International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 10 ǁ October 2022 ǁ PP. 34-40

Handling The Class Imbalance Problem From Different Datasets

Subhasis Misra[1], Rejuwan Shamim[2]and Ratul Dey[3]
[1]*University of Engineering and Management, Jaipur, 303807, India*
[2]*Maharishi University of Information Technology, Noida, 201304, India*
[3]*University of Engineering and Management, Jaipur, 303807, India*
[3] *Research scholar atJaipur National University, Jaipur,India*

Abstract:
Class imbalance is one of the problems in machine learning, deep learning, big data, and data mining. Imbalanced datasets distort the representation of data mining and machine learning strategies because overall accuracy and decision-making are skewed in the majority class, making tests for minority classes misclassified or unexciting. may be considered to exist. This paper proposes an overall review of class imbalance issue arrangements and the main examinations as of late presented by specialists. In this paper, we also discussed the different algorithms used to solve the problem of imbalanced data.

Keywords: *Data mining, Deep learning, Big data, Imbalanced data, Minority class, Majority class.*

Date of Submission: 20-09-2022 Date of acceptance: 04-10-2022

I. Introduction

Over the past few decades, an enormous amount of data is produced and those exponential increase in raw data has been used to train model in the fields of Machine learning, Artificial intelligence, Deep Learning, Big data, and Data Mining. The datasets are either balanced or imbalanced. Nowadays class imbalance has become one of the major problems in the world of technology. The specific condition of an imbalanced dataset, where each class related to a given dataset is conveyed unevenly. The standard way to deal with settling this kind of issue regularly yields unseemly outcomes.

The term "class imbalance" refers to the point where one class's perception is higher than the other class's perception and there is class inequality. The issue of class imbalance is very important. Because it occurs in a variety of fields of incredible climate, imperative, or business importance, and in some cases, major bottlenecks to the performance achievable with class-balanced, traditional methods of learning. because it has been shown to produce.

The standard algorithms that are used in data mining are not typically ready to manage these enormous datasets. As such, classification algorithms should be modified and adjusted considering the arrangements that are being used in big data so they can be used under these circumstances maintaining their predictive capacity. One of the complexities that make troublesome the extraction of helpful information from datasets is the issue of classification with imbalanced data. This issue happens when the number of instances of one class (positive or minority class) is impressively more modest than the number of instances that have a place with different classes (negative or greater part classes). In the present circumstance, the interest in learning is focused on the minority class as the class should be precisely recognized in these issues. Big data is likewise sensitized by this unseen class distribution.

The data is also used by machine learning, incorporating various algorithms to isolate valuable information and models for classroom noise detection, image processing, sentiment analysis, signal processing, traffic accident analysis, social data mining, etc. train the If the data are imbalanced, the minority class will have significantly fewer examples than the majority class, so the tests will be unevenly distributed across the selected classes. As such, dealing with imbalanced datasets is a difficult task in machine learning. Data is also used in machine learning, incorporating various algorithms to isolate valuable information and develop models for classroom noise detection, image processing, sentiment analysis, signal processing, traffic accident analysis, social data mining, and more. training. If the data are imbalanced, the minority class will have significantly fewer examples than the majority class, so the tests will be unevenly distributed across the selected classes. As such, dealing with imbalanced datasets is a difficult task in machine learning.

Over the past decade, deep learning techniques have grown in popularity as they have worked on cutting-edge technologies in speech recognition, computer vision, and other fields. Their new achievements can be attributed to improved data access, hardware, and software upgrades, and leaps in various algorithms that speed up training and advance inferences on new data. Despite these advances, little statistical work has been done exploiting techniques to address class imbalance using deep learning and corresponding architectures. B. A properly rated deep brain network (DNN). Many researchers agree that the topic of deep learning with class imbalance data is understudied.

Data mining: Data mining is the process of analyzing huge data to find out business intelligence that can help companies to solve problems, and seize new opportunities.

Deep Learning: Deep learning is a machine learning method that trains PCs to do what falls into place without any issues for people.

Big data: Big Data is an assortment of data that is gigantic in volume, yet developing exponentially with time.

Imbalanced data: Imbalanced data alludes to those where one class mark has an exceptionally high number of observations and different has an extremely low number of observations.

Majority class: It is just that having the best recurrence in the class distribution of training.

Minority class: The majority class is basically that has the most minimal recurrence in the class distribution of training.

1.1 Algorithm for dealingwith imbalanced datasets.

a. Resampling techniques method:

To manage imbalanced datasets, researchers have introduced various resampling methods. One of the advantages of using these techniques is that they are external approaches that use existing algorithms. If both undersampling and oversampling occur, they can effectively move. Some of the known resampling strategies look like this:

Random Oversampling:Oversampling aims to increase the number of minority classes in the training set. Random oversampling is a simple way to handle resampling by randomly choosing people from minority classes. These randomly selected people are then cloned and added to the new training set.

Random Undersampling: Undersampling is a cycle aimed at reducing the amount of individuals in the majority class in the training set. Random undersampling is a well-known resampling technique that randomly removes documents in the majority class in the training set until the minority class to majority class ratio is at a desired level.

b. Synthetic technique:

The synthetic minority oversampling technique (SMOTE) is probably the best-known method of creating synthetic samples. SMOTE is a validation approach in which minority classes are validated by creating synthetic models rather than by substitution.

It is fundamentally a blend of oversampling the minority (strange) class and undersampling the larger part (ordinary) class, which is found to accomplish better classifier execution than just undersampling the larger part class. In this strategy, synthetic models are produced in a less application-explicit way, by working in the component space as opposed to the information space.

c. **Cost-sensitive algorithms**: Cost-Sensitive Learning is a sort of discovery that thinks about misclassification or different kinds of costs. Cost-sensitive learning is a famous and normal way to deal with addressing class imbalanced datasets. Famous AI libraries, for example, random forest, support vector machines (SVM), logistic regression, decision trees, among others, can be designed utilizing cost-sensitive preparation.

d. **Random Forest**: The essential thought is the same as stowing as in we bootstrap tests, so we take a resample of our preparation data set. And afterward, we reconstruct arrangement or regression trees on every one of those bootstrap tests. The one distinction is that at each split when we split the data each time in a grouping tree, we likewise bootstrap the factors. All in all, main a subset of the factors is considered at every expected split. This makes for a different arrangement of potential trees that can be constructed.Thus the thought is we grow an enormous number of trees. For prediction, we either vote or normalize those trees to get the prediction for another result.

e. **Gradient Boosting**: Unlike the random forest strategy that forms and consolidates a forest of randomly various trees equal, the critical thought of gradient helped choice trees is that they fabricate a progression of trees. Where each tree is prepared so it endeavors to address the errors of the past in the series. Inherent a non-random way, to make a model that commits increasingly few errors as more trees are added. When the model is assembled, making predictions with gradient-supported tree models is quick and doesn't utilize a ton of memory.

f. **Ada Boost:** It is an iterative gathering technique. Ada Boost classifier constructs a solid classifier by joining numerous inadequately performing classifiers to get high accuracy solid classifier. The essential idea driving Ada boost is to set loads of classifiers and train the information test in every cycle to such an extent that it guarantees the exact expectations of uncommon perceptions.

II. Related work:

The authors explain that the purpose of the review paper is to summarize current research on high-class imbalance problems in big data. The scope of the review examines research directed to the last eight years (i.e. 2010-2018) and highlights the intersection of big data and class imbalance problems and the resulting solutions developed by researchers. doing. Additionally, due to legitimate concerns about our emphasis on big data, this paper subdivides one or more datasets (class imbalances in big data) of over 100,000 instances. only related work was considered. [21]
The author describes a work containing two case studies. The dataset used in the first case study is from a different application domain than the dataset used in the second case study. In the first case study, cross-validation (CV) was performed against theMedicare dataset. The second case test used the Slowloris-Big dataset for preparation and the POST dataset for testing. The Medicare dataset is considered high-dimensional (102 features), while the SlowlorisBig and POST datasets are not (11 and 13 features respectively). [1]
The authors describe class imbalanced datasets that occur in real applications where the class distribution of information is highly imbalanced. Again, without losing consensus, we accept that the minority or rare class is a positive class and the majority class is the negative class. The minority class is often very small, say his 1% of the dataset. The basic idea in creating an immediate cost-sensitive learning computation is to directly inject and use the misclassification cost in the learning algorithm. ICET (Turney, 1995) and cost-sensitive decision trees (Ling et al.

The author describes the paper and proposes a general overview of class imbalance problem arrangements and the main investigations as of late presented by researchers. lass imbalance is an interesting issue being investigated as of late by machine learning and data mining researchers. The researchers for taking care of the imbalance problem have proposed various approaches. Be that as it may, there is no general approach legitimate for all imbalanced data sets and there is no unification framework. This paper summarizes various answers for dealing with class imbalance problems.[3]

The author describes they have utilized a SMOTE oversampling and cluster-based under-sampling procedure to adjust cardiovascular data and look at the outcomes. The conventional over-sampling and under-sampling strategies may not forever be appropriate for such datasets. The proposed technique is viewed as valuable for such datasets where the class names are unsure and can likewise assist with defeating the class imbalance issue of clinical datasets and additionally for other data domains.[4]

The author describes the goal is to limit the cost of misclassification, which can be realized by picking the class with the base restrictive risk. Cost-sensitive learning methods attempt to boost a loss function related to a data set. These learning methods are roused by the observation that most genuine applications don't have uniform costs for misclassifications. The real costs related to every sort of blunder are obscure ordinarily, so these methods need to decide the cost matrix in view of the data and apply that to the learning stage. A closely related plan for cost-sensitive students is shifting the inclination of a machine to lean toward the minority class.[5]

The author describes that they removed unnecessary properties so our proposed classification strategy will be more precise. In this data pre-handling, the crude data downloaded from the GOE dataset were log2 standardized utilizing the "biobased" R bundle, eliminated group impacts and undesirable variety utilizing the "affy" bundle, and looked at genuinely or dissected for differential expression utilizing the "Limma" bundle. Subsequent to eliminating cluster impacts in the data pre-handling, 20,663 probes were chosen out of 54,675 probes. We then, at that point, recognized 825 probes that were essentially different with a p-esteem < 0.0001 in COPD subjects contrasted with a sound non-smoker subject.[6]

In the paper, there are by and large two systems to deal with class imbalance classification; 1) data-level approach and 2) algorithm-level approach. The techniques at the data-level approach change the class imbalance proportion with the target to accomplish an equilibrium of the dispersion between classes though, at the algorithm-level approach, the customary classification algorithms are adjusted to further develop the learning task particularly comparative with the more modest class.[7]

The paper says that they acquaint with a modified hybrid strategy to deal with this issue. They utilize mimicked annealing to pick the most ideal subset of significant class records (columns of data). A while later, KNN, DA, SVM, and DT classifiers are used to evaluate the proficiency of our method. They assess our exact outcomes with the two ongoing works. They investigate 51 real datasets from various data repositories for the tests. 24 datasets were utilized. Out of 24 datasets, the strategy out-plays out the technique proposed in 14 datasets and yields comparable performance with the remainder of the datasets. Their approach demonstrates its viability and thus can be applied in real-world settings where the dataset is imbalanced. The proposed procedure is further approved with the introduced technique as far as AUC and G-mean. Our procedure showcased prevalence in 17 datasets while the RCSMOTE strategy yielded better outcomes just in 5 datasets out of 28 datasets.[8]

The paper says it has been shown that traditional AI strategies for handling class imbalance can be stretched out to deep learning models with progress. The survey additionally observes that virtually all exploration in this space has been centered around computer vision assignments with CNNs.[9]

The author describes that they have zeroed in on the AM performance measure, and have shown that under specific conditions, a few straightforward algorithms, for example, module rules with an exact edge and experimentally adjusted ERM algorithms are AM-consistent. Our second objective was to assess the performance of the class imbalance algorithms examined here on a wide scope of genuine information. The author utilized 17 informational collections with differing levels of class imbalance, taken from the UCI rest conservative (Frank and Asun-cion, 2010) and other sources; because of space constraints.[22]

The paper says their methodology demonstrates its viability and subsequently can be applied in true settings where the dataset is imbalanced. The proposed procedure is additionally approved with the introduced technique as far as AUC and G-mean. Our procedure displayed prevalence in 17 datasets while the RCSMOTE technique yielded better outcomes only in 5 datasets out of 28 datasets.[10]

The author says a short survey on the functioning principles of the state-of-the-art it is introduced to undersampling techniques. Taxonomy has been characterized as methodically arranging the proposed techniques and then, comparative analyses of the techniques are directed. In view of the analyses, a couple of theoretical hypotheses are long to highlight the conceivable prospectives of advanced and smart research to plan powerful undersampling techniques for the treatment of class imbalance problems.[11]

Researchers examined different oversampling strategies, such as SMOTE, ADASYN, borderline SMOTE, and safe-level SMOTE, to process imbalanced data sets with different expectation models. Three unique classification techniques are used to evaluate different measures of performance: Naive Bayes, Support Vector Machine, and the northernmost neighbors of the six datasets. In this paper, we use six real datasets represented by the total number of instances in the dataset, the number of major instances, the number of minority instances, the imbalance ratio (IR), and the number of attributes used in the datasets. have considered [12].
The author describes the problem of imbalanced data, and the need to balance the data is discussed in complex ways. We also discussed the different kinds of methods developed by different authors to deal with the problem of imbalance. It is clear from this survey that the majority of developers indicate that the SMOTE algorithm outperforms state-of-the-art algorithms on the class imbalance problem. Imbalance problems exist in many real-world domains, such as medical diagnostics, fraudulent call detection, and the telecommunications sector. Despite the fact that many methods are available to treat imbalance problems in various fields, the field of medical diagnostics still needs a great deal of attention. It suggests that significant improvements in technology are needed to rationalize the imbalance problem [13].

The paper says they have given a probabilistic understanding of the impacts class imbalance has on discriminative models. We ran recreation analyses to verify this hypothesis. In this translation, they exhibited the situations in which observational error-minimizing (linear) classifiers instigated over imbalanced datasets will probably initiate a one-sided separator. Moreover, they evaluated the circumstances when weighted observational expense techniques for alleviating the impacts of imbalance, for example, weighted-SVM and SMOTE, will probably neglect to improve performance.[14]

The author says they have chosen a dataset from Kaggle. This dataset comprises 43431 rows (occasions) with 9 credits (highlights). Each column addresses a web-based vehicle booking done or dropped. The characteristics included are generally mathematical traits. [15]

The paper says a few classifiers that are every now and again utilized for high-dimensional information are exceptionally delicate to the class-imbalance problem and that the problem is exacerbated when information is high-dimensional.[16]

The author says learning from imbalanced informational collections is a significant issue in machine learning. An immediate strategy to tackle the imbalance problem is falsely adjusting the class distributions, and its viability has been empirical.[17]

The paper describes that depending on the examination that has been done, it very well may be presumed that handling class awkward nature in the dataset is significant, particularly in the classification of machine learning. In view of the tests that have been done using the proposed technique. The stacking algorithm between the SVM algorithm and the Random Forest algorithm and the expansion of ADASYN in the resampling system on 5 datasets with various unevenness proportions shows that situation 2 creates preferable execution over situation 1 on one or the other side. the g-mean worth and the AUC esteem notwithstanding the exactness esteem on a few datasets additionally increased. It very well may be reasoned that handling class irregularity in the dataset using two approaches can be an answer to solving the class lopsidedness issue in the dataset.[18]

The paper is concentrated on whether and how class imbalance learning can facil-itate SDP. We investigated five class imbalance learning strategies, covering three sorts (undersampling, threshold-moving, and Boosting-based ensembles), in comparison with the two highest level indicators (Naive Bayes, and Random Forest) in the SDP literature. They have evaluated ten real-world SDP data sets with a wide range of data sizes and imbalance rates. To guarantee that the outcomes introduced in this paper are of practical value, five performance measures were considered, including PD, PF, balance, G-mean, and AUC.[19]

The author brief that they center around two-class imbalanced informational in-dexes, where there is a positive (minority) class, with the most reduced number of occasions, and a negative (majority) class, with the largest number of occurrences. We likewise think about the imbalance ratio (IR), characterized as the number of negative class models that are isolated by the number of positive class models, to sort out the various informational indexes.[20]

Comparative Analysis:

Table 1. Merits and Demerits of different algorithms.

Algorithm	Benefit	**Demerit**
Random Oversampling	It doesn't lose the information.	It improves the probability of overfitting since it replicates the minority class occasions
Random Undersampling	It can assist with working on show time and storage problems to reduce the quantity of training data samples while the training data set is immense.	It can dispose of possibly valuable information which could be significant for building rule classifiers**.**
Synthetic technique:	No deficiency of information	SMOTE is not very practical for high dimensional data
Random Forest	Random Forest can be utilized to settle both classifications as well as regression problems**.**	**:** Random Forest requires substantially more chance to train when contrasted with decision trees as it creates a lot of trees settles on decisions on most of the votes
Gradient Boosting	Often gives prescient accuracy that can't be trumped	Gradient Boosting Models will keep improving to limit all errors. This can overemphasize outliers and cause overfitting
ADA Boost	The precision of weak classifiers can be improved by utilizing Adaboost.	It needs a quality dataset.
RUSBoost	RUSBoost is an algorithm to deal with class imbalance issues in information with discrete class labels. It utilizes a combination of RUS (arbitrary under-sampling) and the standard boosting strategy. RUSBoost is computationally more affordable than SMOTEBoost and results in significantly more limited model training times.	The main downside of RUS, which is the deficiency of information, is extraordinarily overcome by combining it with boosting. While certain information might be missing during a given iteration of boosting, it will probably be included while training models during different iterations.
SMOTEBagging	SMOTEBagging is to create each bag. It makes use of both oversampling and undersampling techniques;	SMOTEBagging algorithm was decided from several parameters, which were the number of bootstrap k-nearest neighbors and the total number of oversampling
MSMOTEBoost	MSMOTEB is a data preprocessing algorithm.	There is a huge problem in bagging and boosting
SMOTEBoost	1.Alleviates overfitting caused by random oversampling as synthetic examples are generated rather than a replication of instances 2.No loss of information	1. While generating synthetic examples, SMOTE does not take into consideration neighboring examples can be from other classes. This can increase the overlapping of classes and can introduce additional noise. 2. SMOTE is not very practical for high-dimensional data.

MSMOTE	SMOTE for learning from imbalanced datasets, based on the SMOTE algorithm. MSMOTE not only considers the distribution of minority class samples, but also eliminates noise samples through adaptive mediation.	SMOTE does not account for important characteristic differences. This is because functionality is decisive for model performance. Future work is needed to address this issue.
SPIDER	No need to define the number of bags	The algorithm stops when the outside-the-bag error estimates stop decreasing.
RareBoost	It tackles the class imbalance problem by simply changing αt’s computation	This constraint is not trivial when dealing with the class imbalance problem.

The above table (Table 1) describesthe working principle of the different algorithms.After all the analysis we find that not a single algorithm gives accuracy for all the data sets. In real world, data are linear or non-linear or categorical, but one single algorithm can’t give an accurate result for all the datasets.

III. Conclusion:

Class imbalance is an intriguing issue being explored as of late by machine learningand data mining specialists. The analysts for tackling the imbalance issue have proposed different methodologies. In any case, there is no broad methodology appropriate for all imbalanced data sets and there is no unification framework.

When confronted with imbalanced data sets there is nobody stops answering for work on the exactness of the forecast model. One might have to evaluate numerous strategies to sort out the most ideal examining methods for the dataset. Much of the time, manufactured strategies like SMOTE will beat the regular oversampling and undersampling techniques.

One of the high-level stowing strategies normally used to counter the imbalanced dataset issue is SMOTE sacking. It follows a totally unique methodology from customary sacking to make each Bootstrap. It produces the positive occurrences by the SMOTE Algorithm by setting a SMOTE resampling rate in every iteration. The arrangement of negative cases is bootstrapped in every iteration. Contingent upon the qualities of the imbalanced data set, the best strategies will shift. Applicable assessment parameters ought to be considered during the model correlation.

This paper summarizes various responses to address class imbalance. This white paper provides an overview of the problem of dataset class imbalance and the problems that come with it.

Reference:

[1]. Hasanin, Tawfiq, et al. "Severely imbalanced big data challenges: investigating data sampling approaches." Journal of Big Data 6.1 (2019): 1-25.

[2]. Ling, Charles X., and Victor S. Sheng. "Cost-sensitive learning and the class imbalance problem." Encyclopedia of machine learning 2011 (2008): 231-235.

[3]. Abd Elrahman, Shaza M., and Ajith Abraham. "A review of class imbalance problem." Journal of Network and Innovative Computing 1.2013 (2013): 332-340.

[4]. Rahman, M. Mostafizur, and Darryl N. Davis. "Addressing the class imbalance problem in medical datasets." International Journal of Machine Learning and Computing 3.2 (2013): 224.

[5]. Longadge, Rushi, and Snehalata Dongre. "Class imbalance problem in data mining review." arXiv preprint arXiv:1305.1707 (2013).

[6]. Mahmudah, Kunti Robiatul, et al. "Machine Learning Algorithms for Predicting Chronic Obstructive Pulmonary Disease from Gene Expression Data with Class Imbalance." BIOINFORMATICS. 2021.

[7]. Ali, Aida, Siti Mariyam Shamsuddin, and Anca L. Ralescu. "Classification with class imbalance problem." Int. J. Advance Soft Compu. Appl 5.3 (2013).

[8]. Desuky, Abeer S., and Sadiq Hussain. "An improved hybrid approach for handling class imbalance problem." Arabian Journal for Science and Engineering 46.4 (2021): 3853-3864.

[9]. 9.Johnson, Justin M., and Taghi M. Khoshgoftaar. "Survey on deep learning with class imbalance." Journal of Big Data 6.1 (2019): 1-54.

[10]. Wang, Shujuan, et al. "Research on expansion and classification of imbalanced data based on SMOTE algorithm." Scientific reports 11.1 (2021): 1-11.

[11]. Devi, Debashree, Saroj K. Biswas, and Biswajit Purkayastha. "A review on solution to class imbalance problem: Undersampling approaches." 2020 International Conference on Computational Performance Evaluation (ComPE). IEEE, 2020.

[12]. Gosain, Anjana, and Saanchi Sardana. "Handling class imbalance problem using oversampling techniques: A review." 2017 international conference on advances in computing, communications and informatics (ICACCI). IEEE, 2017.

[13]. Spelmen, Vimalraj S., and R. Porkodi. "A review on handling imbalanced data." 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018.

[14]. Wallace, Byron C., et al. "Class imbalance, redux." 2011 IEEE 11th international conference on data mining. Ieee, 2011.

[15]. Shukla, Pratyusha, and Kiran Bhowmick. "To improve classification of imbalanced datasets." 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, 2017.

[16]. Lusa, Lara. "The class-imbalance problem for high-dimensional class prediction." 2012 11th International Conference on Machine Learning and Applications. Vol. 2. IEEE, 2012.

[17]. Guo, Xinjian, et al. "On the class imbalance problem." 2008 Fourth international conference on natural computation. Vol. 4. IEEE, 2008.

[18]. Pristyanto, Yoga, et al. "Dual Approach to Handling Imbalanced Class in Datasets Using Oversampling and Ensemble Learning Techniques." 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM). IEEE, 2021.

[19]. Galar, Mikel, et al. "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 42.4 (2011): 463-484.
[20]. Wang, Shuo, and Xin Yao. "Using class imbalance learning for software defect prediction." IEEE Transactions on Reliability 62.2 (2013): 434-443.
[21]. Leevy, Joffrey L., et al. "A survey on addressing the high-class imbalance in big data." Journal of Big Data 5.1 (2018): 1-30.
[22]. Menon, Aditya, et al. "On the statistical consistency of algorithms for binary classification under class imbalance." International Conference on Machine Learning. PMLR, 2013.