.

[6] D. Thammasiri, D. Delen, P. Meesad, and N. Kasap, "A Critical Assessment of Imbalanced Class Distribution Problem: The Case of Predicting Freshmen Student Attrition", *Expert Systems with Applications*, vol. 41, no. 2, pp. 321–330, 2014.

[7] M. Kubat and S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One Sided Selection", in *International Conference on Machine Learning*, vol. 97, 1997, pp. 179–186.

[8] N. Noorhalim, A. Ali, and S. M. Shamsuddin, "Handling Imbalanced Ratio for Class Imbalance Problem Using SMOTE", in *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*, 2017, pp. 19–29.

[9] Z. Peng, F. Yan, and X. Li, "Comparison of The Different Sampling Techniques for Imbalanced Classification Problems in Machine Learning," in *Proceedings - 2019 11th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA 2019*, 2019, pp. 431–434.