Achieving Equilibrium: A Comprehensive Survey on Addressing Class Imbalance in Machine Learning

Vandana Jagtap

Shri Venkateshwara University Gajraula, Amroha, Uttar Pradesh, India
Dr. Vishwanath Karad MIT World Peace University, Pune, India

Rakesh K Yadav
Shri Venkateswara University, Gajraula, Amroha, Uttar Pradesh India

***Abstract*—The present research explores the significant problem of class imbalance in machine learning data sets and provides an in-depth evaluation of different data balancing approaches to solve this difficulty. This is the common problem that hurts the performance of ML models, as biased predictions and lower accuracy lead to. In this study, we examine two different types of imbalanced data sets, including the classification of pneumonia images and the calculation of fraud detection data. Our research begins with a discussion on how to deal with imbalanced numerical data using binary classification methods, including gradient boost machines (GBM), random forests, decision trees, support vector machines (SVM), and log regression. For handling non-balanced numerical data sets, we use strategies such as Random Over Sampling, Random Under Sampling, SMOTE (Synthetic Minority Over- Sampling Technique), and ADASYN (Adaptive Synthetic Sampling). In the second part of the study, we focused on class imbalances in image datasets using transfer learning combined with SMOTE and ADASYN. Classification algorithms such as DenseNet, ResNet, Inception V3, VGG-16, and VGG-19 were used to distinguish pneumonia from normal cases. We evaluate the performance of these models and talk about their effectiveness when combined with data balance strategies. Our research provides insights into the effectiveness of various algorithms and methods to correct class imbalances in machine learning tasks, based on detailed experiments and analyses.**

Keywords— Machine Learning, Transfer Learning, Data Imbalanced, Random Over Sampling, Random Under Sampling, SMOTE, and ADASYN

I. Introduction

A. Selecting a Template (Heading 2)

It is essential to comprehend the diverse characteristics of data to effectively leverage its potential in a multitude of applications. The comprehension comprises categorization according to techniques for handling, retaining, and obtaining entry, leading to a difference between two primary types of data: streaming data and static data. Stream data is known by its real-time processing and continuous flow, and is commonly observed in social media feeds or sensor measurements. Static data, on the other hand, is typically stored in files or databases andis kept constant in size. These Variations affect storage systems, processing techniques, and analysis methodologies, which in turn affect data management strategies [1]. A dataset with balanced data ensures that no class is dominant by presenting each class or group equitably and fairly. Achieving balanced data is crucial for statistical analysis and machine learning in particular, as unbalanced data sets can skew models and hinder decision- making processes. By ensuring that each class is fairly represented, data balance lowers bias, enhances model performance, and increases the accuracy of predictions made across all classes [2, 3]. Supervised learning relies on the fact that each instance of labeled data has a unique class label. Approaches like logistic regression, decision trees, and neural networks use labeled data to reduce errors, generalize trends, and produce accurate predictions in tasks like sentiment analysis and photo categorization. Because unlabeled data lacks accurate grouping, supervised learning is more difficult. Unsupervised learning algorithms, like K-means, discover patterns in unlabeled data. Methods such as self-training enhance model performance by utilizing both labeled and unlabeled data used in Semi-supervised learning. Self-supervised learning predicts data segments based on past input segments. Using supervised techniques for the labeled portion is necessary when combining labeled and unlabelled occurrences in partially labeled data. Semi-supervised learning further enhances performance when both types of data are used. Transfer learning applies labeled data to unlabelled data to transfer previously taught knowledge and improve learning [6]. To put it briefly, understanding data diversity and supervisory techniques is essential for efficiently using data in statistical analysis and machine learning. Every type of data, including balanced, imbalanced, labeled, unlabelled, static, and dynamic data, has unique opportunities and difficulties that have influenced the development of data processing, analytical, and storage technologies [4][5][6][8].

II. Related work

Static Imbalanced labeled data requires certain methodologies to be used in Machine Learning(ML), Neural Network (NN), and Deep Learning (DL) algorithms. For every type of learning algorithm, there are several approaches available to address the unbalanced nature of data, particularly static; these approaches are described in this paper.

Under-sampling: Unbalanced datasets are corrected by using data resampling techniques. One key tactic for addressing the imbalance is the under-sampling approach [9]. Oversampling is an attempt to balance out biased datasets, oversampling procedures boost the samples from minority groups. Often employed in image recognition applications, random oversampling (ROS) is an oversampling approach that involves rotating images or copying or recreating pre-existing data [10]. SMOTE: SMOTE is a synthetic instance creation tool available in multiple versions. Basic SMOTE interpolates between minority class samples by oversampling those instances that are on the edge of the sample. By focusing on data that are near the decision boundary, borderline SMOTE improves classification efficiency [10]. The process of applying knowledge gained from solving one problem to another that is similar but distinct is known as transfer learning in machine learning [11]. Transfer learning is a subfield of deep learning techniques that involves using models that were trained on larger datasets to optimize their performance on a smaller labeled dataset. The idea behind transfer learning (TL) is that people can use skills they've already learned to address contemporary challenges more quickly and accurately [11, 12].

A. Common approaches for addressing data imbalance for dynamic labeled data

Define Particular challenges arise when handling imbalances in dynamically labeled data. Here are some techniques for using machine learning algorithms, neural networks, and deep learning. In dynamic labeled data, Drift Detection Methods (DDMs) are essential for maintaining an equilibrium state. DDMs track changes in class frequencies or the appearance of new classes as well as the efficacy of machine learning models when label distributions change over time.By alerting users when datasets need to be rebalanced through changes to sample plans or model parameters, DDMs ensure proactive adaptation to dynamic contexts, which leads to robust and dependable model performance [17]. Since recurrent neural networks (RNNs) are capable of recognizing temporal links and sequential patterns, they are particularly well adapted to handle input that has been dynamically tagged. Even if labels change or new classes emerge, RNNs remain responsive to changing class frequencies by updating predictions based on data from previous time steps [18]. Because LSTM networks can adapt to shifting label distributions and capture long-range relationships, they are a promising solution for dynamically labeled data balancing. By selectively remembering or forgetting information according to its importance, LSTMs enable adaptation to label changes and continuous adaptation to variations in the data stream [18, 19].

III. Methods

For understanding the domain dynamics for effective machine learning applications, some of the key concepts are very important. Fundamental concepts such as domain, task, feature space,label space, and predictive function. It sets the stage for exploring adaptation methods withina structured framework. Some key concepts essential for understanding domain adaptation in transfer learning are mentioned in this section. It begins by defining a "domain" as consisting of a feature space X and a marginal probability distribution $P(X)$, where X represents feature vectors and $P(X)$ denotes their distribution. An example of document classification is provided to illustrate these concepts. The task comprises a label space Y and an objective predictive function $f(\cdot)$. This function, learned from training data pairs $\{(x_i, y_i)\}$, predicts labels for new instances based on their features. This introduces a probabilistic viewpoint where $f(x)$ can be expressed as $P(y \mid x)$, particularly useful in document classification tasks. The focus of the survey is then highlighted: a single source domain D_S and a single target domain D_T. The source domain consists of data instances x_{S_i} and their corresponding class labels y_{S_i}, while the target domain contains instances x_{T_i} and labels y_{T_i}. Notably, the target domain typically has fewer instances than the source domain. This simplified scenario aids in analyzing domain adaptation techniques across various domains and tasks. Space X and a marginal probability distribution $P(X)$, with differences between source and target domains indicated by disparities in either feature spaces or marginal distributions ($D_S \neq D_T$). For example, in document classification, this could mean variations in term features or distribution patterns between source and target document sets. Likewise, a task consists of a label space Y and a conditional probability distribution $P(Y \mid X)$,with discrepancies between source and target tasks denoted by differences in label spaces or conditional distributions ($T_S \neq T_T$). This could occur when source and target domains have distinct classes or uneven class distributions. Moreover, when there is a relationship between the feature spaces of the two domains, they are considered related, allowing for the transfer of knowledge between them. Mathematically, the relationship between domains and tasks is represented as follows:

Domain Definition: $D = \{X, P(X)\}$

Task Definition: $T = \{Y, P(Y \mid X)\}$

Additionally, the condition $D_S \neq D_T$ implies differences either in feature spaces ($X_S \neq X_T$) or marginal distributions ($P(X_S) \neq P(X_T)$), while $T_S \neq T_T$ indicates discrepancies in label spaces($Y_S \neq Y_T$) or conditional distributions ($P(Y_S \mid X_S) \neq P(Y_T \mid X_T)$).

Overall, this structured framework provides a comprehensive understanding of transfer learning, emphasizing its objective of leveraging knowledge across domains and tasks to improve learning performance.

IV. Results and Discussions

This section has covered two diverse types of datasets and two diverse sorts of methodologies. In the initial part of this discussion, we used an unbalanced numerical dataset for fraud detection. if there are two classes inside the classification. Consequently, we used binary classifiers for the same purpose. Algorithms like Gradient Boosting Machines (GBM), RandomForests, Decision Trees, Support Vector Machines (SVM), and Logistic Regression from the Machine learning domain are used to obtain the classification results. The techniques for data balancing are SMOTE, ADASYN, Random Under Sampling, and Random Over Sampling. In The second part of this post, we made use of our second imbalanced dataset which included pictures. Where the transfer learning method for data balancing is combined with the use of SMOTE and ADASYN. The Gradient Boosting Machines (GBM), Random Forests, DecisionTrees, Support Vector Machines (SVM), and Logistic Regression are the categorization models that are used.

A. Data Balancing with Machine learning

The dataset that we used in our experimentation was available for free download from Kaggle and was used for fraud detection. It contains transactions of credit cards made by customers in country Europe in September 2013. The communications that occurred in two days are shown in this dataset. 492 of the 284,807 transactions were fictitious. A significant amount of imbalance in the dataset means that although other transactions are normal, all transactions fall into the positive class where frauds are 0.172%. The PCA transform yielded numerical values for each of the input variables. The features V1, V2,..., and V28 are the important elements that were obtained using PCA; "Time" and"Amount" are the only features that were not altered using PCA. This feature is used for cost-sensitive learning. The instants that pass among each transaction and the preliminary transaction are stored in the field labeled "Time." The transaction amount is located in the attribute, “Amount class," the outcome of these variables, takes on a value of 1 in the case offraud and 0 in the absence of it.

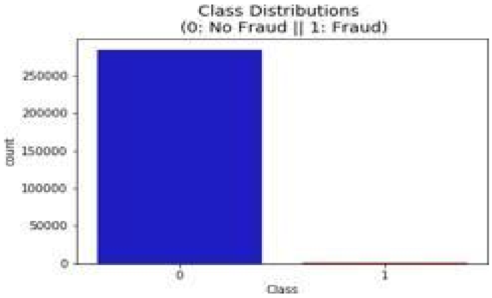

Figure 1: Data Visualization for imbalanced class

We can determine the degree of skewness in these qualities and look at further distribution for the other aspects by looking at the distributions in Figure 1 above. We shall employ techniques that may aid in lessening the skewness of the distributions in this research

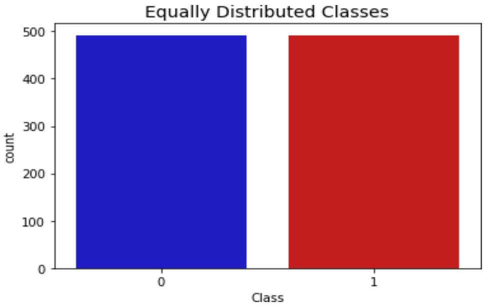

Figure 2: Balanced class after Under-Sampling

Table 1: Statistics of the balanced dataset

Dataset	**Fraud**	**Not Fraud**
Original	492	284,807
Random Over Sampling	284,807	284,807
Random Under Sampling	492	492
SMOTE	284,807	284,807
ADASYN	284,807	284,807

Table 2: Results after balanced data classification by multiple machine learning algorithms

Framework	SVM	LR	DT	RF	GBM
Precision	0.9355	0.9423	0.9356	0.9012	0.9499
Recall/Sensitivity	0.7212	0.7312	0.7411	0.7819	0.7632
Specificity	0.5634	0.6723	0.7212	0.6532	0.6833
F1 Score	0.8365	0.8700	0.8923	0.8987	0.8723
Accuracy	0.8912	0.9198	0.9001	0.9003	0.9845

B. *Data Balancing with Transfer Learning*

The process for resolving data imbalance in transfer learning involves choosing the appropriate model first, fine- tuning it next, resampling the data third, and evaluating the model fourth. To perform transfer learning, the first step is to choose a suitable pre-trained model and transfer strategy based on the properties of the dataset. It is possible to achieve domain adaptability, fine-tuning, and

To obtain a robust evaluation, employ methods such as cross- validation instead of concentrating solely on accuracy.

There are 5,863 X-ray images (JPEG) and two categories (Pneumonia/Normal). A prospective cohort of pediatric patients, aged one to five, at Guangzhou Women and Children's Medical Center, Guangzhou, were utilized for selecting anterior-posterior chest X-ray images. Every chest X-ray had been taken as part of the patient's regular medical treatment.

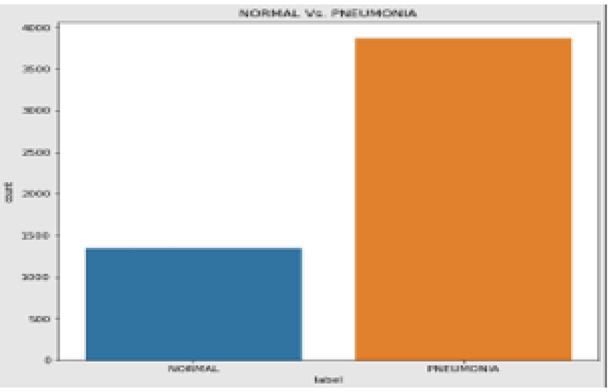

Figure 3: Data visualization

feature extraction. Adjust the model architecture, hyperparameters, and data processing to suit your needs. Use your dataset to train the enhanced model and undertake iterative performance evaluations. These processes make it possible to create models quickly and improve performance, particularly when occupied with modest quantities of labeled data.Weights and hyperparameters must be changed while fine-tuning the pre- trained model utilizing unbalanced data in order to maximize performance. It's crucial to properly adjust parameters like learning rate and regularization strategies to avoid overfitting or underfitting. Early stopping and dropout strategies improve convergence and address class imbalance.

The following step involves resampling the data by changing the sample numbers to balance the classes. To undersample majority classes, strategies include using cluster-based or random techniques; to oversample minority classes, strategies like SMOTE or ADASYNare employed.By lowering bias and variance, resampling improves model accuracy and recall, but it can also generate noise. In the final stage, assess the effectiveness of the transfer learning algorithm on unbalanced data. Use metrics like precision, recall, f1-score, ROC curve, and confusion matrix to account for class imbalance.

V. Conclusion

In conclusion, this research addresses the challenge of dealing with unbalanced data sets in machine learning. It explores various balancing techniques such as random over-sampling, random under-sampling, SMOTE, and ADASYN in numerical and image data sets. For fraud detection, classification systems such as gradient increase machines, random forests, and others were used, showing improvements in data balance performance. In pneumonia detection, transfer learning is used along with SMOTE and ADASYN, with models such as DenseNet, ResNet, and VGG-16. The results highlight improvements in classification accuracy with balanced data. The study emphasizes the importance of addressing class imbalances in a reliable and generalized machine learning model, contributing to real-world applications, and guiding the development of more robust solutions.

We have classified the Pneumonia class and the Normal class using transfer learning models utilizing the previously provided data set. The algorithms that are used for experimentation areDenseNet, ResNet, Inception V3, VGG-16, and VGG-19. The outcomes and the comparison are discussed below.

Given that a multiclass multilabel task has been considered, a few model-related observationsshould be made. First off, Softmax shouldn't be utilized as a production layer since it promotessingle-label forecast. One popular output purpose used in multilabel classes is the sigmoid. When the sigmoid and the loss function are combined, sigmoid optimization of the numerical reliability of the loss function is accomplished. Consequently, this also minimizes the sigmoid. The results of balanced data categorization using several transfer learning models are shown inTable 3.

Table 3: Results balanced data classification by multiple Transfer Learning Models

Framework	VGG-16	VGG-19	ResNet	Inception-V3	DenseNet
Precision	0.9355	0.9423	0.9356	0.9012	0.9499
Recall/Sensitivity	0.7212	0.7312	0.7411	0.7819	0.7632
Specificity	0.5634	0.6723	0.7212	0.6532	0.6833
F1 Score	0.8365	0.8700	0.8923	0.8987	0.8723
Accuracy	0.8912	0.9198	0.9001	0.9103	0.9645

Reference

[1] Wang, Le & Han, Meng & Li, Xiaojuan & Zhang, Ni & Cheng, Haodong. (2021). Reviewof Classification Methods on Unbalanced Data Sets. IEEE Access. PP. 1-1.

[2] Domenici, Andrea & Donno, Flavia. (2009). Static and Dynamic Data Models for the Storage Resource Manager v2.2. J. Grid Comput.. 7. 115-133. 10.1007/s10723-008-9110-3.

[3]]Üstüner, Mustafa & Balik Sanli, Fusun & Abdikan, Saygin. (2016). BALANCED VS IMBALANCED TRAINING DATA: CLASSIFYING RAPIDEYE DATA WITH SUPPORT VECTOR MACHINES

[4] Zhang, Jing & Wu, Xindong & Sheng, Victor. (2016). Learning from crowdsourcedlabeled data: a survey. Artificial Intelligence Review. 46. 10.1007/s10462-016-9491-9

[5] Goldman, S. & Zhou, Y. (2000). Enhancing supervised learning with unlabeled data. In Proceedings of the Seventeenth International Conference on Machine Learning, pp.327-334,San Francisco, CA .

[6] R. J. Lyon, J. M. Brooke, J. D. Knowles and B. W. Stappers, "A Study on Classification in Imbalanced and Partially-Labelled Data Streams," 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 2013, pp. 1506-1511, doi: 10.1109/SMC.2013.260

[7] Bertini, J.R., Lopes, A.d.A. & Zhao, L. Partially labeled data stream classification with the semi-supervised K-associated graph. J Braz Comput Soc 18, 299–310 (2012). https://doi.org/10.1007/s13173-012- 0072-8

[8] J. B. Saxe and J. L. Bentley, "Transforming static data structures to dynamic structures," 20th Annual Symposium on Foundations of Computer Science (sfcs 1979), San Juan, PR, USA, 1979, pp. 148-168, doi: 10.1109/SFCS.1979.47.

[9] Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B Cybern. 2008, 39, 539–550

[10] Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.

[11] a friendly introduction. J Big Data 9, 102 (2022). https://doi.org/10.1186/s40537-022-00652-w

[12] Al-Stouhi, Samir & Reddy, Chandan. (2015). Transfer Learning for Class Imbalance Problems with Inadequate Data. Knowledge and Information Systems. 48. 10.1007/s10115-015-0870-3.

[13] Liu, X., Zhu, S., Yang, F.et al. Research on unsupervised anomaly data detection method based on improved automatic encoder and Gaussian mixture model. J Cloud Comp 11, 58 (2022). https://doi.org/10.1186/s13677-022-00328-zY C a, Padmanabha & Pulabaigari, Viswanath & B, Eswara. (2018). Semi-supervised learning: a brief review. International Journal of Engineering & Technology. 7. 81. 10.14419/ijet.v7i1.8.9977.