International Journal of Artificial Intelligence, Machine Learning and Data Science
Volume.6, Number 2 ;April-June, 2022;
ISSN: 2837-3227| Impact Factor: 6.11
https://zapjournals.com/Journals/index.php/ijaimlds
Published By: Zendo Academic Publishing

NAVIGATING ACADEMIC SUCCESS: A COMPARATIVE ANALYSIS OF IMBALANCE CLASSIFICATION METHODS

Prof. Neha Gupta[1], Dr. Rajat Verma[2]

Article Info

Keywords: Imbalanced classification, Class imbalance, Machine learning, Data mining, Big data

Abstract

Imbalanced classification is a critical challenge within the realms of data mining and machine learning, and over the last few years, it has garnered increasing attention from researchers. The conventional approach to classification involves distributing samples evenly across classes to ensure a balanced dataset. However, this practice often leads to unfavorable performance for the majority class. While classifiers are effective in reducing overall classification errors, they tend to exhibit higher error rates when applied to imbalanced datasets, particularly with respect to minority class examples (Barua & Murase, n.d.). In the age of big data, the complexities of imbalanced learning have become more pronounced, and machine learning and data mining have emerged as key tools to address this challenge.

This study delves into the intricacies of imbalanced classification in the context of big data, emphasizing the critical importance of addressing class imbalance for effective predictive modeling. Finding rare events in machine learning and data mining is inherently a prediction task, and the scarcity of such events can severely impede prediction accuracy due to the lack of balanced data (Reference [2]). In the realm of big data, where datasets are vast and intricate, the issue of class imbalance becomes particularly pronounced. This phenomenon is prevalent in various real-world applications, including but not limited to spam detection, software defect prediction, and fraud detection (Reference [3]).

1. INTRODUCTION

Imbalanced classification is an important problem in the domain of data mining and machine learning. Since the last few years, researchers are giving more attention towards the imbalanced data and its classification. Traditional classification techniques are used to perform a balanced sample distribution approach across classes due to which, the majority of the classes perform unfavourably. Generally, the overall classification error can be reduced by the

[1] Computer Engineering & Applications, Lingayas Vidyapeeth, Faridabad, India.
[2] Computer Science & Engineering, B.S. Anagpuria Institute of Technology & Management, Faridabad, India.

use of classifiers but in case of an imbalanced dataset these classifiers would reveal more classification errors with respect to the examples of minority classes (Barua & Murase, n.d.). In the era of big data, the nature of imbalance learning can be better understood by using machine learning and data mining [1]. It has been observed that finding rare events in Machine learning and data mining groups is a prediction task. As rare events are scant in nature, therefore the prediction task suffers due to the absence of balanced data [2]. The structure of larger datasets (big data) is complex and distinctive. As a result, the disparity of the lower classes is a big problem. Spam detection, software defect prediction and fraud detection are very common examples of unbalanced datasets in the real-world [3].

In an online transaction, electronic fraud detection is a significantly challenging problem due to the imbalancing of classes. Fraudsters have done so many attemptations for closely cloning a legitimate transaction to avoid scrutiny. In this era of big data, differentiation of legitimate and illegal transactions is very difficult due to the overlapping of huge amounts of data. Overlapping problems in machine learning-based fraud transaction detection methods have less attention than the imbalanced classification problems [4]. The philosophy behind it is that the imbalanced data is skewed in favour of the instances of majority class with high training accuracy. The solution to this issue is data creation from the minority class that has the best chance of success [5].

1.1 Class Imbalance Problem

In classification, learning classifiers from skewed or unbalanced datasets leads to a serious problem. In this case, the majority of instances belong to one class. It is due to the other class (minor class) which encompasses the more significant characteristics has a lower number of instances. It is observed from the previous research that traditional classifiers generally categorize all data into the majority class and leave the class with the lowest importance which is unsuited to handle imbalanced learning tasks [6].

When some classes are splendidly belittled, statistical and machine learning techniques are inclined to encounter issues. In spite of being learned, cases of the rare classes are lost amongst the others. As a result, unknown rare cases are misclassified by the resulting classifiers and data could be misrepresented by descriptive models. The learning task becomes significantly more difficult if a small class is hardy to identify due to its other features. Other classes may be significantly overlapped by a small class.

In this paper, various techniques are discussed used to handle the imbalanced data sets used in binary as well as in multi classification problems and also anticipate a relative study of the nearly all accepted methods with their advantages and shortcomings. The remaining portions of the paper are as follows: A few important methods of class-imbalance learning (Literature Review) is discussed in section 2. Existing methods are elucidated in section 3. Essential evaluation metrics are explained in section 4. Finally, the conclusion is given in section 5

2. Literature Review

Contemporary research challenges associated with imbalanced data learning that have roots in application areas of the real-world and also investigate various facets of imbalanced learning, like streaming of mining of data, classification, regression, clustering and big data analytics has been discussed by the author in this paper. By doing this, an error-free overview of new challenges in the above-mentioned fields can be estimated [7].

An open-source toolbox of python which is known as imbalanced-learn proposes a large variety of solutions to handle the issues related to imbalanced dataset that usually emerges in machine learning and pattern recognition [8]. To balance the imbalance dataset in an artificial manner, sampling techniques like the synthetic minority oversampling technique (SMOTE) have been used so that this training dataset can be used by classifiers to build the model.

To reconcile the restraints of SMOTE's for nonlinear problems, a weighted kernel-based SMOTE (WK-SMOTE) model that oversamples the feature space of the support vector machine (SVM) classifier is implemented [9]. On the basis of chromosomal theory of inheritance, the MAHAKIL synthetic oversampling method is introduced for imbalanced software defect datasets [10].

To address the class-imbalance issue in the identification of breast cancer, on the basis of sample selection an algorithm named RK-SVM algorithm was proposed Noise and borderline issues are brought on by SMOTE's blind oversampling. Hussein et al., 2019 proposed the A- SMOTE or advanced SMOTE which works on the basis of distance between original minority class samples and the newly introduced minority class examples.

There was no sufficient solution by a high-class imbalance, random under sampling using conventional binary classifiers for the fraud detection problem. A plan was developed for early prediction of turnover intention of new college graduates by the establishment of a predictive model using public data and machine learning [11]. Data of minority class data can be remodelled into a realistic data distribution if it is sparse for **Generative Adversarial Network (GAN)** (Sharma et al., 2022). An effectual model using ensemble classification (Logit Boost + Random Forest) was built to predict academic performance of students at lower secondary level [12].

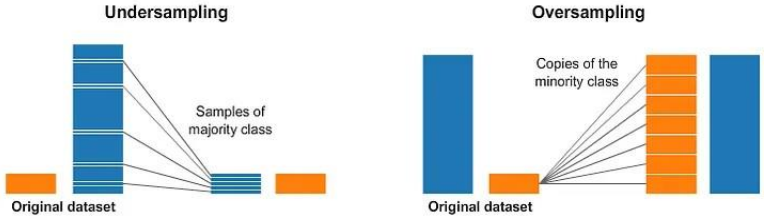
3. Approaches to Handle Imbalanced Dataset

To handle an imbalanced dataset, techniques can be classified into three categories: i) Data level Approach ii) Algorithm level Approach

iii|) Hybrid (Ensemble) Approach

3.1 Data Level Approach

This approach/technique is also known as the resampling method. To make the changes in the training set's distribution, this approach is very useful as it keeps the algorithm's overall structure, including the loss function and optimizer, uninterrupted. This approach is also useful to alter the dataset in order to make popular learning algorithms [13]. Instances of minority and majority classes could be balanced using this method. Two most popular methods of data level approach are under sampling and oversampling techniques of resampling methods. Diagrammetric representation of the resampling method is given in the figure 1.



a) Under sampling Technique b) Oversampling Technique

Fig.1 Resampling Methods

To equalize the count of the majority and minority occurrences, a portion of majority class instances is deleted from the training dataset using under sampling methods. For this, instances of the majority class are removed one by one until the size of the two classes is nearly equal as shown in Fig. 1 (a).

An oversampling method/technique increases the number of instances of minority class until the size of both classes becomes almost equal as shown in Fig. 1(b).

The problem of class imbalance can be reduced up to some extent with the help of some existing approaches like under sampling and oversampling, but each one has its own considerable restraints. Under sampling leads to loss of instances/samples of majority class having meaningful/valuable data while a considerable amount of computational time is entailed by an oversampling method. In a fraud detection model, it is difficult to apply the fusion of these two methods.

Under sampling and Oversampling techniques-based algorithms have their own advantages and disadvantages. So, it is always suggested to use an ensemble/hybrid resampling algorithm which integrates both oversampling and under sampling. It gives you truly accurate results in data processing. There must be a large minimization of imbalancing the samples while diminishing the proportion of majority samples and uprising the number of minority samples,

To make predictions about students at secondary level, an integrated ensemble model comprising features like students' demographic, family, social and academic attributes is developed in this paper. To assess a student at an initial stage, this model is highly beneficial. Out of various models i.e., single, ensemble, and fusion based ensemble classifiers developed in this paper, a model built up with LogitBoost and Random Forest (RF) proves the best model to predict students. [13].

3.2 Algorithm-level Approach

To handle the imbalance classification, a new method is used in this study which utilizes a single-class classifier technique to apprehend the properties of the minority class. An innovative hybrid sampling/boosting method i.e. (RUSBoost algorithm) described by (Seiffert et al., 2010) is used in place of SMOTE Boost for learning from skewed training data. To classify noisy label-imbalanced data, a new technique on the basis of bagging of Xgboost classifiers is proposed by [14]. To combine weighted ensemble classification with a method to handle the issue of class imbalance, Weighted Ensemble with One-Class Classification along with Oversampling and Instance Selection (WECOI) was proposed by [15].

3.3 Hybrid (Ensemble Approach)

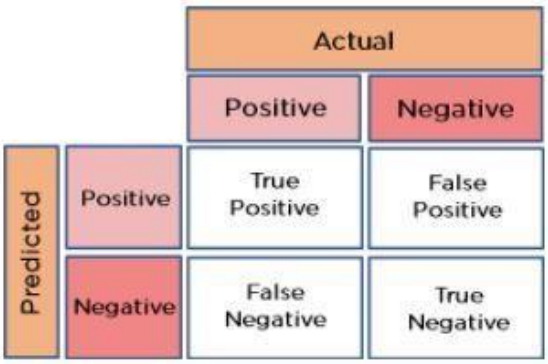
Combination of resampling and ensemble learning techniques are used by hybrid ensemble methods. For this, a comprehensive review of ensemble learning methods was used for imbalanced classification.In ensemble learning, multiple classifiers are integrated to enhance the performance of the model. Since blending is a complex process which takes more time to train the model. Cost-sensitive learning confirms accurate classification of the minority class by various algorithms and also does not affect its computation time and complexity.

The main aim of cost-sensitive learning is to minimize the total cost by computing misclassification cost. Cost-insensitive learning is different from cost-sensitive learning because the former type of learning handles misclassifications uniquely, i.e., the classification cost to predict a sick patient vs healthy is disparate from the classification cost of predicting a healthy patient vs. sick. Thus, the error rate could be minimized and numerous misclassification errors could be neglected by using cost-insensitive learning. Moreover, it is also assumed by cost-insensitive classifiers that all misclassification costs are equal.

4. Evalutaion Metrics

Confusion Matrix is an important parameter to check the performance of a model built for the solution of binary classification as given in Table 4. A negative label (yi=0) is used to mark the majority class while a positive label (yi=1) is used to mark the minority class.

Table 4. Confusion Matrix for binary classification



We used False Positives (FP), False Negatives (FN), Precision (P), Recall (R), Accuracy and F1 Score as base metrics for the purpose of evaluation.

Precision = True Positive/Total Positives = TP/(TP+FP) (1)

Recall=True Positive/(True Positive + False Negative) = TP/(TP+FP) (2)

F1 Score=2*Precision*Recall/(Precision +Recall) (3) Accuracy= (True Positive +True Negative)/Total Values=TP+TN/(TP+FP+TN+FN) (4)

Area under the Precision-Recall Curve is a single statistic that encapsulates the precision recall (PR) curve. It is a useful metric for the prediction's success when the classes are imbalanced. Precision Recall (PR) curves are used as a replacement for receiver operating characteristic (ROC) as it presents the veracious image of performance of an imbalanced dataset.

To make a comparative analysis of various studies, area under the curve (AUC) is used. To get the high score in prediction, each of the four confusion matrix categories (true positive, true negative, false positive, false negative) must be predicted accurately and the same is dependent upon the size of positive and negative values in the dataset.

A special metric Matthews correlation coefficient (MCC) is a statistical statistic used to check the performance of the model. If the value of coefficient is +1, model is in perfect state whereas coefficient=-1 shows that the model is not valid or it is completely failed. A detailed meaning of all metrics for model evaluation is given in Table 2.

Table 2: Evaluation Metrics and their description

Metric	**Description**
Precision	It determines how good the classifier is in detecting fraudulent cases.
Recall	It evaluates the quality of a qualifier.
Accuracy	It measures the efficiency of the algorithm.
F-Measure	It qualifies the quality of a classifier for the rare classes
AUC	It represents the area that exists under a ROC curve
ROC	It is used for evaluating the trade-offs between true positive and false positive error rates in the case of classification algorithms

5. CONCLUSION & FUTURE SCOPE

Most pioneered approaches used to solve the problem of imbalance classification are discussed and evaluated in this paper. Every technique/approach has its own benefits as well as limitations. There are numerous methods such as data level, algorithm-level, hybrid learning, context sensitive learning, deep learning etc. Out of all these methods, data-level methods such as oversampling, under sampling, and hybrids are used on a training set of data. By using under sampling techniques, there is a problem of underfitting as this technique incurs the loss of information whereas there is an issue of model overfitting by using an oversampling approach.

Although hybrid approaches are more effective than resampling, their computation cost is very high and they are also complex in nature. Apart from these, one-class learning and ensemble learning can be used at the classifier level (Bagging and Boosting algorithms).

In the future, advanced techniques such as deep learning and cost-sensitive learning techniques can be used to handle issues of class imbalances in more complex and big datasets. To measure the accuracy and performance of the model, numerous evaluation metrics could be used.

6. REFERENCES

Krawczyk B., "Learning from imbalanced data: open challenges and future directions", Progress in Artificial Intelligence, Vol. 5, pp. 221-232, 2016.

Haxiang G., "Learning from class-imbalanced data: Review of methods and applications", Expert System with Applications,Vol. 73, pp.220-239, Dec. 2016

Huda M., Ahmed R., Siregar M and Maseleno A., "Big Data Emerging Technology: Insight into innovative environment for online learning resources", International Journal of Emerging Technologies in Learning, Vol.13, pp.23-36, Jan. 2018.

Kanika, Singla J., Bashir A.K. and Tariq U., "Handling class imbalance in online fraud detection",Computers, Materials and Continua, Vol.70, pp.2861-2877, Jan.2022.

Desuky A.S. and Hussain S., "An improved hybrid approach for handling class imbalance problem",Arabian Journal for science and engineering, vol46, pp.3853-3864, Jan. 2021

Kostopoulos G., Grawains G. and Kotsiantis S., "Predicting student performance in distance higher education using active learning", Engineering Applications of Neural Networks, vol. 744, pp.75-86, Aug. 2017

Husaini Y.A. and Shukor N.S.A., "Prediction methods on student's academic performance: A review", Jilin Daxue Xuebao/Journal of Jilin University (Engineering and Technology), Vol.41, pp. 196-217, Sep. 2022.

Mathew J., Pang C.K., Luo M. and Leong W.H. "Classification of imbalanced data by oversampling in kernel space of support vector machine", IEEE transactions on Neural Networks and Learning Systems, Vol.99, pp.1-12, Oct.2017.

Bennin K.E., Keung J., Monden A. and Mensah S., "MHAKIL Diversity based oversampling approach to alleviate the class imbalance issue in software detection prediction", IEEE Transactions on Software Engineering, vol.44, pp.534-550, July 2017.

Park J., Kwon S. and Jeong S.P., "A study on improving intention forecasting by solving imbalance data problems: focusing on smote and generative adversarial networks",Journal of Big Data, Vol.36, 2023

Jalota C. "An effectual model for early prediction of academic performance using ensemble classification", Journal of language and linguistics in Society, vol.3, pp.19-33, Mar.2023.

Krawczyk B., "Learning from imbalanced data: open challenges and future directions", Progress in Artificial Intelligence, Vol. 5, pp. 221-232, 2016.

Ruisen L, Songyi D., Chen W., Peng C., Zuodong T., Yanmei Yand Shixiong W., "Bagging of Xgboost classifiers with random undersampling and Tomek Link for Noisy Label-imbalanced data", IOP Conf. Series: Material Science and Engineering, vol.428, 2004.

Czarnowski I., "Weighted Ensemble with one-class classification, oversampling and instance selection (WECOI): An approach for learning from imbalanced data streams", Journal of Computational Science, vol.61, 2022.

Krawczyk B., "Learning from imbalanced data: open challenges and future directions", Progress in Artificial Intelligence, Vol. 5, pp. 221-232, 2016.