Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 26 (2023), No. 1, pp. 177–187
DOI : 10.47974/JSMS-957

Sampling strategies for handling data imbalance problem: An Extensive Review

Bhaskar Kumar Veedhi [§]
Debahuti Mishra [†]
Department of Computer Science and Engineering
Siksha 'O' Anusandhan (Deemed to be University)
Bhubaneswar
Odisha
India

Kaberi Das *
Department of Computer Applications
Siksha 'O' Anusandhan (Deemed to be University)
Bhubaneswar
Odisha
India

Abstract

The imbalanced data classification is a major issue in data mining. Many researchers have proposed various solutions which addressed imbalanced data problem which is broadly categorized into data level and algorithm level. Class distributions are adjusted in data level method. Creating an algorithm or modifying the existing algorithm is an appropriate approach used in algorithm level method. Imbalanced data classification problem can be resolved by means of Sampling, Random over sampling, Random under sampling, Resampling and by SMOTE (Synthetic Minority Oversampling Techniques). Resampling includes k-means clustering, density-based clustering, neural networks and ensemble. However, no algorithm or a method has an ability to remove bias in data classification, thereby integration of kernel methods with sampling methods or integration of sampling and boosting methods or integration Kernel based with Support Vector Machines (SVM) need to be performed a great extent to get the desired accuracy and performance. The main objective of this paper is to focus on various sampling strategies that are based on sampling and re-sampling methods and improving the concept of learning within class imbalanced data. It

[§] *E-mail:* **BhaskarKumarV@hcl.com**
[†] *E-mail:* **debahutimishra@soa.ac.in**
E-mail:* **kaberidas@soa.ac.in (Corresponding Author)

also explains the objectives of the models used by several researchers and emphasized the performance along with the outcomes.

Subject Classification: *68T07.*

Keywords: *Imbalanced data, Classification, Skewed data, Sampling methods, SMOTE.*

1. Introduction

Imbalanced data classification problem exists in many fields such as, data mining, medical data analysis, facial recognition, bioinformatics etc. However, to classify the data in presence of imbalanced data is a challenging task [1]. In most of the domain applications, class imbalance distribution problem occurs often, when a dataset is unbalanced if the classification categories are not equally distributed in the training dataset [2]. Objects in one class will be more in number than the objects in the other class [3]. Solving the imbalance problem is quite difficult task as the hypothesis disobeys the classifier algorithm. And this problem will be challenging for mixed type of data in many real time scenarios and in data mining[4], [5].Though we have many approaches and algorithms, we don't have one defined algorithm which can resolve the problem completely to the root[6], [7]. That is, no single algorithm or classifier suits the best for all imbalances learning problems. Different algorithm suits different data sets. Given an imbalance learning dataset which one to choose from the various available methods is a main aim of this research.

1.1 *Motivations*

Motivation behind studying the sampling strategies is to resolve the issues of data imbalance in real time environment, it's necessary to study the underrepresented group which is biased or ignored. Solving the problem of imbalance helps to predict health issues in medical field and also resolving the problem of data imbalance in financial sector such as banks and in telecommunications domains.

1.2 *Challenges of data imbalance*

Being a vast field, vital issues prevail that are to be addressed for a clear and better understanding. Performance degradation is serious challenge, which is in both binary and multiclass data imbalance, apart from the removal of noisy data. Construction of perfect classifier is quite a challenging task as the classifier will not be able to completely remove the problem of classification in imbalanced data, so the identification of best classification algorithm and a pre-processing activity must rigorously be studied and stressed more for a better performance result. Identification of a specialized technique and to come up with a model for different data sets is the key challenge. And classification problem in multi class is more challenging compared to binary class.

In this paper Section 1 describes the data imbalance problem along with the motivations and challenges. In Section 2, we have discussed the sampling methods and the state of art which represent how various sampling strategies are used to solve the data imbalance problem. Finally in Section 3, we have summarised with all the methods discussed throughout the paper.

2. Study on various strategies for handling data imbalance problems/ issues

2.1 *Sampling methods*

Rebalancing the class distribution is performed in data level approach, which can be done either by under-sampling or over sampling to minimize the bias in training dataset. The purpose of the under-sampling is to eliminate sparse number of examples from majority class to minimize the bias in two classes. Whereas over sampling[8],[9] replicates examples from majority class. In the process of eliminating examples (down sampling) from a class tends to substantial loss of key information. While in examples of replication[10],[11] (over sampling), increase in the number of examples happens by means of duplication but do not provide new examples. Despite of numerous efforts in handling imbalance issues through sampling, one can say these is no concrete standard for appropriate class distribution. But to overcome imbalance issues, a combination of sampling strategies with other strategies has given a proven results in binary class imbalance and even in multi class imbalance problems.

2.1.1 Random over-sampling

Random over-sampling [12] checks the observations randomly, distributes it evenly to over sample. Both old and new samples look alike as its duplicated to meet the requirement of balancing the data, chosen option is replacement or without replacement. Selected data is sampled again, in with replacement, but with some replacement limit. Provision of sampling again from the existing sample is not possible in few scenarios. In over sampling [13], the problem of over fitting is a concern along with noise issue [14]. Both the issues are to be handled to resolve the imbalance problems.

Table 2.1.1

A gist of work done on Random Over Sampling for last 10 years [2010-2021]

Literature	Year	Objective	Methodologies adopted	Findings
Rashmi Dubey *etal. [15]*	2014	Demonstrated K-Medoid under sampling over other data resampling in ADNI dataset.	Over sampling, Under sampling and a hybrid method	It's concluded that under sampling in particular K-Medoids has given a better result than any other resampling method in ADNI study. (Disease name Alzheimer)
Huaxiang Zhang *etal.[16]*	2014	Study each over sampling approach, how differently over sampling rates differ when different algorithms implemented.	Random Walk Over Sampling (RWO)	Datasets with continues attributes has given good results but discrete attributes didn't perform well. Encounters the problem of overfitting.
Myoung-Jong Kim *etal.[17]*	2015	Geometric mean Boost (GMBoost) applied on Bankruptcy prediction datasets	Sampling- GMBoost (Geometric Mean Based Boosting)	Aided in resolving predicting bankruptcy. Given a proven results then Adaboost and cost sensitive methods. Prediction capability is high.
Kyoham Shin *etal. [18]*	2021	Addressing imbalanced and incomplete data classification parallelly.	MI-MOTE - Oversampling Conditional Wasserstein GAN-based oversampling	MI-MOTE proven to be effective in case of missing training data sets is higher. Does not perform well when missing data sets is low.

Contd...

Ankit Vijayvargiya*etal.* *[19]*	2021	Human Knee abnormality detection.	Oversampling	Resolution of data imbalance in sEMG (surface electromyography) using oversampling technique. Human knee abnormality Detected.
Justin Engelmann *etal.[20]*	2021	Resolution of Data imbalance in Credit scoring datafor a binary classification.	Conditional Wasserstein GAN-based oversampling	Further study to be extended to multi class settings for a problem of rating migration analysis.

Table 2.1.2

A gist of work done on Random/ Informed Under-sampling for last 10 years [2010-2021]

Literature	Year	Objective	Methodologies adopted	Findings
Hualong Yu*etal.* *[23]*	2013	Eliminates the noisy genes	Ant Colony optimization (ACO) Under Sampling	Suitable for small sample. Computational and storage cost is high. Efficiency to be improved by modifying formation rule.
Mikel Galar*etal.* *[24]*	2013	Improve the performance with existing base classifier	Ensemble (EUS Boost)	EUS Boost outperformed based on the ensembles.
Peng Cao *etal.[25]*	2014	Lungnodal detection	Ensemble-based hybrid probabilistic sampling	Method only for two class imbalance issues, to be extended for multi-class imbalancedclassification.
J. Hoyos-Osorio*etal.[26]*	2021	Robust to complex data distribution.	Relevant information under sampling	RIUS method combined with clustering technique to curb imbalance data classification named as CRIUS. Extending RIUS and CRIUS to multi class. Underlying structure is not lost.

Table 2.1.3
A gist of work done on SMOTE for last 10 years [2010-2021]

Literature	Year	Objective	Methodologies adopted	Findings
Ming Hao *etal.[29]*	2014	GLMBoost and SMOTE to tackle the imbalanced problem.	GLM Boost with Smote	Clear Comparison is done with a method Random Forest combined with SMOTE on same set of data. Higher performance and also has high computational efficiency than Random Forest combined with SMOTE.
Yuqiang Fan *etal.[30]*	2019	Fault diagnosis of chillers	PCA combined with SMOTE-SVM	PCA combined with SMOTE-SVM aided in improving the diagnosis of fault chillers reducing the time and cost of simulating the fault experiments.
Amir BahadorParsa*etal.[31]*	2019	To Detects the accidents	SMOTE – Accident detection	PNN (Probabilistic Neural Networks) and SVM. To overcome the problem of data imbalance, SMOTE is used.
Hong-Jie Dai *etal.[32]*	2019	Adverse Drug Reactions	WESSMOTE	WESSMOTE approach is used to remove the problem of data imbalance. It has better results than the normal SMOTE.
Asniar*etal.[33]*	2021	Noise Identification	SMOTE – LOF (Local Outlier Factor)	Better results for a smaller data set, future study on different data sets varying a greater number of data samples.

2.1.2 Random under-sampling

Random Under-sampling eliminates the examples of major class to reduce the severity in biased data. Under-sampling [21] outperforms oversampling techniques when data is small in size. Major drawback of random under sampling [22] is probability of discarding of potential data. Loss of key information has to be taken care when eliminating the majority class examples. There are good number of methods which focus on which majority examples to keep and which to eliminate.

2.1.3 Synthetic sampling (SMOTE)

SMOTE (Synthetic Minority Oversampling Technique) [27] is most commonly used advanced version of oversampling. Creates synthetic samples in underrepresented observations [28]. It does not generate the duplicates but creates synthetic data points which are slightly varying from actual data points.

3. Summary

Sampling provides an ideal solution for classification problems other than few non sampling methods as it aids in allowing of new information or removal of unwanted information which results to balance the data. Boosting is another ensemble algorithm which enhances the performance of classifier. Smote by itself helps in solving the problem of data imbalance, however combination of two or more techniques i.e., hybrid approach will result to all time ideal solutions. Methods such as GLMBoost + SMOTE, Radial based SMOTE, PCA combined with SMOTE-SVM has proven results in solving the problem of data imbalance. Though we have many approaches and algorithms, we don't have one defined classification algorithm which can resolve the problem completely. In other words, no single algorithm or classifier suits the best for all imbalances learning problem. Performance of a classification method depends on the type of datasets. For a better performance, it is always good to adopt a hybrid algorithm/classification method. Combination of one or more classifier is an optimal way and has given an outstanding result with accuracy and good computational speed.

References

[1] Hongwei Ding, Leiyang Chen, Liang Dong, Zhongwang Fu, XiaohuiCui,"Imbalanced data classification: A KNN and generative

adversarial networks-based hybrid approach for intrusion detection", *Future Generation Computer Systems*, 131(2022) 240-254.

[2] Wenyang Wang, Dongchu Sun, "The improved AdaBoost algorithms for imbalanced data classification", *Information Sciences*, 563(2021) 358-374.

[3] Jiakun Zhao, Ju Jin, Si Chen, Ruifeng Zhang, Bilin Yu, Qingfang Liu, "A weighted hybrid ensemble method for classifying imbalanced data", *Knowledge-Based Systems*, 203(2020)106087.

[4] PabasaraAthukorala, Madusha Chathurangi & Rajitha Ranasinghe, "A variant of RSA using continued fractions", *Journal of Discrete Mathematical Sciences and Cryptography*, Volume 25, Issue 1 (2022)127-134.

[5] Ruchi Kaushik, Vijander Singh & Rajani Kumari, "Multi-class SVM based network intrusion detection with attribute selection using infinite feature selection technique", *Journal of Discrete Mathematical Sciences and Cryptography*, Volume 24, Issue 8 (2021)2137-2153.

[6] VijayasriIyer, Bhargava Ganti, A. M. Hima Vyshnavi, P. K. Krishnan Namboori&Sriram Iyer, "Hybrid quantum computing based early detection of skin cancer", *Journal of Interdisciplinary Mathematics*, Volume 23, Issue 2 (2020)347-355.

[7] Osama R. Shahin, Rasha M. Abd El-Aziz&Ahmed I. Taloba, "Detection and classification of Covid-19 in CT-lungs screening using machine learning techniques", *Journal of Interdisciplinary Mathematics*, Volume 25, Issue 3 (2022)791-813.

[8] Quoc Hoan Doan, Sy-Hung Mai, Quang Thang Do, Duc-Kien Thai, "A cluster-based data splitting method for small sample and class imbalance problems in impact damage classification", Applied Soft Computing, 120 (2022) 108628.

[9] Ming Zheng, Tong Li, Liping Sun, Taochun Wang, Biao Jie, Weiyi Yang, Mingjing Tang, Changlong Lv, "An automatic sampling ratio detection method based on genetic algorithm for imbalanced data classification", *Knowledge-Based Systems*, 216(2021) 106800.

[10] Xin Gao, Bing Ren, Hao Zhang, Bohao Sun, Junliang Li, Jianhang Xu, Yang He, Kangsheng Li, "An ensemble imbalanced classification

method based on model dynamic selection driven by data partition hybrid sampling", *Expert Systems with Applications*, 160 (2020) 113660.

[11] Dohyun Lee, Kyoungok Kim, "An efficient method to determine sample size in oversampling based on classification complexity for imbalanced data", *Expert Systems with Applications*, 184(2021) 115442.

[12] ChenxiaJin, Fachao Li, Shijie Ma, Ying Wang, "Sampling scheme-based classification rule mining method using decision tree in big data environment", Knowledge-Based Systems, 244(2022)108522.

[13] G.Douzas, F.Bacao, F.Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and smote" *Inf. Sci. 465 (2018) 1–20.*

[14] Rodolfo M. Pereira, Yandre M.G. Costa, Carlos N. Silla Jr., "Toward hierarchical classification of imbalanced data using random resampling algorithms", *Information Sciences*, 578(2021) 344-363.

[15] Rashmi Dubey, Jiayu Zhou, Yalin Wang, Paul M. Thompson, Jieping Ye "Analysis of sampling techniques for imbalanced data: An $n = 648$ ADNI study" *NeuroImage 87* (2014) 220-241.

[16] Huaxiang Zhang, Mingfang Li "RWO-Sampling: A random walk over-sampling approach to imbalanced data classification" *Information Fusion* 20 (2014) 99-116.

[17] Myoung-Jong Kim, Dae-Ki Kang, Hong Bae Kim "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction" *Expert systems with Application* 42 (2015) 1074-1082.

[18] Kyoham Shin, Jongmin Han, Seokho Kang "MI-MOTE: Multiple imputation-based minority oversampling technique for imbalanced and incomplete data classification" *Information Science* 575 (2021) 80-89.

[19] Ankit Vijayvargiya, Chandra Prakash, Rajesh Kumar, Sanjeev Bansal, Joao Manuel R. S. Tavares "Human knee abnormality detection from imbalanced sEMG data" Biomedical Signal Processing and Control 66 (2021) 102406

[20] Justin Engelmann, Stefan Lessmann "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning" *Expert systems with Application* 174 (2021) 114582.

[21] D.Devi, B.Purkayastha, "Redundancy-driven modified tomek-link based under sampling: A solution to class imbalance"*PatternRecognit. Lett.*93 (2017) 3–12.

[22] C.Bunkhumpornpat, K.Sinapiromsaran, "Dbmute: density-based majority under-sampling technique" *Knowl. Inf. Syst.* 50 (3) (2017) 827–850.

[23] Hualong Yu, Jun Ni, Jing Zhao "ACO Sampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data" Neurocomputing 101 (2013) 309-318.

[24] Mikel Galar, Alberto Fernández, EdurneBarrenechea, Francisco Herrera "EUSBoost: Enhancing ensembles for highly imbalanced datasets by evolutionary undersampling" *Pattern Recognition* 46 (2013) 3460-3471.

[25] Peng Cao, Jinzhu Yanga, Wei Li, Dazhe Zhao, OsmarZaianec "Ensemble-based hybrid probabilistic sampling for imbalanced data learning in lung nodule CAD" *Computerized Medical Imaging and Graphics* 38 (2014) 137-150.

[26] J. Hoyos-Osorio, A. Alvarez-Meza, G. Daza-Santacoloma, A. Orozco-Gutierrez, G. Castellanos-Dominguez "Relevant information undersampling to support imbalanced data classification" *Neurocomputing* 436 (2021) 136-146.

[27] Zhaozhao Xu, Derong Shen, TiezhengNie, Yue Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data", *Journal of Biomedical Informatics*, 107(2020), 103465.

[28] José A. Sáez, Julián Luengo, Jerzy Stefanowski, Francisco Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering", *Information Sciences*, 291(2015) 184-203.

[29] Ming Hao, Yanli Wang∗, Stephen H. Bryant "An efficient algorithm coupled with synthetic minority over-sampling technique to classify

imbalanced PubChem BioAssay data"*Analytica Chimica Acta* 806 (2014) 117-127.

[30] Yuqiang Fan, Xiaoyu Cui, Hua Han, Hailong Lu "Chiller fault diagnosis with field sensors using the technology of imbalanced data" *Applied Thermal Engineering* 159 (2019) 113933.

[31] Amir BahadorParsaa, HomaTaghipoura, Sybil Derribleb , Abolfazl (Kouros) Mohammadianc "Real-time accident detection: Coping with imbalanced data" *Accident Analysis and Preventation* 129 (2019) 202-210.

[32] Hong-Jie Dai, Chen-Kai Wang "Classifying adverse drug reactions from imbalanced twitter data" *International Journal of Medical Informatics* 129 (2019) 122-132 2021.

[33] Asniar, Nur UlfaMaulidev, KridantoSurendro "SMOTE-LOF for noise identification in imbalanced data classification" *Journal of King Saud University – Computer and Information Sciences,* (2021).