IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 2, April, 2023, pp. 517-524
Journal homepage: http://iieta.org/journals/ria

A Study on Imbalanced Data Classification for Various Applications

Kunda Suresh Babu, Yamarthi Narasimha Rao*

School of Computer Science and Engineering, VIT-AP University, Amaravathi 522237, Andhra Pradesh, India

Corresponding Author Email: y.narasimharao@vitap.ac.in

https://doi.org/10.18280/ria.370229

Received: 10 January 2023
Accepted: 1 April 2023

Keywords:
deep learning, imbalance data, raw primary data, classification, machine learning, superior accuracy, misclassified

ABSTRACT

In today's world, classification issues with unbalanced data are widespread. The Aim is to solve the issue of low classification learning algorithm accuracy in diverse applications due to a major imbalance of the sample set. In fields including marketing, medical science, information security, and computer vision. Raw primary data is frequently distorted due to a skewed perspective of the data distribution of one class over another. These issues have a negative impact on the categorization process in algorithm development, machine learning, and deep learning. There are classifications with different ratios of specimens in some circumstances, with one class having a large number of specimens and the other having fewer specimens. The latter class is an essential one, yet many classifiers misclassify it. Recent research on unbalanced problems in numerous areas from 2020 to 2021 is in this survey report. Extensive research has been conducted to handle unbalanced data issues utilizing a variety of techniques and approaches. The experimental findings reveal that ADASYN obtains the highest level of accuracy in intrusion detection.

1. INTRODUCTION

When the data distribution between the studied classes is unbalanced, i.e., one of the classes has a large number of samples compared with other classes is referred to as the majority class which leads to the Class Imbalance Problem (CIP) [1]. In prediction/classification tasks, the CIP is particularly unsuitable, while most existing algorithms will prioritize categorization of the dominant class even as avoiding or mislabeling under-represented observations. Because the program's predicting abilities can damage by class imbalances and the overall classification accuracy is strived by the classification algorithm [2]. Many real-world situations, like biology, facial recognition, text mining, anomaly detection, sentimental analysis, physical science, and so on, have a class imbalance problem [3].

To resolve the difficulties of problematic categorization when working with imbalanced data sets, researchers have provided various techniques to improve the algorithm level and data level. At the data level, the most common approach to balance the dataset is eliminating the samples from major classes and increasing the samples in minor classes [4-6]. The integration and cost-sensitive learning approaches are used at the algorithm level largely to improve the core algorithm.

One of the primary ways is rebalancing class distribution which scholars examine when going to deal with a categorization of unbalanced data [7]. Both modern and classic research strategies are designed to generate a more efficient version of the training data while avoiding the problem of class overlap. Certain techniques work with samples in the duplicating zone, specifically those in the borderline areas; nonetheless, resampling rates take control of the degree of class imbalance. As a result, instead of class imbalance, class overlap can have a major effect on outcomes in specific scenarios [8].

Random Under-Sampling (RUS) and Random Over-Sampling (ROS) are approaches for correcting the class imbalance that involves copying or removing samples till a sample size per class equilibrium is reached [9]. The most used technique is Synthetic Minority Over-sampling Technique (SMOTE) and it provides unique synthetic samples and minority class samples overlapped by unique synthetic samples. Other sample approaches that are mainly based on this technique include Adaptive Synthetic Sampling (ADASYN), SMOTE, and borderline-SMOTE altering nearest neighbor.

On either side, RUS has been described as the most effective under-sampling method. Some of these techniques are identified by a heuristic mechanism that aims to eliminate or modify noisy, irregular, or redundant sample labels [10]. One-sided Selection procedures and Neighborhood Clearing Rules, for instance, can be suggested.

Various classification approaches for imbalanced data sets are extensively discussed in this research. The following are the paper's main contributions:

- From 2020 to 2021, we examine current research on imbalance concerns in several domains in this survey report.
- Intrusion detection, medical field, fraud detection, and sentiment analysis are the four primary groups of imbalance concerns.
- Based on their accuracy, we evaluated the performance of the imbalanced data categorization algorithms.

The following is how the rest of the study is done: The unbalanced data classification in sentiment analysis is discussed in Section 2. Unbalanced data classification in the medical profession and intrusion detection were examined in sections 3 and 4. Unbalanced data classification in fraud detection is described in Section 5. The findings are then

examined in Section 6, followed by a discussion of future work in Section 7, and finally the conclusion.

2. DATA BALANCING APPROACHES IN THE SENTIMENT ANALYSIS

To detect sentiment in OS log messages, Gated Recurrent Unit (GRU) networks were suggested by Studiawan et al. [11]. They use the Tomek link method to solve the issue of an unbalanced dataset. It contains a large number of positive messages compared to negative messages which provide a data imbalance problem. Part-of-Speech based Transformer Attention Network (pos-TAN) presents by Cheng et al. [12]. The sentiment categorization method employed the loss function named Focal Loss. This technique is used to reduce the samples from large classes and provides balanced data.

For textual sentiment analysis, a scalable multi-channel dilated combination of bidirectional long short-term memory (CNN–BiLSTM) and architecture of convolutional neural networks was proposed by Gan et al. [13]. To resolve the problem of imbalance, an adaptive weighted loss function was created. Training category weights decrease the influence of category unbalance in the training phase, whereas analysis category weights cause the strategy to focus more on complicated classes in the later training process, according to the adaptive weighted loss function.

For document-level sentiment classification, a Multi-Topic Bidirectional LSTM (MT-BiLSTM) framework in an email was proposed by Liu et al. [14]. Data augmentation is used to reduce the impact of imbalance.

Budhi et al. [15] use machine learning classifiers and multiple preprocessing and textual-based feature methods to construct a fake review detection system. Random sampling approaches are employed to alleviate the imbalance issue. Xia [16] presents support vector machines (SVMs) and conditional random fields (CRFs) for categorizing feelings conveyed in online reviews. Ruz et al. [17] propose employing Bayesian networks to find the relationships between words for sentiment classification on Twitter. SMOTE is employed to find the issue of class unbalances.

Ray et al. [18] presented using several textual features, a Random Forest-based classification algorithm that evaluated sentiments on hotel reviews. The unbalanced class was solved using Random Oversampling and Bootstrap aggregation. In sentiment analysis of Amazon reviews, specifically of cell phones, Recurrent Neural Networks (RNN), Support Vector Machines, Artificial Neural Networks (ANN), and Naive Bayes suggested by Mukherjee et al. [19]. The imbalanced dataset is up-sampled to reduce severe overfitting to the negative sentiment class, resulting in consistent performance.

SSL is a revolutionary method at the same time that incorporates both unnamed and named data. It's critical to leverage the power of unnamed data for spectral signature perception in semi-supervised learning. Iosifidis and Ntoutsi [20] introduced data augmentation, including semantic augmentation by word embedding and corruption, as well as classic oversampling and undersampling strategies, to address the imbalance issue. Dogan and Uysal [21] developed two unique supervised term weighting algorithms for text classification termed TF-MONO and SRTF-MONO. To balance the data, SRTF-MONO can extract the higher features from the unbalanced dataset.

In summary to sentiment analysis, Tomek link method, Adaptive weight loss function, focal loss, random sampling method are employed to balance the dataset.

3. DATA BALANCING APPROACHES IN THE MEDICAL FIELD

Due to the unbalanced dataset, the classification accuracy of breast disease is not effective. To overcome this challenge, Devarriya et al. [22] proposed a two-stage network with genetic programming. The Distance score (D score), focuses on learning about minor and major classes equally and objectively but the f2 score concentrates only on minor classes. For multi-class classification, Huang et al. [23] suggested the association rule-based feature selection technique which produces the feature vector of the image. To balance the dataset, the author employed the RBS Bagging algorithm.

For covid-19 classification, Öztürk et al. [24] developed a support vector machine-based technique. During this process, they utilized SMOTE technique to balance the unbalanced dataset which generates the artificial data on minor classes to increase the amount of dataset. Bria et al. [25] performed a tiny lesion detection process with imbalanced data (high class). For this, they employed the Deep cascade (DC) technique to decrease the high-class samples which improve the accuracy of the classification.

The automatic Gastrointestinal (GI) disease classification technique was introduced by Öztürk and Özkaya [26] based on the convolutional neural network (CNN) framework which requires an effective data balancing technique to categorize the imbalanced datasets. In this classification, the features of the data are extracted by the CNN layers which are transferred to the LSTM network. Finally, the LSTM layers generate the final classification outcome. In the plant disease classification process, Double GAN (Generative Adversarial Network) based technique was employed by Zhao et al. [27]. This approach produces the damaged plant leaves pictures with high resolution for an imbalanced dataset. Then the classification process was conducted by DenseNet121, VGG16, and ResNet50 with this balanced dataset.

To solve the unbalanced data problem, a data oversampling-based technique was suggested by Gan et al. [28] on medical datasets. For binary data classification, they used a deep learning-based framework. Similar to the study [27], GAN based technique was utilized by Shamsolmoali et al. [29] to create artificial images to balance the dataset. For the classification task, they introduced a capsule network. Özdemir et al. [30] performed the multi-class hyperspectral image classification with CNN. Moreover, they improve the classification performance based on the balanced dataset by the use of the existing oversampling techniques such as Clustering, K-Means, Adasyn, and Smote and analyze the performance of these techniques. Another data balancing technique called Focal loss was introduced by Hammad et al. [31] during the detection process of MI. To detect MI, a deep learning-based CNN technique was utilized.

To achieve high accuracy and improve the stability of the technique balancing the dataset is an important task. Therefore, Sayed et al. [32] presented the data augmentation technique with a random over-sampling method (ROS) to balance the melanoma image dataset. By using these data sampling techniques, they achieve better results in melanoma skin cancer prediction tasks without the over-fitting problem.

The summary of the data balancing approach in medical field, Double GAN outperforms most of the algorithms. We can see that other than D2 score and F2 score, no classification boosting algorithm could beat any other algorithm currently in use.

4. DATA BALANCING APPROACHES IN INTRUSION DETECTION

In the intrusion detection process, an imbalanced dataset is very crucial which easily affects the detection rate of intrusions. Therefore, the SGM-CNN-based intrusion detection technique was suggested by Zhang et al. [33]. They combine GMM and SMOTE techniques to balance the large-scale dataset. To under-sampling the major classes, GMM-based clustering was implemented, and to increase the samples in minor classes, SMOTE technique is utilized.

The difficult Set Sampling Technique (DSSTE) algorithm is used as a data balancing technique by Liu et al. [34]. Initially, the imbalanced dataset is separated into an easy set and a difficult set using the Edited Nearest Neighbor (ENN) algorithm. Afterward, the major class samples are decreased by K Means algorithm and the minor class samples are increased by the data augmentation technique. Finally, these two subsets are combined to create a balanced dataset which increases the detection rate of intrusions. Similarly, a hybrid sampling technique was implemented by Jiang et al. [35] for intrusion detection. They used SMOTE technique to increase the number of minority samples and the one-side selection (OSS) technique was used to decrease the noise samples. Due to this way, they achieved a balanced dataset.

To generate the samples, Imbalance Generative Adversarial Network (IGAN) was suggested by Huang and Lei [36] for intrusion detection. A data filter is used by the IGAN to confirm that sampled data are minority classes which increases the performance of the classifier. At last, final intrusion detection was performed with this balanced dataset by the deep neural network. In another work (Bedi et al. [37]) the improved Siam-IDS (I-Siam IDS), was introduced to handle the imbalanced data problem which is a two-layer ensemble network. These techniques recognize majority and minority classes without data-level balancing approaches.

The combination of deep learning and machine learning is presented as an intrusion detection model by Liu et al. [38]. To tackle the imbalanced data issue the oversampling approach named ADASYN was utilized. Due to this the amount of minority samples is increased. For tackling class imbalance in network intrusion detection, the M-AdaBoost-A algorithm was presented by Zhou et al. [39] which incorporates the area under the curve to improve the process.

To address the data imbalance problem, Lee and Park [40] proposed Generative Adversarial Networks (GAN). Because it resamples by defining the desired rare class, the GAN can handle overfitting as well as problems related to noise and class overlaps. Then random forest-based classifier is utilized to classify the intrusions. Another work conducted by Al and Dener [41] utilized the hybrid approach named STL for the unbalanced dataset. It is the combination of Tomek-Links sampling and SMOTE techniques. It drastically reduced the imbalanced data problem. For intrusion classification, the hybrid deep learning approach named Long Short-Term Memory (LSTM) and CNN was applied.

The cost-sensitive stacked auto-encoder (CSSAE) based data sampling technique was presented by Telikani and Gandomi [42] which improves the performance of the classifier. In this technique, the distributions of samples are computed to generate the cost matrix for all classes in the dataset which is given as an input to the stacked sparse auto-encoder for the classification process.

Moreover, an improved CNN-based intrusion detection system was developed by Hu et al. [43] with the utilization of a data sampling technique called the adaptive synthetic sampling (ADASYN) algorithm. This technique helps to maintain the level of data samples in an equal manner and preserve the classifier from being sensitive to the unbalanced dataset.

In summary to intrusion detection, the existing approaches like SMOTE, Imbalance Generative Adversarial Network (IGAN), ADASYN and GMM are employed to balance the dataset. Comparing with other approaches SMOTE, GMM-based clustering and ADASYN are out performs well.

5. DATA BALANCING APPROACHES IN FRAUD DETECTION

Depending on the divide-and-conquer concept, to handle the issues of class imbalance with overlap using a hybrid strategy proposed by Li et al. [44]. Based on this concept, a large amount of major and a few minor class samples are removed to balance the classes in the dataset. This new subset provides reduced learning interference and decreased imbalance ratio. Unbalanced datasets depending on the semantic fusion of k-means were used to evaluate two-level credit card fraud tracking methods and also to strengthen identification precision and speed up detection convergence by artificial bee colony (ABC) by Darwish [45].

To solve the data balance issue, Baesens et al. [46] used the enlarged training set to apply the following over-sampling methods: MWMOTE, ROSE, SMOTE, and ADASYN. These techniques have default parameters that are assigned by the authors. To generate a fresh and more balanced training set used this over-sampling technique. The automotive insurance fraud detection system was conducted by Majhi [47] based on the combination of the fuzzy clustering method (FCM) and the modified whale optimization algorithm (MWOA). In his paper, they use MWOA to solve the imbalanced data problem. To optimize the cluster centroids (AIFDS), this sampling technique was utilized (MWOA). To balance the dataset, this technique removes the exceptions in the majority of class samples.

Hancock and Khoshgoftaar [48] proposed Medicare fraud detection using CatBoost and Light GBM to encode categorical data. Randomly to address the problem of class imbalance, under-sampling is used. Randomly the amount of datasets is reduced when this class ratio is under-sampled, and the negative class threatens data loss. CFXGB (Cascaded Forest and XG Boost) or click fraud detection by Thejas et al. [49]. To balance the unbalanced dataset, under-sampling was performed.

The summary of the data balancing approach in fraud detection, CatBoost outperforms most of the algorithms. We can see that other than ADASYN, no classification boosting algorithm could beat any other algorithm currently in use. While comparing existing techniques, cat boost and ADASYN approach performs well.

6. ANALYSIS OF RESULT

To assess the classifier's performance, the results generated by the classifier is needed to be analyzed with the balanced dataset. The evaluation of the suggested techniques is analyzed by the performance metrics which indicate the capacity of the classifier. Based on this we can make further enhancements depending on their insufficiencies. Accuracy, precision F1-score, and AUC are considered as the evaluation parameters. From the results, it is clear that the best performance metric, that is, those with no bias due to imbalance, is accuracy.

6.1 Evaluation of the techniques

The performance of above mentioned imbalanced data classification techniques is compared in the below table. Table 1 shows the majority of the unbalanced data categorization techniques proposed recently.

Table 1. Methods for metric-based classification of unbalanced data

Year	Reference	Techniques	Application related to	Data set	Accuracy	AUC	Precision	F1-score
2020	Studiawan et al. [11]	Tomek link method	Sentiment analysis	OS log dataset	99.93%	-	99.83%	99.84%
2020	Cheng et al. [12]	Focal Loss	Sentiment analysis	TSB, Waimai, Weibo, NLPCC2014, Yelp 2013 Amazon	97.71%	-	-	-
2021	Gan et al. [13]	Adaptive weight loss function	Sentiment analysis	LPCC2017-ECGC1 and ChnSentiCorp-Htl-unba-10000	97.91%	-	-	97.68%
2020	Liu et al. [14]	Data augmentation	Sentiment analysis	BC3, Enronff, PA Dataset	91.80%	-	-	-
2021	Budhi et al. [15]	Random sampling methods	Sentiment analysis	YelpNYC, YelpZIP, YelpChi Hotel, YelpChi Restaurant	99%	-	-	-
2020	Xia et al. [16]		Sentiment analysis	Chinese and English online review	90%	-	-	-
2020	Ruz et al. [17]	SMOTE	Sentiment analysis	Twitter dataset	85.8%	-	90.6%	87.9%
2021	Ray et al. [18]	Random Over-Sampling, Bootstrap aggregating	Sentiment analysis	Tripadvisor A	92.36%	-	86%	82%
2021	Mukherjee et al. [19]	upsampling	Sentiment analysis	Amazon reviews	96.32%	95.8%	98.50%	95.62%
2020	Iosifidis and Ntoutsi [20]	Traditional under-sampling and oversampling	Sentiment analysis	T-sentiment	93.52%	-	94.16%	94.69%
2020	Dogan and Uysal [21]	SRTF-MONO	Sentiment analysis	Reuters-21578, 20-News groups, WebKB.	-	-	-	-
2020	Devarriya et al. [22]	D2 score and F2 score	Medical field	Breast cancer dataset	99.51%	-	-	-
2020	Huang et al. [23]	RBS Bagging algorithm	Medical field	diabetes and UCI public dataset	74.9%	-	72.7%	73.4%
2021	Öztürk et al. [24]	SMOTE	Medical field	COVID-19	94.23%	-	96.73%	93.99%
2020	Bria et al. [25]	Deep cascade	Medical field	IN breast database	-	-	-	-
2021	Öztürk and Özkaya [26]	CNN-LSTM	Medical field	kvasir dataset	98.05%	-	98.05%	98.05%
2021	Zhao et al. [27]	Double GAN	Medical field	Plant village dataset	99.53%	-	-	-
2020	Gan et al. [28]	Oversampling	Medical field	Heart, ILPD, Dermatology dataset, and (CCRF)	88.87%	-	-	-
2021	Shamsolmoali et al. [29]	GAN	Medical field	MNIST, CelebA, and CIFAR-10 datasets	88.9%	-	92.7%	92.1%
2021	Özdemir et al. [30]	Smote, Adasyn, K-Means Cluster methods	Medical field	EEE-Dataport Machine Learning Repository	96.69%	-	95.01%	97.44%

2021	Hammad et al. [31]	Focal loss	Medical field	Physikalisch-Technische Bundesanstalt (PTB) dataset	98.8%	-	98.31%	97.92%
2021	Sayed et al. [32]	random over-sampling method	Medical field	Xuzhou Hyspex dataset in	98.37%	-	95.01%	97.4%
2020	Zhang et al. [33]	SMOTE and GMM-based clustering	Intrusion detection	UNSW-NB15 and CICIDS2017 dataset	99.74%	-	91.66%	95.53%
2020	Liu et al. [34]	DSSTE	Intrusion detection	NSL-KDD and CSE-CIC-IDS2018	82.84%	-	-	81.66%
2020	Jiang et al. [35]	SMOTE	Intrusion detection	UNSW-NB15 and NSL-KDD dataset	77.92%	-	84.95%	84.35%
2020	Huang et al. [36]	GAN	Intrusion detection	CICIDS2017, UNSW-NB15,and NSL-KDD dataset	77.16%	-	84.85%	84.47%
2021	Bedi et al. [37]	I-Siam IDS	Intrusion detection	NSL-KDD and CIDDS-001 datasets	89%	95%	-	-
2021	Liu et al. [38]	ADASYN	Intrusion detection	NSL-KDD and CIS-IDS2017 datasets	99.91%	-	-	-
2020	Zhou et al. [39]	M-Adaboost	Intrusion detection	AWID and NSL-KDD datasets.	99.9%	-	88.04%	91.93%
2021	Lee et al. [40]	GAN	Intrusion detection	CICIDS 2017 dataset	99.83%	-	98.68%	95.04%
2021	Al and Dener [41]	SMOTE and Tomek-Links	Intrusion detection	CIDDS-001 and UNS-NB15 data set	99.83%	-	99%	99%
2021	Telikani and Gandomi [42]	cost-sensitive stacked auto-encoder	Intrusion detection	NSL-KDD and KDD ’99 datasets	99.06%	-	99.02%	99%
2020	Hu et al. [43]	adaptive synthetic sampling	Intrusion detection	NSL-KDD datasets	84.08%	-	-	-
2021	Li et al. [44]	Random undersampling	Fraud detection	an electronic transaction dataset and a fraud detection dataset	-	-	-	-
2020	Darwish et al. [45]	semantic fusion of k-means cluster	Fraud detection	transaction’s dataset	98.9%	-	-	-
2021	Baesens. et al. [46]	ROSE, MWMOTE, ADASY, and SMOTE	Fraud detection	credit card transaction dataset	-	91.02%	-	-
2021	Majhi [47]	fuzzy clustering method	Fraud detection	automobile insurance dataset	86.38%	-	-	-
2021	Hancock and Khoshgoftaar [48]	Random Undersampling	Fraud detection	Healthcare Common Procedure Coding System (HCPCS) code.	-	95.7%	-	-
2021	Thejas et al. [49]	Under-sampling	Fraud detection	TalkingData, Avazu, and Kad dataset	-	96.45%	96%	96%

They help to provide better classification accuracy. In sentiment analysis, Studiawan et al. [11] achieve the highest accuracy with the imbalanced data. To balance the dataset, they used the Tomek link method which supports the classification method to gain 99.93% accuracy. In the medical field, Zhao et al. [27] achieve the highest accuracy with this Double GAN approach which supports the classification method to gain 99.53% accuracy. In Intrusion detection, Liu et al. [38] achieve the highest accuracy with imbalanced data. To balance the dataset, they used ADASYN which supports the classification method to gain 99.91% accuracy. And the last one fraud detection, Darwish [45] achieve the highest accuracy with this Semantic fusion k approach which supports the classification method to gain 98.9% accuracy. Finally, it is concluded that the ADSYN approaches yield the best results from the above discussion. In medical field, Huang et al. [23] presented a RBS Bagging algorithm, which yields an accuracy of 74.9%, 72.7% of precision and 73.4% of F1-score. Comparing to other approaches RBS Bagging algorithm performances was poor. Then Shamsolmoali et al. [29]

presented GAN approach to balance the dataset which supports the classification method to gain 88.9% accuracy, which is the second least accuracy obtain in medical data. In intrusion detection system, Huang et al. [36] introduced GAN method which achieve a lowest accuracy of 77.16%. Jiang et al. [35] introduced SMOTE approach which achieve a second lowest accuracy of 77.92%. At last in fraud detection Majhi et al. [47] introduced fuzzy clustering method to balance the unbalanced dataset which yields an accuracy of 86.38%.

7. FUTURE WORK

Imbalanced data sets classification models are becoming more common, and several researchers have suggested a variety of approaches depending on the characteristics of imbalanced data sets. However, there are still too many issues of imbalanced data sets that need to be simplified and overcome, which the list will review for future research on this type of data set.

7.1 Improvement of imbalanced data classification in sentiment analysis

To look at the impact of unbalanced datasets on behavioral features and the possibilities of enhancing them with a new hybrid method. Additional data augmentation approaches as a tool for meaningfully enlarging a training set and addressing issues of class unbalance. Concentrate on how to find significant augmentations for a specific domain, as well as which segments of the population should be augmented, to reduce noise generation and data quality degradation. We strive to focus on various data augmentation strategies during this process to reduce noise creation and data quality degradation while also expanding the training set. To address the issue of data imbalance, we used an augmentation strategy.

7.2 Improvement of imbalanced data classification in the medical field

A larger dataset with more balanced and labeled data will be created. The classification algorithm will be tested to see if it can deliver good performance on such datasets. The development of new data balancing techniques will improve performance. To test our approach on different unbalanced datasets so that we can analyze our classification findings with a broader range of data. To assess the accuracy of classification and effectiveness of the approaches suggested, researchers decided to evaluate the work of such fitness functions with multiple class datasets. Work on huge data to improve the accuracy of the suggested system, which can be done through augmentation techniques or additional data. In future work, we can also make better classification accuracy on the discovery of new balancing techniques and employing an augmentation technique.

7.3 Improvement of imbalanced data classification in intrusion detection

Intend to apply the DL approach for extracting the feature and model training on the network traffic data in the following step, leveraging DL capabilities in extracted features to mitigate the effects of unbalanced data and obtain higher classification accuracy. After sampling the imbalanced training set samples through the approach, the deep learning method outperformed machine learning. In future work, the correct classifier techniques must be applied to achieve good results. There are a variety of approaches to dealing with unbalanced data, but the major focus should be on selecting the right feature extraction approach.

7.4 Improvement of imbalanced data classification in fraud detection

To discover a trade-off among evaluation metrics, a strategy for the class imbalance problem is being developed. Building a Big Data-driven ecosystem and testing the model to the study with huge amounts of data. Because they couldn't investigate significant volumes of data, the size of the data was an evident constraint of the study. As a result, they intend to use Big Data technology to create a scalable ecosystem. In our future work, we plan to develop a model that can handle the class imbalance problem and provide a large amount of data to provide big data.

Research gap

Following are explanations of how this research fills the gap found in the previous study.

- According to existing investigations, the majority of the studies have used a single dataset, which makes each computational algorithm adequate for diagnosing the specific disorder.
- Another issue examined in previous research is that the majority of them either did not address class imbalance or only used one method to manage it. To the best of our knowledge, no prior research studies have examined various balancing methods across various datasets.
- The computational approaches presented in the investigations have not been able to achieve acceptable prediction accuracy, which represents a significant research gap evident in earlier works.

Managerial implications

Every research begins with the idea that it will have some social implications or improve the social and organizational conditions. There are some additional goals associated with the study as well, such as assisting other academics and identifying areas that have escaped their attention up until now. The implications of this study are similar. It presents the results of numerous studies on unbalanced data for various domains. Imbalanced datasets create challenges for predictive modelling, but they're actually a common and anticipated problem because the real world is full of imbalanced examples. Balancing a dataset makes training a model easier because it helps prevent the model from becoming biased towards one class. Even though the topic in issue has received extensive study, there is still a need for more research to be done. The following are some research themes: the efficiency of various training methods in improving training efficiency. Impact of corporate support infrastructure on training efficacy. Motivation's impact on an unbalanced collection.

8. CONCLUSION

In today's world, the imbalance is a relatively typical problem that leads to significant deviations in the performance of classical classifiers. In this work, the issue of uneven data and the requirement for data balancing is deeply analyzed. This review paper provides a comprehensive analysis of the imbalance problem based on recent research papers. In many real-world fields, such as telecommunications, fraudulent call detection, and medical diagnosis, the imbalance problem arises. Although there are numerous approaches for dealing with the imbalance issues in different areas, the fraud detection sector still requires major attention.

Many methods and techniques have been suggested to counteract the negative impacts of data imbalance because imbalanced data limits the effectiveness and precision of classifiers. It has been done and discussed to compare current approaches, as well as the application domains in which they are used. These methods have been fine-tuned to obtain the generalized learning model in numerous proposed papers with validation of desired outcomes. Contrasting other evaluation metrics, such as AUC, Accuracy, precision and F1-Score metrics that are appropriate for the multi-class imbalanced area.

REFERENCES

[1] Vuttipittayamongkol, P., Elyan, E., Petrovski, A. (2021). On the class overlap problem in imbalanced data classification. Knowledge-Based Systems, 212: 106631. https://doi.org/10.1016/j.knosys.2020.106631

[2] Koziarski, M. (2020). Radial-based undersampling for imbalanced data classification. Pattern Recognition, 102: 107262. https://doi.org/10.1016/j.patcog.2020.107262

[3] Xu, Z., Shen, D., Nie, T., Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on the random forest for medical imbalanced data. Journal of Biomedical Informatics, 107: 103465. https://doi.org/10.1016/j.jbi.2020.103465

[4] Koziarski, M., Woźniak, M., Krawczyk, B. (2020). Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise. Knowledge-Based Systems, 204: 106223. https://doi.org/10.1016/j.knosys.2020.106223

[5] Al Majzoub, H., Elgedawy, I., Akaydın, Ö., Köse Ulukök, M. (2020). HCAB-SMOTE: A hybrid clustered affinitive borderline SMOTE approach for imbalanced data binary classification. Arabian Journal for Science and Engineering, 45(4): 3205-3222.

[6] Ren, R., Yang, Y., Sun, L. (2020). Oversampling technique based on fuzzy representativeness difference for classifying imbalanced data. Applied Intelligence, 50(8): 2465-2487.

[7] Lin, E., Chen, Q., Qi, X. (2020). Deep reinforcement learning for imbalanced classification. Applied Intelligence, 50(8): 2488-2502.

[8] Feng, F., Li, K. C., Shen, J., Zhou, Q., Yang, X. (2020). Using cost-sensitive learning and feature selection algorithms to improve the performance of imbalanced classification. IEEE Access, 8: 69979-69996. https://doi.org/10.1109/ACCESS.2020.2987364

[9] Gao, X., Ren, B., Zhang, H., Sun, B., Li, J., Xu, J., He, Y., Li, K. (2020). An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling. Expert Systems with Applications, 160: 113660. https://doi.org/10.1016/j.eswa.2020.113660

[10] Gao, L., Zhang, L., Liu, C., Wu, S. (2020). Handling imbalanced medical image data: A deep-learning-based one-class classification approach. Artificial Intelligence in Medicine, 108: 101935. https://doi.org/10.1016/j.artmed.2020.101935

[11] Studiawan, H., Sohel, F., Payne, C. (2020). Anomaly detection in operating system logs with deep learning-based sentiment analysis. IEEE Transactions on Dependable and Secure Computing, 18(5): 2136-2148. https://doi.org/10.1109/TDSC.2020.3037903

[12] Cheng, K., Yue, Y., Song, Z. (2020). Sentiment classification based on part-of-speech and self-attention mechanism. IEEE Access, 8: 16387-16396. https://doi.org/10.1109/ACCESS.2020.2967103

[13] Gan, C., Feng, Q., Zhang, Z. (2021). Scalable multi-channel dilated CNN–BiLSTM model with attention mechanism for Chinese textual sentiment analysis. Future Generation Computer Systems, 118: 297-309. https://doi.org/10.1016/j.future.2021.01.024

[14] Liu, S., Lee, K., Lee, I. (2020). Document-level multi-topic sentiment classification of email data with bilstm and data augmentation. Knowledge-Based Systems, 197: 105918. https://doi.org/10.1016/j.knosys.2020.105918

[15] Budhi, G.S., Chiong, R., Wang, Z. (2021). Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. Multimedia Tools and Applications, 80(9): 13079-13097.

[16] Xia, H., Yang, Y., Pan, X., Zhang, Z., An, W. (2020). Sentiment analysis for online reviews using conditional random fields and support vector machines. Electronic Commerce Research, 20(2): 343-360.

[17] Ruz, G.A., Henríquez, P.A., Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. Future Generation Computer Systems, 106: 92-104. https://doi.org/10.1016/j.future.2020.01.005

[18] Ray, B., Garain, A., Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. Applied Soft Computing, 98: 106935. https://doi.org/10.1016/j.asoc.2020.106935

[19] Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S.M., Sangwan, R.S., Sharma, R. (2021). Effect of negation in sentences on sentiment analysis and polarity detection. Procedia Computer Science, 185: 370-379. https://doi.org/10.1016/j.procs.2021.05.038

[20] Iosifidis, V., Ntoutsi, E. (2020). Sentiment analysis on big sparse data streams with limited labels. Knowledge and Information Systems, 62(4): 1393-1432.

[21] Dogan, T., Uysal, A.K. (2020). A novel term weighting scheme for text classification: TF-MONO. Journal of Informetrics, 14(4): 101076. https://doi.org/10.1016/j.joi.2020.101076

[22] Devarriya, D., Gulati, C., Mansharamani, V., Sakalle, A., Bhardwaj, A. (2020). Unbalanced breast cancer data classification using novel fitness functions in genetic programming. Expert Systems with Applications, 140: 112866. https://doi.org/10.1016/j.eswa.2019.112866

[23] Huang, C., Huang, X., Fang, Y., Xu, J., Qu, Y., Zhai, P., Fan, L., Yin, H., Xu, Y., Li, J. (2020). Sample imbalance

disease classification model based on association rule feature selection. Pattern Recognition Letters, 133: 280-286. https://doi.org/10.1016/j.patrec.2020.03.016

[24] Öztürk, Ş., Özkaya, U., Barstuğan, M. (2021). Classification of Coronavirus (COVID-19) from X-ray and CT images using shrunken features. International Journal of Imaging Systems and Technology, 31(1): 5-15. https://doi.org/10.1002%2Fima.22469

[25] Bria, A., Marrocco, C., Tortorella, F. (2020). Addressing the class imbalance in deep learning for small lesion detection on medical images. Computers in biology and medicine, 120: 103735. https://doi.org/10.1016/j.compbiomed.2020.103735

[26] Öztürk, Ş., Özkaya, U. (2021). Residual LSTM layered CNN for classification of gastrointestinal tract diseases. Journal of Biomedical Informatics, 113: 103638. https://doi.org/10.1016/j.jbi.2020.103638

[27] Zhao, Y., Chen, Z., Gao, X., Song, W., Xiong, Q., Hu, J., Zhang, Z. (2021). Plant disease detection using generated leaves based on DoubleGAN. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19(3): 1817-1826. https://doi.org/10.1109/tcbb.2021.3056683

[28] Gan, D., Shen, J., An, B., Xu, M., Liu, N. (2020). Integrating TANBN with cost-sensitive classification algorithm for imbalanced data in medical diagnosis. Computers & Industrial Engineering, 140: 106266. http://dx.doi.org/10.1016/j.cie.2019.106266

[29] Shamsolmoali, P., Zareapoor, M., Shen, L., Sadka, A.H., Yang, J. (2021). Imbalanced data learning by minority class augmentation using capsule adversarial networks. Neurocomputing, 459: 481-493. https://doi.org/10.1016/j.neucom.2020.01.119

[30] Özdemir, A., Polat, K., Alhudhaif, A. (2021). Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods. Expert Systems with Applications, 178: 114986. https://doi.org/10.1016/j.eswa.2021.114986

[31] Hammad, M., Alkinani, M.H., Gupta, B.B., El-Latif, A., Ahmed, A. (2021). Myocardial infarction detection based on the deep neural network on imbalanced data. Multimedia Systems, 1-13.

[32] Sayed, G.I., Soliman, M.M., Hassanien, A.E. (2021). A novel melanoma prediction model for imbalanced data using optimized SqueezeNet by bald eagle search optimization. Computers in Biology and Medicine, 136: 104712. https://doi.org/10.1016/j.compbiomed.2021.104712

[33] Zhang, H., Huang, L., Wu, C.Q., Li, Z. (2020). An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in the imbalanced dataset. Computer Networks, 177: 107315. https://doi.org/10.1016/j.comnet.2020.107315

[34] Liu, L., Wang, P., Lin, J., Liu, L. (2020). Intrusion detection of imbalanced network traffic based on machine learning and deep learning. IEEE Access, 9: 7550-7563. https://doi.org/10.1109/ACCESS.2020.3048198

[35] Jiang, K., Wang, W., Wang, A., Wu, H. (2020). Network intrusion detection combined hybrid sampling with the deep hierarchical network. IEEE Access, 8: 32464-32476. https://doi.org/10.1109/ACCESS.2020.2973730

[36] Huang, S., Lei, K. (2020). IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks. Ad Hoc Networks, 105: 102177. https://doi.org/10.1016/j.adhoc.2020.102177

[37] Bedi, P., Gupta, N., Jindal, V. (2021). I-SiamIDS: An improved Siam-IDS for handling class imbalance in network-based intrusion detection systems. Applied Intelligence, 51(2): 1133-1151.

[38] Liu, C., Gu, Z., Wang, J. (2021). A hybrid intrusion detection system based on scalable K-means+ random forest and deep learning. IEEE Access, 9: 75729-75740. https://doi.org/10.1109/ACCESS.2021.3082147

[39] Zhou, Y., Mazzuchi, T.A., Sarkani, S. (2020). M-AdaBoost-a based ensemble system for network intrusion detection. Expert Systems with Applications, 162: 113864. https://doi.org/10.1016/j.eswa.2020.113864

[40] Lee, J., Park, K. (2021). GAN-based imbalanced data intrusion detection system. Personal and Ubiquitous Computing, 25(1): 121-128.

[41] Al, S., Dener, M. (2021). STL-HDL: A new hybrid network intrusion detection system for an imbalanced dataset in a big data environment. Computers & Security, 110: 102435. https://doi.org/10.1016/j.cose.2021.102435

[42] Telikani, A., Gandomi, A.H. (2021). Cost-sensitive stacked auto-encoders for intrusion detection in the Internet of Things. Internet of Things, 14: 100122. https://doi.org/10.1016/j.iot.2019.100122

[43] Hu, Z., Wang, L., Qi, L., Li, Y., Yang, W. (2020). A novel wireless network intrusion detection method based on adaptive synthetic sampling and an improved convolutional neural network. IEEE Access, 8: 195741-195751. https://doi.org/10.1109/ACCESS.2020.3034015

[44] Li, Z., Huang, M., Liu, G., Jiang, C. (2021). A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with an overlap in credit card fraud detection. Expert Systems with Applications, 175: 114750. https://doi.org/10.1016/j.eswa.2021.114750

[45] Darwish, S.M. (2020). A bio-inspired credit card fraud detection model based on user behavior analysis suitable for business management in electronic banking. Journal of Ambient Intelligence and Humanized Computing, 11(11): 4873-4887.

[46] Baesens, B., Höppner, S., Ortner, I., Verdonck, T. (2021). robROSE: A robust approach for dealing with imbalanced data in fraud detection. Statistical Methods & Applications, 30(3): 841-861.

[47] Majhi, S.K. (2021). Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. Evolutionary Intelligence, 14(3): 35-46. https://link.springer.com/article/10.1007/s12065-019-00260-3

[48] Hancock, J.T., Khoshgoftaar, T.M. (2021). Gradient boosted decision tree algorithms for medicare fraud detection. SN Computer Science, 2(4): 1-12. https://link.springer.com/article/10.1007/s42979-021-00655-z

[49] Thejas, G.S., Dheeshjith, S., Iyengar, S.S., Sunitha, N.R., Badrinath, P. (2021). A hybrid and effective learning approach for Click Fraud detection. Machine Learning with Applications, 3: 100016. https://doi.org/10.1016/j.mlwa.2020.100016