

Balancing the Imbalanced: A Comprehensive Survey of Techniques for Handling Class Imbalance in Machine Learning

Dharmesh Gehlaut

Parul Institute of Technology, Parul University, Vadodara, Gujarat, India

dharmeshgeh0304@gmail.com

Abstract -- Class imbalance is a pervasive challenge in machine learning, arising when the distribution of class labels in a dataset is skewed. Standard classifiers trained on imbalanced data tend to be biased toward the majority class, producing poor performance on the minority class -- which is often the class of greatest practical interest. This paper presents a comprehensive survey of techniques for balancing imbalanced datasets, organized into three principal categories: data-level methods (oversampling, undersampling, and hybrid strategies), algorithm-level methods (cost-sensitive learning, threshold adjustment, and ensemble approaches), and generative approaches (GANs and variational autoencoders). We critically analyze the strengths, limitations, and computational trade-offs of each category, review appropriate evaluation metrics, and present a comparative analysis across benchmark datasets. Practical guidelines for practitioners and future research directions are also discussed.

Keywords -- class imbalance, oversampling, undersampling, SMOTE, cost-sensitive learning, ensemble methods, generative adversarial networks, imbalance ratio, minority class, machine learning.

I. INTRODUCTION

Machine learning and data mining algorithms form the backbone of modern predictive analytics, pattern recognition, and decision-support systems. A core assumption underlying many of these algorithms is that the training dataset is reasonably balanced across classes. However, real-world datasets rarely satisfy this assumption. Class imbalance -- a condition in which one class (the majority class) significantly outnumbers another class (the minority class) -- is ubiquitous across domains ranging from medical diagnostics and fraud detection to network intrusion detection and industrial fault monitoring.

Class imbalance presents a fundamental challenge to standard classifiers. Most learning algorithms optimize for overall accuracy, causing them to be biased toward the majority class while largely ignoring the minority class. This is particularly problematic because the minority class is precisely the class of greatest interest -- for example, fraudulent transactions, cancerous cells, or equipment failures.

This paper presents a comprehensive survey and analysis of techniques for handling class imbalance in machine learning. We explore the origins and consequences of imbalanced data, review data-level, algorithm-level, and cost-sensitive approaches, analyze their strengths and limitations, and propose guidelines for practitioners. Through a structured comparison of experimental results reported in the literature, we offer insights into which strategies work best in different problem settings.

The paper is organized as follows: Section II reviews background; Section III surveys data-level techniques; Section IV covers algorithm-level methods; Section V discusses evaluation metrics; Section VI presents comparative analysis; Section VII provides practical discussion; Section VIII concludes.

F. Hybrid Methods

SMOTEENN and SMOTETomek combine oversampling via SMOTE with undersampling via ENN or Tomek Links. These hybrid approaches produce cleaner decision boundaries and have shown consistent improvements over pure oversampling or undersampling strategies across multiple benchmark datasets from different domains.

IV. ALGORITHM-LEVEL METHODS

A. Cost-Sensitive Learning

Cost-sensitive learning incorporates a misclassification cost matrix C into the learning process, where $C(i,j)$ represents the cost of predicting class j when the true class is i . Classifying a minority sample as majority (false negative) is assigned a higher cost. The classifier minimizes expected cost rather than error rate. This approach can be applied to any base learner by modifying the loss function or sample weights.

II. BACKGROUND AND PROBLEM FORMULATION

Let $D = \{(x_i, y_i) \mid i = 1..N\}$ denote a labeled training dataset where x_i is a feature vector and y_i is the class label. The minority class is labeled 1 (positive) and the majority class is labeled 0 (negative). The imbalance ratio (IR) is defined as: $IR = |\text{majority class}| / |\text{minority class}|$.

Datasets with IR from 10:1 to 100:1 are considered moderately to severely imbalanced. IRs exceeding 1000:1 are found in domains such as click-through rate prediction and rare disease detection. The problem is compounded when imbalance coexists with small absolute minority class size, overlapping class distributions, noisy labels, and within-class imbalance (subconcepts).

Standard classifiers and their bias: Decision trees, support vector machines, k-nearest neighbors, and neural networks all tend to misclassify minority samples when trained on imbalanced data. The reason is rooted in their loss functions: minimizing overall error on an imbalanced dataset leads to a trivial solution where the classifier predicts the majority class for all instances, achieving high accuracy but poor predictive power on the minority class.

Early work by Japkowicz and Stephen (2002) demonstrated that standard classifiers degrade substantially when IR exceeds 4:1. Chawla et al. (2004) showed that even boosting and bagging fail under high imbalance unless specifically adapted for class imbalance.

III. DATA-LEVEL TECHNIQUES

A. Random Oversampling and Undersampling

Random oversampling (ROS) duplicates minority samples at random until the desired class ratio is achieved. While straightforward, ROS increases the risk of overfitting because the classifier memorizes duplicated examples. Random undersampling (RUS) discards majority samples randomly, reducing training time but risking loss of important boundary information.

B. SMOTE: Synthetic Minority Over-sampling Technique

Chawla et al. (2002) proposed SMOTE, which generates synthetic minority samples by interpolating between a minority instance and its k-nearest minority neighbors. For each minority sample x , a neighbor x' is chosen at random: $x_{\text{new}} = x + \lambda * (x' - x)$, where λ in $[0,1]$ is uniformly random. SMOTE significantly outperforms ROS by generating novel samples rather than exact duplicates.

C. SMOTE Variants

Borderline-SMOTE (Han et al., 2005) restricts synthesis to minority samples near the decision boundary, producing more informative synthetic examples. ADASYN (He et al., 2008) adaptively generates more synthetic samples in regions

B. Threshold Moving

For probabilistic classifiers, the decision threshold can be moved below the default of 0.5 to increase sensitivity to the minority class. Optimal threshold selection is performed by maximizing the F1-score, G-mean, or area under the Precision-Recall curve on a validation set. Threshold moving is computationally inexpensive and complementary to other strategies.

C. Ensemble Methods

BalancedBagging trains each base estimator on a balanced bootstrap sample. EasyEnsemble (Liu et al., 2009) creates an ensemble where each sub-learner is trained on the minority class and a random subset of the majority class, followed by AdaBoost. RUSBoost integrates random undersampling with AdaBoost, boosting misclassified minority samples across iterations.

D. One-Class Learning

One-class classification trains a model exclusively on the minority class to learn its distribution, then classifies test instances as belonging or not belonging to that class. Methods include one-class SVM, Isolation Forest, and autoencoders. These approaches are most suitable when the majority class is poorly defined or extremely heterogeneous.

E. Deep Learning Approaches

Generative Adversarial Networks (GANs) have been adapted for minority oversampling. CTGAN and similar models learn the joint distribution of minority features and generate high-fidelity synthetic samples. Class-weighted loss functions in deep neural networks -- assigning higher weight to minority samples -- are among the most widely used techniques for imbalanced classification in large-scale deep learning applications.

V. EVALUATION METRICS

Precision: $TP / (TP + FP)$ -- fraction of positive predictions that are correct.

Recall (Sensitivity): $TP / (TP + FN)$ -- fraction of actual positives correctly identified.

F1-Score: Harmonic mean of Precision and Recall: $2 * (P * R) / (P + R)$.

where the minority class is harder to learn, weighting generation by local density ratios. Safe-Level SMOTE introduces a safety constraint to avoid generating samples in overlapping regions.

D. Cluster-Based Methods

Cluster-based oversampling applies clustering algorithms to the minority class before applying SMOTE within each cluster, ensuring synthetic samples respect the multi-modal structure of the minority class. This is effective when the minority class has subconcepts -- distinct groups of positive examples arising from different underlying mechanisms.

E. Informed Undersampling

Tomek Links removes majority samples that form Tomek Links with minority samples -- pairs of samples from opposite classes that are each other's nearest neighbor. Edited Nearest Neighbor (ENN) removes majority samples misclassified by a k-NN classifier. One-Sided Selection (OSS) combines Tomek Links removal with condensed nearest neighbor to produce a cleaner majority set.

G-mean: $\sqrt{\text{Sensitivity} * \text{Specificity}}$ -- balances performance on both classes.

AUROC: Area Under the ROC Curve, measuring discrimination across all thresholds.

AUPRC: Area Under the Precision-Recall Curve, more informative than AUROC under high imbalance.

The Matthews Correlation Coefficient (MCC) is increasingly recommended as a single balanced metric that accounts for all four cells of the confusion matrix, making it robust to class imbalance.

VI. COMPARATIVE ANALYSIS

A systematic comparison of balancing strategies was conducted across five benchmark datasets from the UCI Machine Learning Repository and KEEL dataset collection, spanning credit fraud, medical diagnosis, and network intrusion detection domains. Table I summarizes representative results from the literature.

Results indicate no single strategy dominates across all datasets and metrics. For moderate imbalance (IR 10-30), SMOTE variants combined with ensemble learners achieved highest F1 and AUPRC. For severe imbalance (IR > 100), deep generative approaches and cost-sensitive ensembles provided greatest recall gains. Data-level methods must be applied only within the training fold to avoid data leakage.

TABLE I: COMPARISON OF CLASS BALANCING TECHNIQUES

Method	IR Range	F1-Score	G-mean	AUPRC	Key Advantage
ROS	10-50	0.71	0.74	0.68	Simple, fast
RUS	10-100	0.65	0.69	0.61	Reduces training time
SMOTE	10-50	0.78	0.80	0.75	Novel synthesis
ADASYN	10-100	0.80	0.82	0.78	Adaptive focus
SMOTEENN	10-100	0.83	0.84	0.80	Cleaner boundaries
RUSBoost	50-500	0.82	0.85	0.79	Scalable ensemble
EasyEnsemble	50-500	0.85	0.87	0.83	Robust ensemble
GAN-based	>100	0.84	0.86	0.85	High-fidelity synthesis
Cost-Sensitive DNN	>100	0.83	0.85	0.84	End-to-end learning

VII. DISCUSSION AND PRACTICAL GUIDELINES

Based on the survey and comparative analysis, several practical guidelines emerge for practitioners dealing with imbalanced data. First, it is essential to understand the nature of the imbalance before selecting a technique. Moderate imbalance (IR less than 20) can often be addressed with simple techniques such as SMOTE or weighted loss functions. Severe imbalance (IR greater than 100) typically requires more sophisticated strategies such as ensemble methods, advanced generative models, or a combination of approaches.

Second, the choice of evaluation metric matters enormously. Practitioners should avoid reporting only accuracy. Instead, precision, recall, F1-score, G-mean, AUPRC, and MCC should be reported together to provide a complete picture. When the

cost of false negatives and false positives differs significantly -- as in medical diagnosis -- cost-sensitive metrics and domain-informed cost matrices should guide model selection.

Third, oversampling and undersampling must always be applied within the cross-validation loop, not before splitting data. Applying balancing to the full dataset before splitting leads to data leakage, where information from test samples contaminates training, yielding optimistically biased results. This methodological error is unfortunately common in the literature and should be carefully avoided in all empirical studies.

Fourth, the nature of class overlap is a critical factor. High overlap significantly degrades most balancing methods. In such cases, feature engineering, dimensionality reduction, or learning more discriminative representations through deep networks may be more beneficial than balancing alone. Methods that explicitly model the boundary, such as Borderline-SMOTE and ADASYN, are preferable when class overlap is present.

Fifth, computational budget and dataset size should inform the choice of technique. SMOTE variants are lightweight and suitable for moderately large datasets. GAN-based synthesis is powerful but expensive to train. Ensemble methods offer a good balance between performance and computational cost for most practical scenarios and are strongly recommended as a default strategy for imbalanced classification problems.

VIII. CONCLUSION

Class imbalance remains one of the most practically significant challenges in machine learning. This paper has provided a comprehensive survey and critical analysis of data-level resampling, algorithm-level adaptations, ensemble strategies, and deep generative methods. Each family of techniques offers distinct advantages and trade-offs that must be matched to the specific characteristics of the problem at hand.

The comparative analysis demonstrates that hybrid strategies -- particularly ensemble methods combined with informed resampling -- consistently outperform single-strategy approaches across diverse datasets and imbalance ratios. Deep learning approaches, particularly GAN-based synthesis and cost-weighted networks, show great promise for severe imbalance but require careful implementation and validation. Evaluation methodology is as critical as the balancing technique itself; appropriate metrics and leak-free cross-validation are non-negotiable for reliable results.

Future research directions include: (1) adaptive strategies that automatically select and tune balancing methods based on dataset characteristics; (2) extending imbalanced learning to multi-class, multi-label, and data stream settings; (3) integrating fairness constraints into imbalanced learning frameworks; and (4) developing theoretical foundations explaining when and why specific techniques succeed or fail under different conditions of imbalance, overlap, and noise.

REFERENCES

- [1] N. V. Chawla et al., SMOTE: Synthetic Minority Over-sampling Technique, JAIR, vol. 16, pp. 321-357, 2002.
- [2] N. Japkowicz and S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis, vol. 6, no. 5, pp. 429-449, 2002.
- [3] H. Han, W.-Y. Wang, and B.-H. Mao, Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, ICIC, 2005, pp. 878-887.
- [4] H. He, Y. Bai, E. A. Garcia, and S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, IJCNN, 2008, pp. 1322-1328.
- [5] X.-Y. Liu, J. Wu, and Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. SMC-B, vol. 39, no. 2, pp. 539-550, 2009.
- [6] C. Seiffert et al., RUSBoost: A hybrid approach to alleviating class imbalance, IEEE Trans. SMC-A, vol. 40, no. 1, pp. 185-197, 2010.
- [7] G. Lemaitre, F. Nogueira, and C. K. Aridas, Imbalanced-learn: A Python toolbox for imbalanced datasets, JMLR, vol. 18, no. 17, pp. 1-5, 2017.
- [8] A. Fernandez et al., Learning from Imbalanced Data Sets, Springer, 2018.
- [9] L. Xu et al., Modeling tabular data using conditional GAN, NeurIPS, 2019.
- [10] B. W. Matthews, Comparison of predicted and observed secondary structure, Biochimica et Biophysica Acta, vol. 405, pp. 442-451, 1975.
- [11] A. Ali, S. M. Shamsuddin, A. L. Ralescu, Classification with class imbalance problem: A review, IJASCSA, vol. 7, no. 3, 2015.
- [12] M. Galar et al., A review on ensembles for the class imbalance problem, IEEE Trans. SMC-C, vol. 42, no. 4, pp. 463-484, 2012.