International Journal of Computer Sciences and Engineering
Vol.**11**, Issue **5**, pp.**06-12**, May **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Tackling Imbalance Datasets: Methods, Techniques & Comparisons

Shivam Kumar[1*], Deepanshu Ahuja[2], Sandeep Kumar[3]

[1,2,3]Dept. of Computer Science & Engineering, Sharda University, University, Greater Noida, India

**Corresponding Author: shivambhardwajdbgs@gmail.com*

Received: 22/Mar/2023; **Accepted**: 02/May/2023; **Published**: 31/May/2023. **DOI**: https://doi.org/10.26438/ijcse/v11i5.612

***Abstract*:** Over the past many years of continuous research and learning from data, i.e.duplication and Extraction continues to be a spotlight of enormous research. A classification data set with skewed class proportions is referred to as imbalanced. This term originated as a debate over the skewed distributions of binary tasks. Imbalanced data are those datasets that have an uneven distribution of observations across the target class, i.e First class category will have a very higher number of observations while the other class will have less number of observations. The emergence of the massive data era, along with the growth of machine learning and data mining (Data Science), as going deeper into the field of learning with imbalanced datasets, alongside the challenges which are emerging. Data-level methods and algorithm-level methods are repeatedly used and getting improved and popularity of hybrid approaches increased due to the extraction of earlier approaches (data level and algo level) and reduced weaknesses with powerful points.

In order to advance the field of addressing imbalanced datasets and compare existing approaches and methodologies, this paper attempts to discuss the open questions and challenges that need to be resolved. This essay discusses each of them and offers ideas for potential directions for further investigation. The main issue with an unbalanced class distribution is when bad training habits cause bias in favour of the majority class. Deep learning algorithms and machine learning algorithms perform training on datasets which are underrepresented in some categories. Conventional methods advise to perform undersampling on majority class category and oversampling minority class category before the learning stage.By including learning modules with clever representations of samples from majority and minority samples, this research investigates various traditional and contemporary strategies to address this issue. The works of several researchers are compiled in a very logical approach and numerical opportunities and also future difficulties for the field's future research are discussed.

***Keywords*:** Multiclass, Classification, Imbalance, Prediction, Majority, Minority, Synthetic Minority Over-sampling Technique(smote), Simplified Swarm Optimization(SSO), Particle Swarm Optimization (PSO), Adaptive Synthetic (ADASYN), Diversified One-vs-One strategy(DOVO), Diversified Error Correcting Output Codes (DECOC).

1. Introduction

We begin by defining an unbalanced dataset in the context of this scenario. We specifically define an unbalanced dataset as one in which the target class has an asymmetric distribution of observations. Supervised learning, unsupervised learning, semi-supervised learning, or a mix of the first two or all of these are examples of learning processes. Regression, classification, and clustering problems are all examples of imbalanced learning tasks. Output The accuracy of each method's results compared to those obtained using pure, unambiguous datasets will be examined.

The distribution of data is frequently assumed to be balanced in traditional machine learning techniques. However, conventional machine learning techniques frequently perform poorly when the class distribution is unbalanced. On unbalanced data, a suggested strategy for imbalanced learning performed satisfactorily. Oversampling techniques have been successful and are one of the hot topics in non-equilibrium learning, but they frequently have the drawback of destroying the distribution of the original data. For instance, the synthetic minority oversampling technique (SMOTE) linearly interpolates in the sample space to produce minority-class samples, whereas the majority-class sample space is frequently filled by newly formed samples. Additionally, intrusive sampling has an effect on later data processing, which affects classification performance.[2]

In addition to posing significant new challenges to the field of data research, imbalanced learning is also widely applied in real-world data, from civilian applications like financial and biomedical data analysis to security and defense-related applications like surveillance and security, i.e. Military data analysis. The recent notable growth in publications in the topic and the development of specialized seminars, conferences, symposiums, and special issues demonstrate the rising interest in this asymmetrical learning. Consider a

typical case study in the biological data analysis of cancer patients using the ML technique as a basic illustration of imbalanced learning. Consider a "mammography data set," which is a collection of pictures from a series of mammography examinations conducted on several individuals and is mostly utilized in breast cancer detection (reference from last sem project). For such a dataset, the resulting class will be Positive or Negative for images representing cancer patients or healthy patients, respectively. According to past experience, there will likely be more non-cancer patients than cancer patients. In actuality, there are about 10,923 Negative samples (majority class) and 260 Positive samples (minority class) in this dataset. In our dataset, we would prefer a ML model that offers a balance of accuracy from both majority and minority class categories. However, many common learning algorithms have classifier accuracies that are highly unbalanced, with accuracies near 100% for the majority class and between 0% and 10% for the minor class[4].

Traditional classifiers like decision trees and logistic regression cannot handle the imbalanced classes they contain, so simply seeking high accuracy on imbalanced datasets can be counterproductive[10]. This makes classes more likely to be large, and classes with few data points are treated as noise and are often ignored. As a result, the minority class has a higher misclassification rate compared to the majority class. Therefore, accuracy metrics are less important when evaluating the performance of models trained on imbalanced data.

This paper's focus extends beyond simply categorising issues into unbalanced categories. Instead, we describe different forms of Methods to tackle Imbalance datasets, Comparing the methods widely used in this domain and also, which algo which suits what type of Datasets. We present our position on these open issues in a promising way & Research directions to explore and address them [17].

Oversampling techniques are more appealing to multiclass classification problems because they equalize the quantity of samples from majority and minority classes in the training set while resampling the samples of the class with less data. According to the author, a straightforward random undersampling technique performs better than composite undersampling techniques[12]. SMOTE, a prominent oversampling approach, generates synthetic samples of the class with fewer samples in order to equalize the majority and minority classes in the training set. The primary idea behind the SMOTE approach is to construct samples based on the interpolation between multiple distinct features rather than just replicating the minority class samples, which is why it is considered to be focused on the "feature space" rather than the "data space".

A PSO optimized oversampling approach has been developed to overcome the aforementioned issues by optimizing synthetic samples from categories to increase the number of minority categories and as nearly as possible mimic the original data distribution. SMOTE is specifically designed to create synthetic samples.[13]. The best synthetic samples are then chosen using a decision tree classifier using an evolutionary technique called PSO in a way to increase samples of minority class. We compare the suggested approach's efficacy with that of the conventional oversampling method and the hybrid sampling method.

2. Literature Survey

In 2020, Fathy et al.[6] proposed a paper in which they said IoT industry is leading , transforming in the manufacturing world which is also known as smart manufacturing. A key pillar of intelligent manufacturing includes use of IoT data and use of machine learning (ML) to automate failure prediction, thereby reducing time involved in maintenance and cost resources and therefore improving quality of product, failures in this today's world almost always outweigh good performance examples. This difference is seen clearly in collected data from IoT enabled devices. imbalance datasets restrict the success of ML in failure prediction and thus poses a serious obstacle to the progress of smart functioning. And is the first to provide a framework for evaluating the effectiveness of these corrective actions in the context of manufacturing. Applying the proposed framework, we present a comprehensive comparative analysis evaluating the performance of different combinations of algorithmic components using real-world manufacturing datasets. Gain critical insight into the effectiveness of each component and the interrelationships between datasets, application context, and ML algorithm design.

Kaur et al.[8] proposed that A challenge with balanced data categorization occurs when the proportional class sizes of the datasets are considerably different from one another. The remaining samples fall into the other class, and at least one class is mapped with a small number of samples (referred to as the minority class) (called majority class). Fundamentally, classifier performance on this topic often matches a few number of classes (majority classes) in an uneven data set. Performance bias is the difference in behavior between solutions for the majority and minority classes. For most classes, solutions are usually more precise. However, there is a less accurate solution that runs on the minority side. In real-world applications like mistake, fraud, and bleeding detection in medical diagnostics, the issue of uneven data distribution is well-posed. Neural networks, safe-level SMOT, cost-sensitive algorithms, and neighborhood cleaning rules are the most widely used techniques for unbalanced data.

Krawczyk et al. proposed in 2009 [1] that It will be crucial to deepen your understanding and support decision-making due to the ongoing data availability expansion and many complex systems,such as: B. Basic discovery of knowledge and investigation and analysis from raw data, including surveillance, security, the Internet, and finance. creating a procedure The difficulty of providing learning done from (the imbalanced learning problem) has drawn criticism from both academia and industry, despite the fact that pre-existing data engineering and learning approaches had achieved

considerable major triumph in many applications. It's a difficulty that has just recently gotten noticed, but it is growing. The issue of unbalanced learning focuses on how learning algorithms function when there are only some points in data and significant category distribution biases. The intrinsic complexity of unbalanced datasets, learning from them necessitates novel ideas, strategies, and insights for effectively converting enormous volumes of raw data into representations of knowledge and information. Tools and algorithms are needed. This paper offers a thorough review of recent studies on learning from unbalanced data. Our attention is drawn to the nature of the issue, cutting-edge technology, and a critical analysis of the measures in use today to evaluate student learning outcomes in unbalanced learning scenarios. In order to encourage more study in this field, we also identify notable possibilities, problems, and potentially significant research avenues for learning from unbalanced data.

Krawczyk et al.[9] We now have a better grasp of the nature of imbalanced learning while addressing the difficulties that are arising thanks to the growth in field of mining of data and machine learning, as well as advent of the big data age. Methods based on algorithms and data feeds are constantly being developed, and hybrid strategies are becoming more and more common. Analysis of class asymmetries and other challenges posed by the nature of data are recent themes, as well. Researchers are concentrating on effective, adaptable, and real-time computing approaches in response to new practical situations. The goal of this essay is to examine the unresolved issues and difficulties that must be overcome in the process to advance the learning from machine learning. The whole spectrum of learning from imbalanced data is extended by seven key study topics in this field: classification, regression, clustering, data flow, big data analytics, and applications, such as in computer science and social media. For each of them, this page provides a discussion and recommendations for potential future study topics.

In 2021 Susan et al.[5] proposed that Their article focuses on one of the main problems that currently prevent data mining researchers from conducting experiments on existing datasets of the real-world. The main problem is the unequal category(class) distribution which leads to difference(bias) in favor of the majority category class since there are inadequate samples trained from minority category class. These days, underrepresented datasets are utilized to train deep learning and machine learning algorithms. However, because data for these classes is readily available, some other classes contain surplus samples. Conventional methods suggest to undersample the majority category class and/or oversample the minority category class in a way to fairly balance the distribution of class prior to learning . Although the difficulty of the unequal class category is frequently disregarded by some academics interested in learning technology, it is now necessary to incorporate correction of balance and data cutting strategies within learning process itself. In order to address this issue, the learning module in this study is given by smart representation of samples from both majority and minority classes category. This study looks at a wide range of classic and modern ways that do this. Both hybrid sampling techniques that choose and retain the challenging samples while discarding the simple samples are taken into consideration, as well as the use of natural evolutionary method algorithms to sapling intelligently and smartly. In a logical manner, the findings of diverse researchers are presented, and several opportunities and challenges for future study areas are explored.

A brand-new neighborhood-based undersampling (N-US) technique was put forth by Goyal et al. in 2020 [16]. As far as the authors are aware, there hasn't been any research that uses neighborhood-based undersampling methods for SDP. Therefore, this research adds to a fresh undersampling strategy. H.N.-US within the SDP sphere. This study also looks at N-potential US's as a trustworthy partner for SDP classifiers. Because ensembles have the inherent ability to handle unbalanced data, as shown by the literature, this study also considers the connections of the proposed N-US method. module. Algorithm-level, data-level, or ensemble approaches can all be used to handle unbalanced datasets. Resampling can be used to address data-level unbalanced datasets and solve this issue (oversampling and undersampling)[10].

3. Methodology

A. Remove Duplicate Data

The datasets are many times contain many similar and duplicate data points. Lets for food ordering ex. ‘Where is my Food’ and ‘Where is the Food Order’ means the same. Removing such duplicate messages helps in reducing size of majority .

Merging the Minority Classes: Multiple classes may have overlapping traits. Some algorithms consolidate some of these minority classes. This trick improved score by over 10% sometimes.

Resample the training set: The simplest strategy to balance an unbalanced dataset is done by simply oversampling the occurrences from minority class or undersampling the instances from majority class (data -level method and algorithm-level method).

B. Undersampling

An attempt to randomly removing data instances from majority class until the classes are balanced. There is a potential for loss of information leading to poor performance of model training.

Following is the pseudocode of Undersampling:
Step - 1 Count different classes present in dataset.
Step - 2 Divide all the dataset by classes.
Step - 3 UnderSample the majority class.
Step - 4 Concat the UnderSampled class with the minority class.
Step - 5 Train the model with UnderSampled dataset.

C. Oversampling

This is the process of randomly duplicating instances of minority classes. This approach can lead to overfitting and inaccurate predictions of test data.

Following is the Pseudocode of Oversampling:
Step -1 Count the Different number of classes present in the dataset.
Step -2 Divide the dataset by classes.
Step -3 OverSample the minority class.
Step -4 - Concat the UnderSampled class with the minority class.
Step -5 - Train the model with an OverSampled dataset.

D. SMOTE

Smote takes a sample from each minority class and introduces an artificial sample along joining any or every k-nearest neighbors of the minority classes, to generate a synthetic sample. More importantly, this approach effectively forces minority class decision-making domains to become more general. Read this article for a quick explanation. woefully, the above explained technique doesn't work efficiently with text data datasets, as numerical vectors produced from text data are high- dimensional . By using cutting-edge techniques like the Synthetic Minority Oversampling Technique (SMOTE), minority classes may be used to make new instances which are synthetic.

Following is the pseudo code for SMOTE:
Step -1: By calculating the geometrical distance between each sample in set A and each instance of x in minority category set A, one may determine the k-nearest neighbors of x.
Step -2: The rate N is acceptable in light of the imbalanced proportion. The set A 1 is created by selecting N samples (i.e. x1, x2,..., xn) at random from an x's k-nearest neighbors for each x in A. .
Step -3: A replacement example is generated using the following formula for each example x(k) in A 1 (k=1, 2, 3...N):
$x` = x + rand(0, 1) * \mid x - x_k \midx_k \mid$
where the random integer between 0 and 1 is represented by the rand (0, 1) variable.

E. Data Augmentation

This technique is commonly used in computer vision. An image dataset is about transforming (rotating, translating, scaling, adding noise) the images in the dataset to create a new image. For text, data augmentation can be done by breaking the document into sentences, shuffling and reassembling to generate new text, or replacing adjectives, verbs, etc. with synonyms to generate another text with the same meaning. can be done.It can also use pre-trained word embeddings or NLTK's(Natural Language Toolkit) word mesh to find synonyms for words.

F. PSO Learn

We examine learning sampling techniques and associated PSO algorithmic work in this part. Different numbers of minority samples are produced by oversampling methods like ADASYN depending on the distribution. To create minority samples, SMOTE randomly chooses one sample from the closest neighbors and interpolates between them. Therefore, reducing the imbalance ratio can help the issue of overfitting to the majority class. While SMOTE values every sample in the minority class equally, the modeling procedure is likely to misclassify samples that are close to the boundary. In this situation, it is suggested to use Borderline SMOTE, which combines SMOTE with knowledge of the border samples. Experiments showed that the performance of the model may be improved by leveraging the data about samples situated in the border to create additional samples. Recent research on oversampling attempted to integrate SMOTE for big dataset problems into distributed computing settings like Spark. Samples are produced by doing interpolation between minority category samples and their k-minority category class furthest neighbors, according to Gosain's proposed FSMOTE. Mean-SMOTE, a modified version of SMOTE that Binghao developed, was used to classify non-p2p traffic. Hamdy used SMOTE to forecast the severity of specific types of bugs. The Easy-SMT approach, which Wu invented, combines SMOTE-based oversampling strategy with EasyEnsemble to efficiently separate the imbalance issue into the balanced learning subproblem.

To increase the imbalance ratio, undersampling techniques often discard noisy or redundant samples from majority classes. Han, a recent study, used the Gaussian mixture model to account for undersampling. But if we omit samples from the majority class, we frequently lose some information. Noisy data also has a negative impact on how well neighborhood information-based algorithms work. To overcome the oversampling and undersampling constraints, Tomek Links and his ENN are combined. More specifically, Tomek Links purges samples whose closest neighbours fall outside of their own category from the augmented dataset. Smote + ENN [18] uses ENN[18] to predict the label of each sample in the augmented data set. Samples are dropped if the prediction does not match the actual specification. We discovered that sample encroachment, which destroys the local structural data of the majority class sample and interferes with the undersampling process, still occurs in the hybrid sampling method. In order to enhance the data distribution, we must optimize the samples created after oversampling.

The PSO approach, which is based on the behavioral traits of the population which are used to address optimization problems. When utilizing the approach of PSO , one may think of potential solutions to any optimization problem as particles in the search space. Each particle is given a fit value, that is also a velocity that determines the direction and length of its motion. Particles fits the solution space better than the current best particle. The PSO method has previously been applied to non-equilibrium learning in pertinent articles. Use his PSO for sample subset and feature selection in [1] and [2]. Hu recently employed PSO to identify the WELM (Weighted Extreme Learning Machine) parameters' ideal weights. Studies revealed that the -PSO technique might enhance his -WELM's generalization and performance on unbalanced

datasets. This approach improves the imbalanced percentage while maintaining original category distribution by using the PSO algorithm to optimize the sample distribution after SMOTE. [7]

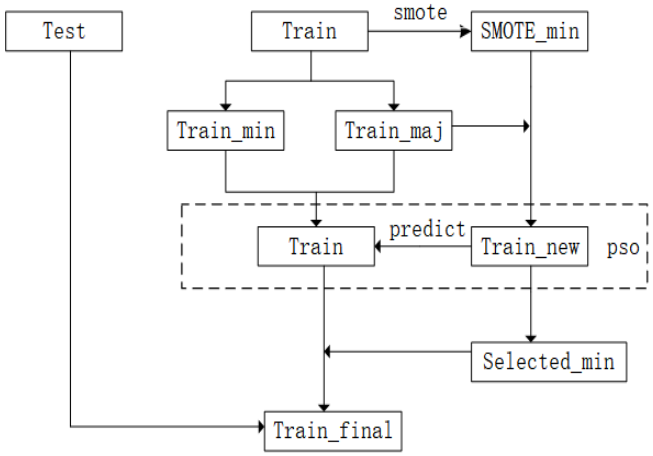

Fig.1: Framework of oversampling method(PSO-optimized).

A. Software Tools

Numpy, scipy, and scikit-learn are just a few of the important libraries in Python that programmers and data scientists use to perform data mining methods. These libraries include built-in features that help with categorization and decision-making. In the past, sampling algorithms had to be coded from scratch in the C++, Java, or C languages. Instead of depending solely on the capabilities of prepackaged functions, this gives programmers more freedom in how they choose to sample data. SSO-PSO majorly flourished in Java, and his SSO-SMOTE-SSO are two recent works that use "code from scratch" approaches. In order to assist categorize unbalanced datasets, the R language has capabilities including training set selected, discretization, and feature selection that are popular among data mining enthusiasts.

Programmers now have less freedom in parameter selection because of the advent of contemporary toolboxes designed specifically for managing unbalanced datasets. One of the open-source Python toolboxes for learning from unbalanced datasets is imbalanced-learn. The oversampling, undersampling, and hybrid sampling methods in this toolkit include SMOTE and Tomek Links. includes cutting-edge methods. A Java-based open-source programme called Knowledge Extraction Based on Evolutionary Learning was developed. It has an evolutionary classification scheme for unbalanced data sets. Open source software for multiclass imbalance datasets is called Multi-imbalance. 18 algorithms are included in this toolkit for multiclass learning under unfair conditions. AdaBoost, DECOC, and DOVO ensemble algorithms intelligently combine sample weights and base classifiers to adapt to multiclass settings. The authors' performance research demonstrates that the best software tools for classifying multi-class unbalanced datasets are DECOC and DOVO. Bi and others suggested DECOC, or Diversified Error Correcting Output Codes. Zhang76. DOVO is a Diversified One-vs-One strategy for multiclass classification proposed by Kang et al.77.DECOC is a type of multiple classifier and ensemble system. A dichotomous classifier converts the codewords into class labels by using the maximum distance between classes to create codewords for various classes. On the other hand, when a subset of the data is most useful, DOVO chooses the classifier with the lowest error rate. Subsets are made by choosing samples from each pair of categories. The strongest classifier outputs from each subgroup are combined to produce the final forecast[7].

4. Results and Discussion

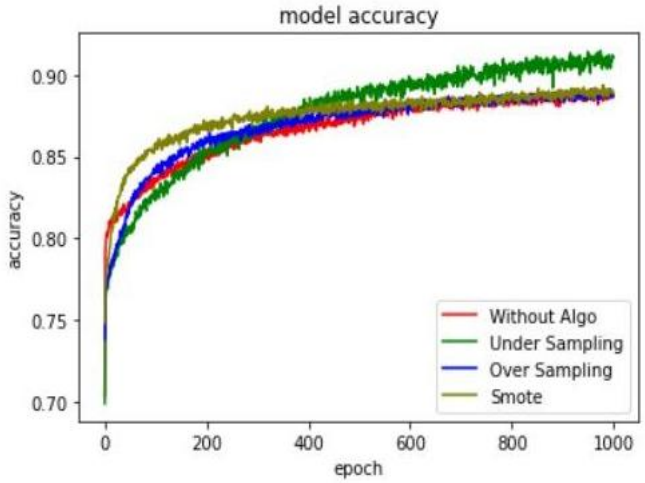

Fig. 2: Graph showing model accuracy of all the methods

The above mentioned graph which is produced using matplotlib shows the accuracy concern of all the codes and approaches which we have discussed.

The pseudo code explained earlier for Undersampling transformed into a program gives accuracy better than earlier without any algo approach. Below shown fig 3 shows the accuracy and other aspects of Undersampling Algorithm.

```
Classification Report using UnderSampling:
              precision    recall  f1-score   support

           0       0.77      0.71      0.74       374
           1       0.73      0.78      0.76       374

    accuracy                           0.75       748
   macro avg       0.75      0.75      0.75       748
weighted avg       0.75      0.75      0.75       748
```

Fig 3a: Undersampling method precision

Table 1: Table showing accuracy of Undersampling method

Classification Report using UnderSampling:				
	precision	recall	f1-score	support
0	0.77	0.71	0.74	374
1	0.73	0.78	0.76	374
accuracy			0.75	748
macro avg	0.75	0.75	0.75	748
weighted avg	0.75	0.75	0.75	748

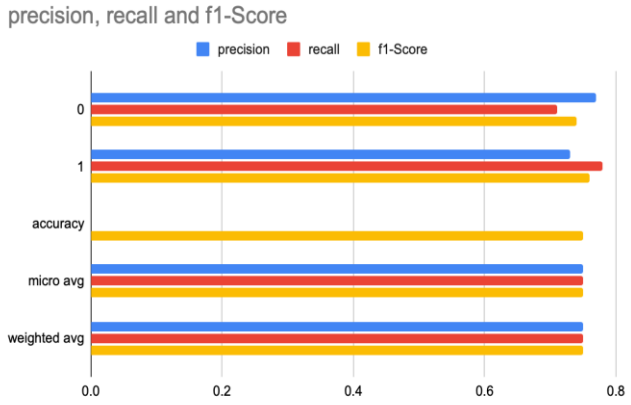

Fig 3b: Graph showing undersampling method accuracy with dataset

The Earlier explained pseudocode for oversampling approach accuracy report with other attributes is shown below in fig 4.

Classification Report using UnderSampling:

	precision	recall	f1-score	support
0	0.77	0.71	0.74	374
1	0.73	0.78	0.76	374
accuracy			0.75	748
macro avg	0.75	0.75	0.75	748
weighted avg	0.75	0.75	0.75	748

Fig 4a: Oversampling method precision

Table 2: Table showing accuracy of Oversampling method

Classification Report using OverSampling:				
	precision	recall	f1-score	support
0	0.86	0.70	0.77	1033
1	0.75	0.88	0.81	1033
accuracy			0.79	2066
macro avg	0.80	0.79	0.79	2066
weighted avg	0.80	0.79	0.79	2066

The accuracy report of SMOTE algorithm approach is shown in fig 5.

Classification Report using Smote:

	precision	recall	f1-score	support
0	0.79	0.82	0.81	1033
1	0.81	0.79	0.80	1033
accuracy			0.80	2066
macro avg	0.80	0.80	0.80	2066
weighted avg	0.80	0.80	0.80	2066

Fig 5a: SMOTE method accuracy

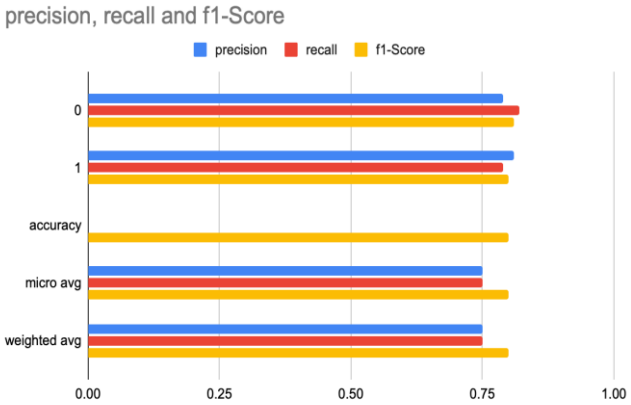

Fig 5b: Graph showing SMOTE method accuracy with dataset.

5. Conclusion & Future Scope

According to the proposed study, there are many challenges and difficulties in broad area of unbalanced learning which needs the attention and active development of the scientific community. There are several untested directions in this vast field of machine learning. In the future, we anticipate that all the issues and challenges raised in this work will be answered, leading to a greater understanding of the phenomenon of learning system imbalance.

References

[1] Krawczyk, B. Learning from imbalanced data: open challenges and future directions.*Prog Artif Intell* 5, pp.**221–232, 2016.** https://doi.org/10.1007/s13748-016-0094-0

[2] S. Sridhar and A. Kalaivani, "A Two Tier Iterative Ensemble Method To Tackle Imbalance In Multiclass Classification," 2020 International Conference on Decision Aid Sciences and Application (DASA), pp.**1248-1254, 2020.** doi: 10.1109/DASA51403.2020.9317019.

[3] Yang P, Yoo P D, Fernando J, *et al*. Sample subset optimization tech- niques for imbalanced and ensemble learning problems in bioinformatics applications, *IEEE Transactions on Cybernetics*, Vol.**44**, no.**3**, pp.**445- 455, 2014.**

[4] Wang K J , Makond B , Chen K H , *et al*. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients, *Applied Soft Computing*, 2014, Vol.**20**, pp.**15-24, 2014.**

[5] Susan, S., & Kumar, A. (2021). The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art. *Engineering Reports*, *3*(4), e**12298, 2021.**

[6] Y. Fathy, M. Jaber and A. Brintrup, "Learning With Imbalanced Data in Smart Manufacturing: A Comparative Analysis," in IEEE Access, Vol.**9**, pp.**2734-2757, 2021.** doi: 10.1109/ACCESS.2020.3047838.

[7] Neshat, M., Sepidnam, G. & Sargolzaei, M. Swallow swarm optimization algorithm: a new method to optimization. *Neural Comput & Applic* 23, pp.**429–454, 2013.** https://doi.org/10.1007/s00521-012-0939-9

[8] Kaur, Harsurinder & Pannu, Husanbir & Malhi, Avleen. (2019). A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Computing Surveys. 52. pp.**1-36, 2019.** 10.1145/3343440.

[9] Krawczyk, B. Learning from imbalanced data: open challenges and future directions.*Prog Artif Intell* 5, pp.**221–232, 2016.** https://doi.org/10.1007/s13748-016-0094-0

[10] W. Obaid and A. B. Nassif, "The Effects of Resampling on Classifying Imbalanced Datasets," 2022 Advances in Science and Engineering Technology International Conferences (ASET), pp.**1-6, 2022.** doi:10.1109/ASET53988.2022.9735021.

[11] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, Amanda Gonsalves, Data imbalance in classification: Experimental evaluation, Information Sciences, Vol.**513**, pp.**429-441, 2020.** ISSN 0020-0255,https://doi.org/10.1016/j.ins.2019.11.004.

[12] Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, *52*(4), pp.**1-36, 2019.**

[13] Goyal, A., Rathore, L., & Kumar, S. (2021). A Survey on Solution of Imbalanced Data Classification Problem Using SMOTE and Extreme Learning Machine. In *Communication and Intelligent Systems*, pp.**31-44, 2021.** Springer, Singapore

[14] Sowah, R. A., Kuditchar, B., Mills, G. A., Acakpovi, A., Twum, R. A., Buah, G., & Agboyi, R. (2021). HCBST: An Efficient Hybrid Sampling Technique for Class Imbalance Problems. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(3), pp.**1-37, 2021.**

[15] Liu, Y., Loh, H. T., & Sun, A. (2009). Imbalanced text classification: A term weighting approach. *Expert systems with Applications*, *36*(1), pp.**690-701, 2009**.

[16] Goyal, S. (2022). Handling class-imbalance with KNN (neighborhood) under-sampling for software defect prediction. *Artificial Intelligence Review*, *55*(3), pp.**2023-2064, 2022.**

[17] Tsai, C. F., Lin, W. C., Hu, Y. H., & Yao, G. T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, *477*, pp.**47-54, 2019.**

[18] Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), pp.**20-29, 2004.**

AUTHORS PROFILE

Shivam Kumar, Pursuing B.tech in specialization of Artificial Intelligence & Machine Learning from Sharda University, Greater Noida. My area of interest is machine learning and java and python development. My hobbies are traveling, reading books and swimming.

Deepanshu Ahuja, Pursuing B.tech in specialization of Artificial Intelligence & Machine Learning from Sharda University, Greater Noida. My area of interest is machine learning and java and python development. My hobbies are traveling, trading and swimming.

Dr. Sandeep Kumar, working as an Associate Professor in department of computer science & engineering at Sharda University, Greater Noida. I do have more than 17 years of academic experience in the field of Computer Science. My research area includes Data Mining, Fractal Graphics and AI. My core subjects are Real Time System, Distributed System, Computer Graphics, Operating System, Computer Organization & Architecture, Artificial Intelligence, and Data Mining & Warehousing. My strengths are my passionate approach and effective planning attitude. And, I am a fitness freak too.

