Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 001-012
https://doi.org//10.47514/kjcs/2025.2.1.001

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Integrated Approaches for Handling Imbalanced Data: Techniques, Applications and Future Directions

Muhammad Shamsu Usman[1], Aminuddeen Abubakar[2], Abubakar Yushau[3] and Usman Abdullahi Musa[4]
[1, 2,4] Department of Computer Science, Federal University Dutse, [3]Department of Computer Science, Jigawa State Polytechnic Dutse.
muhammad.su@fud.edu.ng[1], aminuddeen@fud.edu.ng[2], sadiqyushau@gmail.com[3] and usman.m@fud.edu.ng[4]

Abstract

Class imbalance poses a significant challenge in machine learning, particularly in critical fields such as healthcare, fraud detection, and natural language processing. The underrepresentation of minority classes often results in models that fail to detect rare but crucial events, such as identifying rare diseases or fraudulent transactions. This paper aims to review advancements in integrated techniques for addressing class imbalance, focusing on their applications and challenges. A comprehensive review of data-level methods, including SMOTE (Synthetic Minority Oversampling Technique), algorithm-level adjustments like cost-sensitive learning, and hybrid techniques such as SMOTEBoost and RUSBoost was conducted. These approaches were analyzed for their effectiveness in improving model performance on minority class detection. Findings indicate that integrated techniques enhance the detection of minority instances while maintaining overall model accuracy. However, computational demands and fairness issues remain key challenges, necessitating further research into interpretable and efficient solutions.

Keywords: Imbalanced Data, Machine Learning, Oversampling, Cost-Sensitive Learning, Hybrid Approaches

1. Introduction

Imbalanced data, characterized by disproportionate class distribution, pose significant challenges in areas such as medical diagnosis, credit scoring, and fraud detection, where accurate classification of minority classes is essential (Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G., 2017). In such datasets, traditional machine learning algorithms tend to favor the majority class, resulting in poor performance in the minority class (He, H., & Garcia, E.A. , 2009). This imbalance undermines the reliability of the model as it often leads to a large number of false negatives, especially in high-risk domains such as healthcare and finance (A. Fernández, G. Salvador, G. Mikel, P. Ronaldo, K. Bartosz & H. Francisco. , 2018). Recent research has highlighted the potential of integrated approaches, which combine different methods to effectively handle imbalanced data. This review aims to analyze these methods in depth, assessing how to sample combination, cost-sensitive techniques and combination techniques improve classification results. Additionally, it explores practical applications, recent advances, and future directions in imbalanced data management.

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 001-012
https://doi.org//10.47514/kjcs/2025.2.1.001
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

2. Related Works

2.1 Definition and Implications of Imbalance

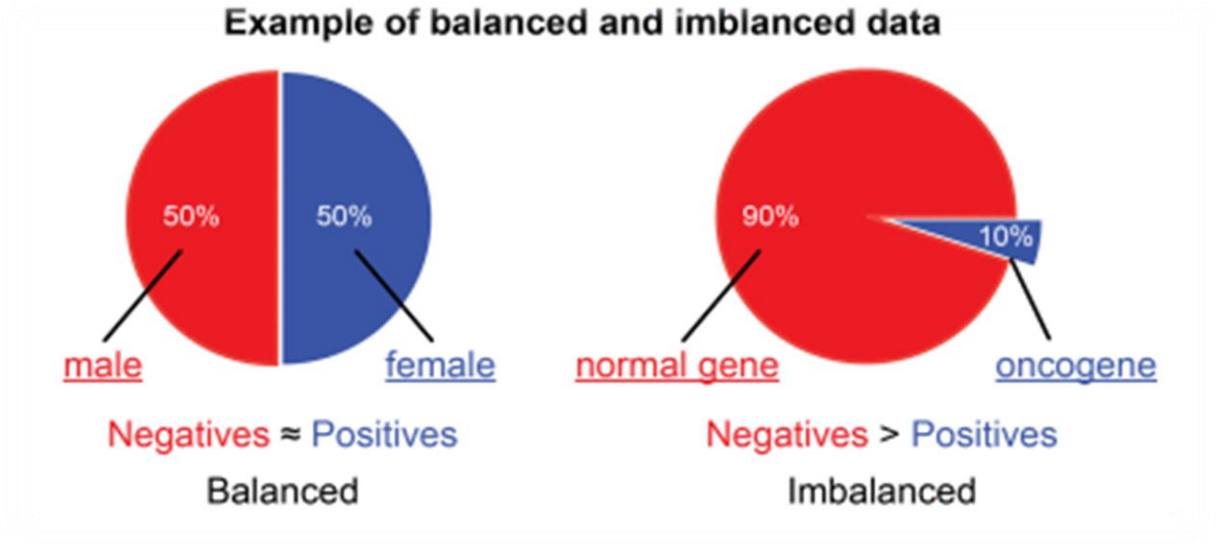

Figure 1: Balanced and Imbalance (Tripathi, 2019)

Imbalanced data refers to data sets in which one class (majority) has significantly more cases than in the other class (minority), such as in medical diagnostic data sets, where positive cases (e.g, rare diseases) are underrepresented (J. M. Johnson & T. M. Khoshgoftaar, 2019). This imbalance often results in:

1. **Biased learning process:** Models trained on imbalanced data tend to achieve high overall accuracy by focusing on the majority class, often to the detriment of the performance of the minority class (Krawczyk, 2016).
2. **High false negative rate:** Minority classes tend to have higher misclassification rates, which can be particularly detrimental in high-risk applications such as fraud detection and serious disease diagnosis (Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G., 2017).

2.2 Evaluation metrics for imbalanced data

Traditional metrics such as accuracy are inadequate for evaluating model performance in imbalanced datasets because they ignore the performance of minority classes. Therefore, alternative measures such as F1 score, Matthews's correlation coefficient (MCC), G-mean, and ROC-AUC are needed for more reliable evaluation. Recent studies highlight the effectiveness of MCC, especially for severely imbalanced cases, as it provides a balanced measure that combines true positives, false positives, true negatives, and false negatives (Chicco, D., & Jurman, G., 2020).

2.3 Data Level Technique

Data level technique aims to modify a dataset to reduce class imbalance by generating synthetic samples, removing data, or using a combination of both.

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 001-012
https://doi.org//10.47514/kjcs/2025.2.1.001
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

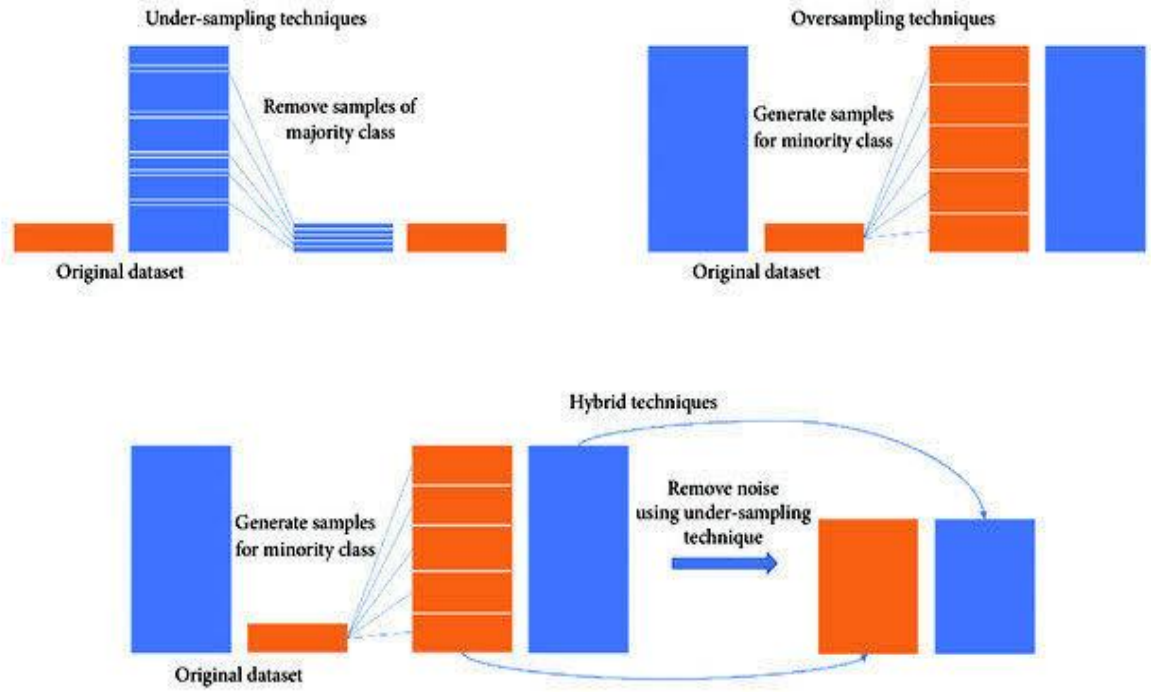

Figure 2: Data Level techniques (C. L. Liu & Y. H. Chang, 2022)

3. Methodology

Oversampling adds synthetic minority cases to the dataset. Synthetic minority sampling technique (SMOTE), one of the most widely used methods, generates synthetic data points by interpolating between existing minority samples (N. V. Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer, 2002). However, SMOTE can lead to overfitting, especially with small datasets, as the synthetic versions can unintentionally reinforce existing noise (Blagus R., & Lusa L., 2013). To address this issue, variants of SMOTE have been developed:

1. **Borderline-SMOTE:** This technique focuses on minority cases near the decision boundary to generate synthetic samples, improving the classification of particularly difficult-to-classify cases (Han, Hui & Wang, Wen-Yuan & Mao, Bing-Huan. , 2005).
2. **ADASYN (*Adaptive Synthetic Sampling*):** ADASYN generates synthetic samples based on the complexity of the data, focusing more on difficult-to-classify cases (He, H., & Garcia, E.A. , 2009).

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 001-012
https://doi.org//10.47514/kjcs/2025.2.1.001
Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

3.1 Undersampling

Undersampling reduces the number of instances of the majority class, effectively balancing the dataset by retaining only representative instances. Random sum in complex datasets. More sophisticated undersampling techniques, such as NearMiss sampling (RUS) is a popular approach, but it risks discarding valuable information, especially and cluster-based subsampling, selectively discard instances that add little to the accuracy of classification (M Beckmann, N. F. F. Ebecken & B. S. L. Pires de Lima, 2015).

3.2 Hybrid sampling

The hybrid approach combines oversampling and undersampling techniques to minimize the risk of overfitting and loss of information. Methods such as SMOTEENN and SMOTETomek have become popular, where SMOTEENN uses Edited Nearest Neighbors to clean synthetic samples and SMOTETomek uses Tomek alignment to reduce noise (Ramentol, E., Caballero, Y., Bello, R. et al., 2012). In medical imaging and predictive maintenance, hybrid sampling is especially advantageous as it provides high reliability.

4 Results and Discussion

Algorithm-level techniques modify the learning algorithm to improve performance on imbalanced data, rather than modifying the dataset itself.

4.1 Cost-sensitive learning

Cost-sensitive learning imposes a higher penalty on errors in the minority class, allowing models to prioritize the performance of the minority class. This approach is particularly effective in ensemble methods such as random forests, where cost sensitivity is built into each decision tree, balancing the trade-off between bias and variance (Xipeng Pan et al, 2023).

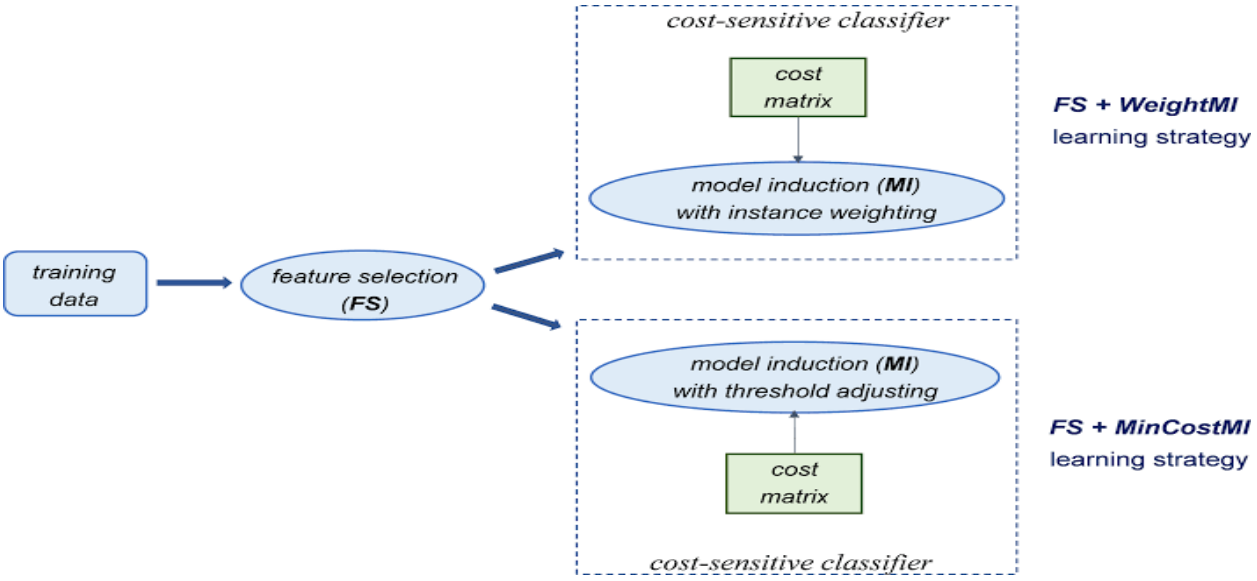

Figure 3: Cost-Sensitive Learning (B. Pes & G. Lai, 2021)

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

However, determining the appropriate cost can be difficult and often requires experimentation to achieve optimal results. Recent research on cost-sensitive deep learning models has shown promise, particularly in healthcare and finance, where imbalanced datasets are common. For example, cost-sensitive neural networks have been shown to improve the memorability of minority classes without excessively false positives (Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G., 2017).

4.2 Ensemble Learning

Ensemble methods combine predictions from multiple models, leveraging diverse perspectives to improve reliability and performance. Techniques such as AdaBoost and Balanced Random Forests modify the sample distribution to ensure balanced training, with equal emphasis on majority and minority classes (C. Chen, A. Liaw & L. Breiman, 2004). Boosting methods, such as RUSBoost and AdaCost, incorporate cost-sensitive adjustments to focus more on minority cases, resulting in better recall of the minority class. Ensemble methods are widely used in fraud detection, where accurate classification of rare cases is required, as well as in high-dimensional datasets, where balanced clustering has been shown to be beneficial (Z. Zhou and X. Liu, 2006).

4.3 One-class learning

One-class learning treats the minority class as the target class, completely ignoring the majority class. This approach, widely used in anomaly and fraud detection, focuses only on identifying deviations from the "normal" (Khan, S.S., & Madden, M.G., 2013). Popular methods include One-class support vector machines (OC-SVMs) and autoencoders, which are particularly effective when the minority data represent rare events or anomalies (Bernhard Schölkopf et al., 2001).

4.4 Hybrid Techniques

Hybrid Techniques integrates both data-level and algorithm-level approaches, allowing for an improved balance between the sampling and model training stages. These approaches are increasingly popular in complex and high-dimensional datasets.

4.5 Cost-sensitive learning and sampling techniques

Cost-sensitive learning combined with sampling techniques is an effective strategy to improve model reliability on imbalanced datasets. For example, SMOTE combined with cost-aware neural networks has achieved remarkable success in medical diagnosis, as it improves minority precision and recall without overfitting (Sun, Y., Wong, A. K., & Kamel, M. S., 2007). Similarly, cost-sensitive methods associated with ADASYN have improved fraud detection systems by reducing false negatives (Tran Khanh Dang et al, 2021).

4.6 Sampling and Ensemble Techniques

Ensemble techniques combined with sampling methods provide a robust solution to severe imbalances. SMOTEBoost, a variant that integrates SMOTE with boosting, has been successfully applied in many areas including finance and text classifications, where model sensitivity to underclasses is paramount (Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W, 2003). Balanced random forests, which use balanced

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

bootstrapping, have also gained importance in imbalanced data applications, demonstrating high stability in large data sets (C. Chen, A. Liaw & L. Breiman, 2004).

4.7 Deep Learning-based Hybrid Methods

Deep learning models, especially convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks have shown promise in handling imbalanced data when incorporated into ensemble methods. Generative adversarial networks (GANs) have recently been used to generate realistic synthetic samples for minority classes, which helps to address imbalance without overfitting (A. Antoniou, A. Storkey, & H. Edwards, 2017) (Buda, M., Maki, A., & Mazurowski, M. A, 2018). Hybrid deep learning methods have improved performance in areas such as medical imaging and sentiment analysis, where class distributions are naturally imbalanced (Jabbar, M. A., Arulmozhivarman, P., & Liyanage, C., 2020).

4.8 Applications of Integrated Imbalance Handling Techniques

4.8.1 Healthcare and Medical Diagnostics

Healthcare data is often severely class-imbalanced due to the underrepresentation of rare diseases. Ensemble techniques such as SMOTE combined with cost-sensitive neural networks have improved the accuracy of early disease detection, demonstrating the impact of ensemble methods on clinical decision making (López, V., Fernández, A., García, S., Palade, V., & Herrera, F., 2013). Similarly, deep learning models augmented with GANs have shown promise in medical imaging, improving the detection rate of rare diseases (Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G., 2017).

4.8.2 Financial Fraud Detection

Financial fraud detection is an excellent example of a high-risk application where managing imbalanced data is essential. Fraudulent cases are rare compared to non-fraudulent transactions, but misclassification can lead to significant financial losses. Cost-sensitive ensemble methods, such as Balanced Random Forests and RUSBoost, have been widely applied, focusing on reducing false negatives by imposing higher penalties on misclassified fraud (Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G., 2015). Additionally, hybrid methods combining SMOTE with cost-sensitive neural networks have shown significant improvements in recall rates, thereby reducing fraud cases missed by the model (Sun, Y., Wong, A. K., & Kamel, M. S., 2007). More recently, GAN-based oversampling has been incorporated into fraud detection systems, generating realistic synthetic fraud examples that improve model performance in identifying new fraud patterns (Douzas, G., & Bacao, F., 2018). By simulating these rare events, GANs help models better generalize to real-world situations, especially when historical data is limited.

4.8.3 Natural Language Processing (NLP) and Text Classification

Imbalanced datasets are also common in NLP applications, such as spam detection, sentiment analysis, and topic classification. In these cases, oversampling techniques such as SMOTE, as well as combined deep learning models, have been shown to be particularly effective (Wang, S., & Manning, C. D., 2012). For example, in sentiment analysis, rare instances of negative sentiment are often underrepresented, affecting the model's performance in detecting negative opinions (Buda M, Maki A & Mazurowski M. A., 2018). Deep learning methods, including LSTMs combined with GAN-generated text samples, have improved text classification accuracy for minority classes by generating synthetic minority samples. These techniques

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

help the model learn the semantic features of underrepresented classes, achieving balanced performance across categories (Kumar, V., Basu, M., & Pan, Y., 2019). Hybrid techniques that integrate cost-aware learning with attention mechanisms also show promising results, especially in the case of context-dependent minority sentences.

4.8.4 Industrial Applications: Predictive Maintenance

Predictive maintenance applications often deal with imbalanced data sets, where failure events are rare compared to normal operational data. In such cases, misclassification can lead to serious errors, highlighting the importance of detecting outliers (Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z., 2014). Ensemble methods combined with cost-sensitive learning have been shown to be effective, with models such as Balanced Random Forests achieving higher sensitivity to failures (López, V., Fernández, A., García, S., Palade, V., & Herrera, F., 2013). Recent studies have leveraged hybrid deep learning methods, such as LSTM with SMOTE, to augment data over time. This integration improves the ability to predict equipment failures by focusing on sequences that represent warning signs (Jabbar, M. A., Arulmozhivarman, P., & Liyanage, C., 2020). By prioritizing minority class sequences, these methods help prevent unplanned downtime, which is critical for industrial operations.

4.9 Recent Advances and limitations

4.9.1 Advances in hybrid techniques

In recent years, hybrid methods that combine deep learning with traditional machine learning methods have emerged, showing promise in handling very large imbalanced datasets in a number of domains. Techniques that integrate GANs with CNNs or LSTMs have become popular in the fields of images and sequential data, where they generate synthetic samples that improve minority class detection without significant overfitting (A. Antoniou, A. Storkey, & H. Edwards, 2017). Studies on medical and financial datasets demonstrate that these combined methods systematically improve model robustness, reduce false negatives, and improve recall (Buda M, Maki A & Mazurowski M. A., 2018).

4.9.2 Limitations of Ensemble Approaches

Despite these advances, ensemble approaches still face limitations. Ensemble approaches often require complex tuning and are computationally intensive, especially on large-scale datasets. Techniques such as SMOTE-based oversampling and GANs can lead to overfitting in some cases, especially when the synthesized samples are too similar to existing data points, affecting generalization (Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G., 2017). In addition, cost-sensitive methods rely on precisely defined cost matrices, which are difficult to optimize and often require domain-specific knowledge. This sensitivity to cost parameterization can hinder their application in many different domains (Xipeng Pan et al, 2023). Additionally, deep learning-based ensemble models require large datasets and computational resources, making them impractical for small or resource-constrained environments.

4.9.3 Interpretability Challenges

Another important challenge is interpretability. While ensemble and deep learning models offer performance improvements, they often operate as “black boxes,” limiting information about the model’s

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

decision-making process. This lack of transparency can be problematic in regulated industries such as healthcare and finance, where interpretability is essential (Chicco, D., & Jurman, G., 2020).

5 Future Directions

5.1 Model Interpretability and Explainability

Future research should prioritize model interpretability, especially in hybrid and deep learning approaches. Methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive Explanations) have been developed to increase transparency, providing insights into model predictions for minority classes (Ribeiro, M. T., Singh, S., & Guestrin, C, 2016). Enhanced interpretability would enable more effective use of integrated imbalance handling techniques, especially in sensitive applications like healthcare and finance.

5.2 Transfer Learning and Domain Adaptation

Transfer learning is emerging as a potential solution for handling imbalance, especially in domains with limited data. By transferring knowledge from a source domain with abundant data to a target domain with imbalanced data, models can learn to generalize better (Pan, S. J., & Yang, Q., 2010). Transfer learning has shown promise in NLP and image processing, where models trained on balanced datasets can be fine-tuned for minority-focused tasks (Kumar, V., Basu, M., & Pan, Y., 2019).

5.3 Adaptive and Automated Techniques

With advancements in AutoML (Automated Machine Learning), developing adaptive frameworks that automatically adjust to imbalanced data conditions could significantly reduce the burden of manual tuning (He, H., Zhao, Z., & Chen, W., 2019). Automated techniques could streamline hyperparameter optimization for cost-sensitive learning, ensemble selection, and hybrid configuration, making these methods accessible for broader use.

5.4 Integration with Edge Computing

For applications requiring real-time processing, such as predictive maintenance and fraud detection, integrating imbalanced data handling with edge computing could provide more efficient solutions. Edge computing frameworks can leverage lightweight, hybrid techniques to process imbalanced data close to the source, enhancing speed and reducing data transfer costs (Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L., 2016).

5.5 Ethical and Fairness Considerations

Finally, ethical considerations are crucial when applying imbalance handling techniques in sensitive domains. Future research should address how integrated approaches can contribute to fair and unbiased decision-making, particularly in applications like criminal justice and hiring, where class imbalance might unintentionally reinforce social biases (Mehrabi, N., Morstatter, F., Saxena,

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

N., Lerman, K., & Galstyan, A., 2021). Developing fairness-aware algorithms that handle both class imbalance and ethical considerations will be essential for responsible AI development.

5 Conclusion

This review explores integrated approaches to handling imbalanced data in machine learning, analyzing the strengths and limitations of combining data-level, algorithm-level, and hybrid techniques. Integrated approaches such as Hybrid oversampling with cost-sensitive ensembles, GAN-based deep learning models, and SMOTE-Booster ensembles have shown promise in important domains including healthcare, finance, natural language processing, and industrial maintenance. Although effective, these approaches pose challenges that Includes interpretability, computational complexity, and reliance on domain-specific tuning. To address these issues, future research should focus on interpretability, ethical considerations, and the development of adaptive and automated frameworks that can generalize across domains. As machine learning continues to expand into important areas with imbalanced datasets, the development of robust and accessible imbalance management techniques will be essential. By addressing these challenges, integrated approaches will better serve high-stakes domains, supporting more accurate and fair decision-making processes.

5.1 Analysis and Discussion

Managing imbalanced data is essential in domains such as healthcare, finance, and predictive maintenance, where misclassification can have serious consequences. Integrated approaches, combining data-level methods (e.g. SMOTE), algorithm-level methods (e.g. cost-aware learning), and ensemble methods (e.g. GAN-based ensemble generation) have improved minority class detection by addressing data distribution and algorithm bias. These techniques provide notable performance gains but have specific challenges and limitations.

5.2 Performance gains and challenges

Integrated techniques, such as cost-sensitive ensembles combined with oversampling, have been shown to be effective in high-stakes applications by improving minority class recall without sacrificing majority class accuracy (Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G., 2015). However, methods such as GANs, while effective in generating realistic minority samples, require high computational resources, making them impractical for real-time or resource-constrained applications (A. Antoniou, A. Storkey, & H. Edwards, 2017). Additionally, sampling methods can sometimes overfit the minority class if not carefully managed.

5.3 Interpretability and Ethical Issues

In sensitive domains such as healthcare and finance, interpretability remains important. Ensemble models, especially those based on deep learning; often act as “black boxes,” making it difficult to verify their decisions (Chicco, D., & Jurman, G., 2020). Tools such as SHAP and LIME have

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

helped improve transparency but are not always sufficient for regulated sectors (Ribeiro, M. T., Singh, S., & Guestrin, C, 2016). Additionally, bias issues arise because models can unintentionally reinforce existing imbalances in data sets, affecting fairness, especially in areas such as recruitment or criminal justice (Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A., 2021).

5.4 Emerging Approaches

Recent trends such as transfer learning and domain adaptation offer promising solutions to imbalance by leveraging knowledge from balanced domains to handle rare classes in new datasets (Pan, S. J., & Yang, Q., 2010). AutoML and adaptive algorithms can also reduce the tuning burden associated with hybrid models, making them more accessible for diverse applications (He, H., & Garcia, E.A. , 2009). For latency-sensitive applications, integrating imbalance management with edge computing can provide more efficient real-time solutions (Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L., 2016).

5. 5 Practical recommendations

Practical use of integration techniques requires a balance between performance, interpretability, resource constraints and ethical considerations:

1. **Real-time requirements:** For fast-paced applications, lightweight, adaptable methods such as cost-sensitive packages may be more appropriate.
2. **Regulated areas:** In areas where transparency is required, integration methods using interpretable tools (e.g. SHAP) should be preferred.
3. **Limited Data Case:** For domains with sparse data, simpler transfer or ensemble learning models that integrate sampling with cost sensitivity can be effective without high computational requirements.

Class imbalance, common in domains such as healthcare, fraud detection, and natural language processing, poses significant challenges for machine learning because models are often biased toward the majority class, resulting in poor performance on minority classes, which are often the most important classes to detect, such as rare diseases or fraudulent transactions. To address this problem, ensemble methods are used, combining data-level, algorithm-level, and ensemble techniques to improve prediction accuracy on imbalanced datasets. Data-feedback methods, such as SMOTE (Synthetic Minority Oversampling Technique), rebalance the dataset by generating synthetic samples for minority classes, helping the model learn minority patterns without duplicating the data. Algorithmic-level methods, such as cost-aware learning, adjust misclassification penalties, favoring minority classes during training. Ensemble methods, such as SMOTEBoost and RUSBoost, combine resampling with ensemble methods, balancing the accuracy of minority and majority classes. This review highlights the advantages and limitations of these methods, addresses their applications, and addresses the need for explainable, efficient, and fair models. Future research aims to improve these methods for better performance and ethical use in high-risk applications.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

References

A. Antoniou, A. Storkey, & H. Edwards. (2017). *Data augmentation generative adversarial networks.* arXiv preprint arXiv:1711.04340.

A. Fernández, G. Salvador, G. Mikel, P. Ronaldo, K. Bartosz & H. Francisco (2018). *Learning from Imbalanced Data Sets.*10.1007/978-3-319-98074-4.

B. Pes & G. Lai. (2021). Learning Strategies for High-Dimensional and Imbalance Data: A comparative Study. *PeerJ Computer Science*.

Bernhard Schölkopf et al. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation VOL 13 (7)*, 1443-1471.

Blagus R., & Lusa L. (2013). SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics, 14*, 106 - 106.

Buda M, Maki A & Mazurowski M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw. doi: 10.1016/j.neunet.2018.07.011.*

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249-259.

C. Chen, A. Liaw & L. Breiman. (2004). *Using Random Forest to Learn Imbalanced Data.*

C. L. Liu & Y. H. Chang. (2022). Learning from Imbalanced Data with Deep Density Hybrid Sampling. *IEEE Transactions on System, Man, and Cybernetics: Systems, Vol. 52, No. 11*, 7065-7077.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. . *European Conference on Principles of Data Mining and Knowledge Discovery.* Berlin, Heidelberg: Springer.

Chicco, D., & Jurman, G. (2020). *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.* BMC Genomics.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2015). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE transactions on neural networks and learning systems, 29(8)*, 3784-3797.

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications, 91*, 464-471.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220-239.

Han, Hui & Wang, Wen-Yuan & Mao, Bing-Huan. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *Adv Intell Comput. 3644.*, 878-887.

He, H., & Garcia, E.A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering, 21*, 1263-1284.

He, H., Zhao, Z., & Chen, W. (2019). AutoML: A survey of the state-of-the-art. *Journal of Computer Science and Technology, 34(6)*, 1138-1165.

J. M. Johnson & T. M. Khoshgoftaar. (2019). Survey on deep learning with class imbalance. *J Big Data 6, 27. https://doi.org/10.1186/s40537-019-0192-5*.

Jabbar, M. A., Arulmozhivarman, P., & Liyanage, C. (2020). Predictive maintenance model using LSTM for highly imbalanced datasets. *IEEE Access, 8*, 134246-134257.

Khan, S.S., & Madden, M.G. (2013). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review, 29*, 345 - 374.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Prog Artif Intell 5*, 221–232.

Kumar, V., Basu, M., & Pan, Y. (2019). Deep transfer learning for imbalanced data in natural language processing. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, (pp. 5125-5132).

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141.

M Beckmann, N. F. F. Ebecken & B. S. L. Pires de Lima. (2015). A KNN Undersampling Approach for Data Balancing. *Journal of Intelligent Learning Systems and Applications, Vol.7 No.4*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR), 54(6)*, 1-35.

N. V. Chawla, K. W. Bowyer, L. O. Hall & W. P. Kegelmeyer. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research, Volume 16*, 321-357.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22(10)*, 1345-1359.

Ramentol, E., Caballero, Y., Bello, R. et al. (2012). SMOTE-RSB *: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowl Inf Syst 33*, 245–265.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 1135-1144).

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal, 3(5)*, 637-646.

Sipos, R., Fradkin, D., Moerchen, F., & Wang, Z. (2014). Log-based predictive maintenance. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1867-1876.

Sun, Y., Wong, A. K., & Kamel, M. S. (2007). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23(04)*, 687-719.

Tran Khanh Dang et al. (2021). Machine Learning Based on Resampling Approaches and Deep Reinforcement Learning for Credit Card Fraud Detection Systems. *Appl. Sci. 2021, 11(21)*.

Tripathi, H. (2019). *What Is Balanced and Imbalanced Dataset?* Medium.

Wang, S., & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 90-94.

Xipeng Pan et al. (2023). Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations. *https://doi.org/10.1016/j.media.2023.102867.*

Z. Zhou and X. Liu. (2006). *OnMulti-Class Cost-Sensitive Learning.* American Association for Artificial Intelligence.