A K-means Clustering Based Under-Sampling Method for Imbalanced Dataset Classification

1st Chih-Ming Huang
dept. of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
andrewh232@gmail.com

2nd Chuan-Sheng Hung
dept. of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
cshung@g-mail.nsysu.edu.tw

3rd Yao-Yuan Hsu
dept. of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
yuanster@g-mail.nsysu.edu.tw

4th You-Cheng Zheng
dept. of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
yczheng7652@gmail.com

5th Cheng-Han Yu
dept. of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
yu199483@gmail.com

6th Chun-Hung Richard Lin
dept. of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
lin@cse.nsysu.edu.tw

7th Shi-Huang Chen
dept. of Computer Science and Information Engineering
Shu-Te University
Kaohsiung, Taiwan
shchen@stu.edu.tw

***Abstract*—This paper explores the challenges of applying imbalanced datasets to machine learning models. There is a significant disparity in the quantities of different classes, leading to biases in the learning process. Consequently, many models favor predicting the majority class to enhance accuracy, neglecting potentially crucial minority class data. This bias results in an overly optimistic perception of predictive outcomes, leading to erroneous decision-making. This paper proposes a classification sampling method that solves the issue of original imbalanced data. The process is divided into two main parts. The first part involves data preprocessing, utilizing the K-Means algorithm to cluster majority class data. Additionally, it replaces the traditional Euclidean distance with the Hellinger distance as the similarity metric for data clustering. The second part utilizes the representative data obtained in the previous step as training input, thereby reducing the imbalance between different classifications. Finally, experimental results using the imbalanced Kawasaki Disease (KD) dataset, with an imbalance rate of 64.35%, demonstrate that the proposed Hellinger method improves Precision (PPV) by 20.2% compared to the XGBoost method when Recall is above 90%. This effectively addresses the classification bias of traditional learning methods towards imbalanced data, enhancing predictive outcomes for imbalanced datasets. This approach holds potential applications in medical disease prediction, financial fraud detection, and other related domains in the future.**

Keywords—Imbalanced Dataset, Sampling, Correlation, K-means, Euclidean Distance

I. Introduction

Classification problems hold a central and pivotal position within machine learning. These challenges encompass the rigorous analysis of data features to render predictions regarding their probable category. Generally, they are categorized into two principal domains: binary classification and multi-class classification. Notably, binary classification, for instance, boasts a broad spectrum of applications across various domains, encompassing fault detection and the diagnosis of diseases.

When delving into the realm of machine learning, the endeavor of constructing models utilizing imbalanced datasets emerges as a frequently encountered yet formidable challenge, particularly when one is dealing with datasets characterized by equilibrium[1]. The underlying rationale for the intricacy inherent in this task is grounded in the tendency of models trained on imbalanced data to manifest predictive biases. This bias arises as a consequence of the model's excessive fixation on achieving accuracy, thereby inducing it to misconstrue predictions belonging to the majority class as the definitive overarching verdict of the model. Consequently, the model favors the majority class in its predictive endeavors, resulting in recurrent inaccuracies when predicting data from the minority class. Such a predicament constitutes a routine circumstance across diverse real-world scenarios, encompassing financial fraud detection, medical prognostication, and insurance claims assessment, thereby establishing it as a widely pervasive and pragmatic issue.

In medical data analysis, it is discernible that, in most cases, the quantity of positive samples is significantly lower than that of negative samples compared to individuals who have not been diagnosed with the condition. The genesis of this issue can be attributed to two primary factors: (1) the inherent low prevalence rate of the ailment, making the acquisition of positive patient data a formidable task, and (2) inherent deficiencies and lapses in the data collection process, leading to a conspicuous disparity in the quantity of positive data to negative data.

Predictive models tend to incline their outcomes towards the more frequently occurring negative cases when encountering the problem of imbalanced data. While this invariably results in a commendably high overall model accuracy, it simultaneously neglects the critical aspect of accurately predicting positive cases. Such a model, which neglects the prediction of positive cases, evidently fails to serve a meaningful purpose in practical medical diagnosis. We genuinely require a model capable of correctly identifying the

979-8-3503-3094-6/24/$31.00 ©2024 IEEE

rare positive cases instead of a mere pursuit of overall accuracy.

With the ubiquitous integration of artificial intelligence across diverse domains, medicine, and healthcare are no exception. Presently, the utilization of artificial intelligence within the medical sphere primarily revolves around augmenting the diagnostic capabilities of physicians and forewarning potential health perils. However, the current application of artificial intelligence in the medical field is mostly to assist doctors in diagnosing patients, reminding them of the potential risks of disease, and preventing patients from missing the best treatment time. At the same time, It will also incur extra effort and costs for patients and medical resources.

Processing methods predominantly encompass two primary directions, the algorithmic and data levels [2-4]. Techniques at the algorithmic level often incur comparatively elevated developmental costs. In contradistinction, methods at the data level can be further delineated into over-sampling and under-sampling. Nevertheless, owing to the inherent constraints of algorithmic attributes, the under-sampling approach is susceptible to overfitting the dataset. Conversely, the under-sampling regimen may readily engender the forfeiture of pivotal attributes, resulting in a substantial decline in the discriminative accuracy of the majority class data while enhancing the precision of minority class data. The purview of this study is centered on the data under-sampling facet. We aspire to employ a clustering algorithm to enhance data curation for under-sampling. Through more precise data curation, we can more effectively discern salient from inconsequential data attributes. This can avert the loss of vital features, concurrently ameliorating the predictive accuracy, recall, precision, and other metrics associated with minority class data.

This paper will sequentially discuss related work, such as Imbalance Datasets and Clustering Algorithms. It will also elaborate on the system architecture, experimental results, and conclusions of this study.

II. Related work

A. Imbalance Datasets

In recent years, data acquisition has become increasingly attainable and diverse. Nonetheless, post-data acquisition, a multitude of data still harbor inherent limitations, with data set imbalance being a prevalent predicament. Within the conventional realm of machine learning, we often make the presumption that the categories within the data set exhibit similarity in numbers. Regrettably, practical data does not align with such idealism, and substantial disparages between category counts frequently manifest. Consequently, the predicament of imbalance extends beyond the binary classification paradigm. So long as there exists a substantial magnitude differential among various classifications, it qualifies as an imbalanced data set.

The problem of data set imbalance often occurs in various types of data such as medical, financial, and transportation. These data usually have the characteristics that positive data are challenging to collect, so there are considerable differences in magnitude between different data types. The magnitude of the huge difference in models has also led to misunderstandings in the application of traditional machine learning, and the accuracy used by general machine learning to measure the quality of model standards is not suitable as a measurement indicator, which is the so-called accuracy paradox, on the contrary, precision (PPV), recall (Recall, Sensitivity) and ROC curve should be used as the criterion for judging the quality of the model [5]. The current methods for processing imbalanced data sets are mainly divided into two major directions: data-level and algorithm-level methods. Next, the data-level issues will be explained.

B. Imbalanced Data Set Processing (Data Level)

Data-level methods mainly operate on the original data to change the category distribution, allowing traditional algorithms to process subsequent data, such as over-sampling and under-sampling. Combined over and under sampling method [6].

- Over-sampling: The over-sampling method operates on a smaller number of classification samples, expands the number of minority categories in a specified way, and balances the quantitative gap between different categories by increasing minority category samples to solve the problem—the problem of data set imbalance. Representative algorithms include the random upsampling algorithm, SMOTE (Synthetic Minority Over-sampling TEchnique) [7] and its extended related algorithm Borderline-SMOTE [8], as well as ADASYN (Adaptive Synthetic Sampling Approach) [9] and ADOMS (Adjusting the Direction Of the synthetic Minority classes examples)[10]. Among them, the SMOTE algorithm is the most representative. Its main steps are: (1) Use KNN (K-Nearest Neighbors) to select K neighboring minority class samples for the minority class sample (2) Randomly select a sample from them to calculate and compare with the original sample After multiplying the gap (3) by a random weight between 0 and 1, a new sample is synthesized with the original sample (4) and the first two steps are repeated until the number of samples is satisfied.
- Under-sampling: The under-sampling approach employs a contrary strategy to the over-sampling method. It meticulously culls a substantial number of classification samples, thereby pruning the instances from the majority category and diminishing their count. This endeavor attains a quantitative equilibrium among distinct categories. Notable algorithms encompass the random under-sampling technique, the NearMiss algorithm [11], and its expanded counterpart. While the under-sampling method capably addresses the challenge of category imbalance and mitigates the risk of overfitting, it carries the potential for inadvertent removal of pivotal key attributes alongside the majority category samples. Consequently, the model may overlook plausible feature combinations, resulting in a marked decline in the accuracy of classifications for the majority class. Hence, the preservation of critical features during this process stands as the focal concern this paper aspires to resolve.
- Hybrid method of Undersampling and Oversampling: Through the amalgamation of diverse sampling methods, it mitigates the over-replication of minority class data and curtails the deletion of prominent class attributes. This aids in harmonizing the data within the

model, thus sidestepping overfitting of the training data and the excessive loss of pivotal features. Consequently, it yields commendable results in straightforward imbalance scenarios, notably through the synergy of SMOTE in tandem with ENN (Edited Nearest Neighbor) or SMOTE in conjunction with Tomek Links[12-13].

C. Clustering Algorithm

Clustering algorithms constitute a category within unsupervised learning. Prominent clustering algorithms encompass K-means[14] and DBSCAN (Density-Based Spatial Clustering of Applications with Noise)[15], primarily relying on data for their operation. To execute clustering, they establish a distance calculation approach between each data point, employing these inter-data distances as the foundation for ascertaining data similarity. Ultimately, after numerous iterative partitions, the conclusive clustering outcome emerges.

The quintessential and emblematic clustering algorithm is K-means. This algorithm iteratively computes the distances between the data points and the cluster center, designating the closest centroid as the novel cluster center. This process persists until convergence is achieved. In addition to using distance as a clustering judgment, there are also algorithms that use data distance to divide the calculation range and further calculate the density of data clusters. The proximity between data groups is harnessed as a discerning criterion, further delineating anomalous data points during the clustering process. When it comes to normal data, the exemplar algorithm is DBSCAN.

- K-means: Among the unsupervised clustering methods, a highly representative classical algorithm continuously refines the classification group to which each datum pertains. This refinement is achieved through an iterative process that meticulously assesses the distance between each datum and the cluster center, persisting until data convergence is achieved. The K-means algorithm delineates its operation through the ensuing sequence of steps: (1) Initialization of cluster centroids. (2) Calculation of distances between data points and cluster centroids. (3) Reclassification of data points based on the computed distances to the nearest cluster centroid. (4) Recalibration of the mean center for each cluster. (5) Repetition of steps 2 to 4 until the gravitational center of each cluster exhibits no further alteration, as elucidated in Algorithm 1.

Algorithm 1：$K - means\ (D, K)$

```
Input：Input data D = $\{d_1, d_2, \cdots d_n\}$ ; Number of cluster $K$
Output：$K * set\ of\ cluster$
Initialisation： $K$ initialise cluster centers: $m_1, m_2, \cdots, m_k \in D$
repeat
 for each iteration do
   $C_j \leftarrow \emptyset$ for all $j = 1, \cdots, K$
   compute distance between data and cluster centers $r_{nk}$
      for $N \leftarrow 1$ to $n$
         $j \leftarrow argmin_j\{r_{Nk}\}$
         $C_j \leftarrow C_j \cup \{d_N\}$
      endfor
 for $k \leftarrow 1$ to $K$
   $m_k \leftarrow \frac{sum(d_i)}{|C_k|}$ for $d_i \in C_k$
 endfor
 until converged
```

- K-Medoids: While the K-means algorithm exhibits simplicity and ease of application, its adaptability across diverse scenarios remains relatively constrained. This constraint arises from the derivation of cluster centroids in the K-means algorithm, which relies on the arithmetic mean of data points within each cluster—a rather straightforward process. However, this method is susceptible to the influence of outliers in the data, causing the cluster centroids to shift towards these extreme values. Consequently, when employing different distance metrics as the foundation for clustering, K-means encounters challenges in defining an appropriate cluster centroid, rendering its calculations futile in scenarios employing distinct distance measures. Conversely, the enhanced K-medoids algorithm adeptly mitigates the interference posed by outliers. Furthermore, by abstaining from the averaging of data within clusters during calculations, K-medoids selects the true data point with the minimal sum of distances to serve as the cluster centroid. This capability enables K-medoids to seamlessly accommodate various distance metrics for clustering, thereby enhancing the diversity and flexibility of its applications.

D. Similarity Algorithm

Hellinger distance: The Hellinger distance is a metric that extends the Bhattacharyya Coefficient (BC) and quantifies the structural resemblance between probability distributions, as exemplified by (1).

$$H(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\Sigma_{i=1}^{n} \left(\sqrt{P_i} - \sqrt{P_i}\right)^2} \qquad (1)$$

$H(P, Q)$ represents the Hellinger distance between distributions P and Q . P_i and Q_i represent the probabilities associated with the ith event or outcome in the distributions P and Q, respectively. These probabilities should sum to 1 for both distributions. The summation $\sum_{i=1}^{n}$ is taken over all possible events or outcomes in the distributions. Here, the Hellinger distance is essentially a measure of the "closeness" of two probability distributions. It ranges between 0 (indicating identical distributions) and 1 (indicating completely dissimilar distributions). In practice, the Hellinger distance is often used in applications such as statistical hypothesis testing, clustering, and machine learning to compare the similarity or dissimilarity of data distributions.

III. Architecture

A. Cluster Sampling Method

The concept of cluster sampling draws its lineage from the realm of statistical stratified sampling. By dividing the data matrix into multiple independent strata, the mutual differences between each stratum are significant, and each piece of data can only belong to a single stratum. As it were, this stratified division culminates in an independent perusal of data drawn from diverse strata, thus ensconcing an endeavor to harmonize the sampled data with utmost fidelity to the original data distribution of the parent populace. Through this stratagem, the sampled data attains a paramount significance in the context of the parent population. The resultant model, constructed through sampled data, bestows a profound enhancement in both its representational prowess and its veracious precision.

However, the riddle of clustering within the parental data enclave emerges as another formidable quandary. It is imperative to lucidly elucidate the confluences and disparities pervading the data cosmos to coax forth an exemplary clustering. Significantly, when the data dimensions increase, the representativeness of the Euclidean distance for each other's data also decreases. Therefore, we introduce the concept of similarity calculation to help us more accurately distinguish the similarities and differences between data.

B. Similarity Clustering

The use of similarity is to solve the problem that Euclidean distance is not representative enough for the similarity between data, so the similarity results between data are used as the calculation basis for clustering, but this also derives the following two problems, (1) When using non-Euclidean distance, how should we select a reasonable cluster center for the next step of clustering calculation, and (2) how to reasonably convert the similarity so that it has the property of distance: the more The closer the distance between similar data is to each other, and there is no distance less than 0, then it can be added to the clustering algorithm for calculation.

In order to solve the first problem, this study refers to the idea of K-medoids and uses similarity distance as the distance basis of K-medoids. The final similarity clustering algorithm process can be divided into the following five steps: (1) Select clusters Class center (2) Calculate the similarity distance between the data and the cluster center (3) Update the cluster to which the data belongs according to the similarity distance between the data and the cluster center (4) Update the sum of similarity distances between each data, and Use the data with the smallest similarity distance as the center of the new round of clustering (5) and repeat steps 2 to 4 until the clustering results no longer change, as shown in Fig. 1.

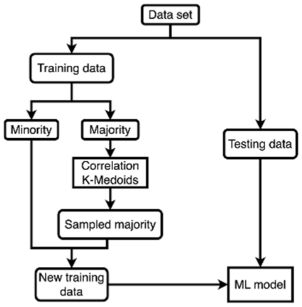

Fig. 1. Research architecture diagram of cluster sampling method

In transforming similarity to adhere to the essential attributes necessary for distance computation within clustering algorithms while preserving the outcomes of similarity calculations, the Hellinger distance is additionally employed for gauging the resemblance between two probability distributions. It inherently embodies the concept of distance.

C. Data Sampling Within Clusters

Although the clustering algorithm has completely classified the similarity of the data for us, if we do equal sampling for clusters of different sizes, it will obviously lack representation for different features in the data. To faithfully mirror the proportions of distinct majority category attributes within the dataset, this study has opted to employ the quantity of data within each cluster as the arbiter, judiciously apportioning the anticipated number of sampled data uniformly across clusters in accordance with their respective sizes. As illustrated in Figure 2 below, this approach entails extracting a substantial volume of data from the larger clusters, thereby enhancing the fidelity of the original data distribution.

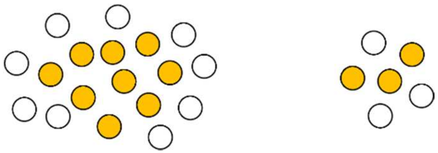

Fig. 2. Cluster size sampling diagram

After ascertaining the quantity of data to be chosen for each cluster, this study shall employ the standard deviation method to partition the data points residing within the cluster, demarcating the cluster's data domain into the ensuing three parts: 0 times to 1 times the Standard Deviation, 1 times to 2 times the Standard Deviation, and each partition's data is sampled in accordance with the proportion of the normal distribution. A total of 68% of the requisite data shall be selected from the 0 times to 1 times Standard Deviation range, with the remaining 32% of the data being drawn from the interval spanning 1 to 2 times the Standard Deviation, as illustrated in Figure 3 below. Through such meticulous segmentation in the selection of cluster data, this study aspires to augment the diversity of data within the cluster, mitigating the impact of an insufficient number of data points or an undue concentration thereof.

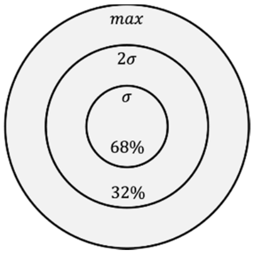

Fig. 3. Schematic diagram of data segment selection

In addition, when the number of data in a cluster is too small, statistical indicators such as calculating distance standard deviation have lost their representative significance. Consequently, we establish a threshold for the data count within each cluster. In the case of clustering outcomes yielding fewer than 30 data points, this study refrains from executing segmented data selection. Instead, subsequent to ascertaining the requisite data count for the cluster, it opts for a randomized data selection approach, thereby mitigating the risk of ineffective standard deviation assessment.

IV. Experimental Results

The dataset utilized in this investigation originates from Kawasaki disease patients (KD) within the age bracket of 0 to 5 years, who were treated at Chang Gung Memorial Hospital over the decade spanning from 2010 to 2019. In parallel, fever-afflicted patients (FC) within the same age range served as the control cohort. This dataset comprises three personal attributes: gender, age, and month of consultation, along with seventeen hematological markers and two urinary test parameters. A cumulative of twenty-two distinct attributes were harnessed to establish and analyze the model. Systematically structuring the dataset's content and distribution contributes to a deeper comprehension of its composition and arrangement, which proves advantageous for subsequent research and experimentation.

In this examination, the count of Kawasaki disease-positive patients tallied at 1142, whereas the number of fever patients constituting the control group reached 73,499. The statistical illustration underscores the comparability between these numerical values, with the discrepancy approximating a ratio of nearly 65 to 1.

According to research, men have a higher chance of being infected with Kawasaki disease than women, and the gender distribution in this data set is consistent with this research result. There are 687 cases of male Kawasaki disease patients, while women Kawasaki disease patients had only 455. For fever patients, although the number of male patients is higher than that of female patients, there is no obvious difference like Kawasaki disease. There are 41,465 male fever patients, while female fever patients have 41,465 fever patients. There are 32034 transactions. Age is an important factor in determining whether Kawasaki's disease is present. The chance of developing Kawasaki disease decreases as you get older. There is almost no Kawasaki disease when you are over 6 years old. This is why this study selected 0. Children aged 5 years old are the targets of analysis. Looking at Kawasaki disease's age distribution, younger children have a relatively higher proportion of diagnosed Kawasaki disease.

The age distribution of general fever patients gradually decreases with age. This may be because the immune system of newborn children is still developing, and their ability to resist infection is relatively weak, making them more susceptible to infection. Fever, but although the proportion decreases with age, it can still be seen that the incidence rate of Kawasaki disease patients decreases with age much faster than that of general patients with fever.

The clustering algorithm employed in this study employs the Hellinger similarity measure. During the experimental phase, only the majority of categorical data will undergo classification. The number of categorized groups is a multiple of 'n' times the quantity of minority category data, with 'n' ranging from 1 to 20, progressively increasing. Post-cluster sampling, the ratio of minority categories to majority categories is 1:4. following clustering, this research will meticulously select the clustering combination that exhibits the most outstanding Recall performance among the results. This selection criterion requires Recall to exceed 80%, 85%, and 90% by practical necessities.

Figure 4 depicts the outcomes of the clustering operation employing the Hellinger distance metric. Here, the x-axis represents the 'n' value in the clustering multiples, while the y-axis denotes the computed values of various indices.

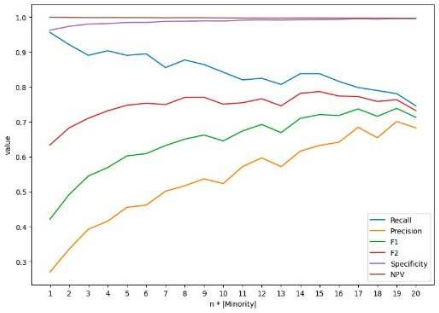

Fig. 4. Hellinger distance clustering sampling result chart

The Hellinger distance yields the highest Precision results when the clustering multiplicities are as follows: 16, 9, 4, and 1 times, as indicated in Table I It is discernible that even when Recall exceeds 80%, Precision maintains an accuracy of 64.1%. Furthermore, with Recall surpassing 90%, Precision still upholds an accuracy of 41.5%. This underscores the study's commitment to identifying all superficial patients while preserving the model's ability to discern ordinary patients. This also retains the model's ability to identify ordinary patients, reducing the occurrence of false positives so that the model can effectively help doctors make diagnoses in practical applications.

TABLE I. HELLINGER DISTANCE HAS THE BEST RESULTS UNDER HIGH RECALL

	80%	**85%**	**90%**	**95%**
Recall	81.6%	86.4%	90.4%	95.6%
Specificity	99.3%	98.9%	98.1%	96.2%
PPV(Precision)	64.1%	53.7%	41.5%	27.0%
NPV	99.7%	99.8%	99.9%	99.9%
F1	71.8%	66.2%	56.9%	42.1%
F2	77.4%	77.0%	73.2%	63.4%
N	16	9	4	1

Based on the preceding analysis, it is observed that as the clustering parameter 'n' increases, the Recall value gradually diminishes. Conversely, the Precision metric exhibits an inverse behavior, steadily augmenting as the clustering parameter increases. These two indices, however, share a commonality in their rates of change, which decelerate as the clustering parameter escalates. Precision and Recall gradually converge as the parameter approaches the value of 20, settling at approximately 0.7.

In conclusion, this study supplements the traditional XGBoost results with the similarity clustering algorithm for comparative analysis. It can be deduced that there has been a discernible enhancement following the adoption of the similarity clustering algorithm. Focusing on the Precision results of Hellinger distance clustering sampling, this study has achieved a remarkable 20.2% improvement in Precision compared to XGBoost under a 90% Recall rate. For Recall rates of 80% and 85%, which exceed 90%, and with Precision serving as the evaluation criterion, it becomes evident that the proposed Hellinger distance method in this paper exhibits an average enhancement of 12.03% when juxtaposed with the XGBoost approach. Consequently, this paper effectively elevates the Precision of predictions by applying the Hellinger distance methodology.

V. CONCLUSION

In the contemporary age of rapid progress, the issue of imbalanced data has emerged as a salient concern that commands unwavering attention. Whether in finance, medicine, fault prediction, or other projects, unbalanced data may appear. Nevertheless, it is noteworthy that many of the preeminent machine learning algorithms today have yet to be suitably tailored and calibrated for this particular difficulty. Concurrently, this problem poses a formidable impediment to the efficacy of machine learning prognostications in practical, real-world applications. Within this context, the cluster sampling methodology expounded upon in this study is not bound by the confines of specific domains. It is our aspiration that this approach can be readily employed in diverse scenarios where the challenge of imbalanced data rears its head, encompassing instances such as the diagnosis of ailments within the realm of medicine, the prediction of financial malfeasance, and the anticipation of faults in the industrial manufacturing sector, among others.

In contrast to the conventional XGBoost technique, the methodology delineated in this paper bequeaths an enhancement in Precision averaging 12.03%. With a Recall rate of 90%, Precision experiences a remarkable surge, amounting to an impressive 20.2%. Consequently, the technology in this manuscript holds the potential to enable developers to obtain more efficacious predictions pertaining to imbalanced data, all while curbing development costs. Furthermore, it stands to augment the accuracy, recall, Precision, and other associated metrics of the model.

REFERENCES

[1] H. He and E. A. Garcia, “Learning from Imbalanced Data,” IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[2] G. Haixiang et al., “Learning from class-imbalanced data,” Expert Systems With Applications, vol. 73, no. 73, pp. 220–239, May 2017, doi: 10.1016/j.eswa.2016.12.035.

[3] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” Neural Networks, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[4] P. Branco, L. Torgo, and R. P. Ribeiro, “A Survey of Predictive Modeling on Imbalanced Domains,” ACM Comput. Surv., vol. 49, no. 2, p. 31:1-31:50, Aug. 2016, doi: 10.1145/2907070.

[5] J. Stefanowski, “Overlapping, Rare Examples and Class Decomposition in Learning Classifiers from Imbalanced Data,” in Emerging Paradigms in Machine Learning, S. Ramanna, L. C. Jain, and R. J. Howlett, Eds., in Smart Innovation, Systems and Technologies. Berlin, Heidelberg: Springer, 2013, pp. 277–306. doi: 10.1007/978-3-642-28699-5_11.

[6] M. Harrison, Machine learning pocket reference: working with structured data in Python, First edition. Beijing ; Boston: O’Reilly, 2019.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[8] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” in Advances in Intelligent Computing, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, pp. 878–887. doi: 10.1007/11538059_91.

[9] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Jun. 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[10] S. Tang and S. Chen, “The generation mechanism of synthetic minority class examples,” in 2008 International Conference on Information Technology and Applications in Biomedicine, May 2008, pp. 444–447. doi: 10.1109/ITAB.2008.4570642.

[11] J. Zhang and I. Mani, “kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction,” Proceedings of workshop on learning from imbalanced datasets, vol. 126, pp. 1–7, Aug. 2003.

[12] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[13] M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, “Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data,” in 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), May 2016, pp. 225–228. doi: 10.1109/ICOACS.2016.7563084.

[14] J. Macqueen, “SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS,” MULTIVARIATE OBSERVATIONS, 1967.

[15] M. Ester, H.-P. Kriegel, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” 1996.