International Journal of Engineering &Technology, *8 (3) (2019) 390-397*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A review on data preprocessing methods for class imbalance problem

Haseeb Ali [1] *, Mohd Najib Mohd Salleh [1], Kashif Hussain [2], Arshad Ahmad [3], Ayaz Ullah [3], Arshad Muhammad [4], Rashid Naseem [5], Muzammil Khan [6]

[1] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
[2] Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Cheng-du, Sichuan, China
[3] Department of Computer Science, University of Swabi, Swabi, KPK, Pakistan
[4] Faculty Computing and Information Technology, Sohar University, Oman
[5] Department of Computer Science, City University of Science and Information Technology, Peshawar, KPK, Pakistan
[6] Department of Computer and software Technology, University of Swat, KPK, Pakistan
**Corresponding author E-mail: najib@uthm.edu.my*

Abstract

Data mining methods are often impaired by datasets with desperate nature. Such real-world datasets contain imbalanced data distri-butions among classes, which affects the learning process negatively. In this scenario, the number of samples pertaining to one class (majority class) surpasses adequately the number of samples of other class (minority class) – resulting in ignorance of the minority class by classification methods. To address this, various useful approaches related to data preprocessing are considered mandatory for developing an effective model by using contemporary data mining algorithms. Oversampling and undersampling are two of the fundamental approaches for preprocessing data in order to balance the distribution among dataset. In this study, we thoroughly discuss about the preprocessing techniques and approaches, as well as, challenges faced by researchers to overcome the weaknesses of resampling techniques. This paper highlights the basic issues of classifiers, which endorse bias for majority class and ignore the minority class. Additionally, we synthesize viable solutions and potential suggestions on how to handle the problems in prepro-cessing of data effectively, also present open issues that call for further research.

Keywords: *Imbalanced Data; Re-Sampling; Majority Class; Minority Class; Oversampling.*

1. Introduction

Applications of machine learning and data mining infer through an important challenge that is how to win desired classification accuracy from the data which is highly skewed in nature. When class distribution in a particular training dataset is imbalanced, the training of a learning model becomes even difficult. Imbalanced data problem occurs due to unequal distribution of instances within the classes in a dataset, like one class having a very large number of instances and other class has considerably fewer. The class represented by small number of instances is actually an important class and usually ignored by classifiers, while the classifiers present accuracy by addressing entirely as majority class. In most of the data mining algorithms, uncommon examples are hard to identify than common examples [1, 2]. That said, the ability of developing good model significantly depends on its ability to learn from highly imbalanced data, because the cost of misclassification of minority class is much higher than the cost of misclassifying majority class [3, 4]. Real-world applications frequently encounter this problem like in medical diagnosis, e-mail filtering, financial crisis prediction, fraud detection [5-7]. Misclassification of minority class influences the higher cost in classifying the fraudulent transactions, if the classifier declared a fraud transaction as normal.

To solve the imbalanced data problems, a variety of approaches and methods are proposed in literature, which can be separated into three types: data-level methods, algorithmic-level methods, cost-sensitivity methods [8, 9]. Data-level methods are pre-processing of data before the learning process or constructing any classifier, in which resampling of data is performed externally, to balance the ratio of instances in minority and majority class. Algorithmic-level methods are the creation or modification of algorithms in which minority class is considered. These methods reinforce the learner towards minority class, not allowing to bias for majority class [10]. In cost-sensitivity methods, the cost of misclassification is reduced as well as total cost of errors is minimized [11]. Pre-processing methods are mainly focused in this paper as preprocessing can be performed independently without considering any classifier [12]. A comparative study of various renowned techniques, approaches and combinations of methods of data preprocessing with ensemble classifiers that perform better than other methods are discussed in this study.

Data-level methods are also known as external-level methods, as they manipulate the training data externally. It is done by re-sampling of data externally in order to balance the distribution of instances in majority and minority classes. The most common approaches to balance the ratio of instances in all classes are undersampling and oversampling [13]. Removing the samples from majority class is known as undersampling or generating the artificial data in minority class to enhance the number of samples for balancing the ratio that can be used for learning, so-called oversampling. Random elimination of samples from the majority class can result in loss of the potential data which is useful for learning process. Similarly, duplicating the samples in minority regions randomly may cause overfitting, or sometimes, it may generate noise in the data impacting the classifier negatively. It is because, external or data-level methods can be composed of informal methods which consider the distribution of samples, However, the random methods only determine the samples to be duplicated or eliminated [12]. The informed methods take into account the critical area of the input space, like safe areas [14], sparse areas [15] or areas which are closer to decision boundary [16]. Consequently, generation of noise can be avoided so that the informed methods may tackle the imbalances within the classes.

Besides, preprocessing techniques are more versatile and can be applied globally, but the algorithmic approaches are more confined to specific classifiers. On the other hand, the cost sensitive methods are problem specific, also require to be implemented by the classifier [9]. There is plenty of work performed by research community on preprocessing of data to overcome the imbalanced class issues because it enhances the learnability of data by any classifier. Various techniques and methods which address these problems have been proposed for their solutions. This research study covers most of the challenges faced by researchers in preprocessing of data and solutions proposed by them. In this research, we will focus on answering the following research questions (RQs):

RQ 1: Why we need resampling methods for solving the imbalanced data problems?

RQ 2: What are the problems faced when resampling data? What are solutions?

RQ 3: Where is the research trend in this particular research area? Moreover, what are the future prospects?

The rest of this paper is composed as our upcoming section presents the research methodology occupied for this study. In Section 3, a brief description of the techniques and methods proposed for the solutions of imbalanced data is presented. Section 4 highlights the research gaps and evaluation matrices. In section 5, discussion on this research is presented, while Section 6 concludes research and highlights the potential future directions in current line of research.

2. Research methodology

An organized review of literature, which exist in the particular area, can be directed in binary ways: manually [17] and automatically [18]. In this research, this study preferred manual strategy since the automatic approach concurred some limitations and for this study, automatic search engines were not feasible [19]. Based on research methodology given in [20], this study also performed in two-stage research process with the purpose of compiling the most important and relevant papers which published in last two decades. Seven library databases are used in order to search and collect this most relevant data in the primary stage: Elsevier, Cambridge, IEEEXplore, Springer, Wiley, Sage and ACM. In which social science and most natural science research fields are covered. Keywords used for the studies are most relevant to this paper, listed above, which resulted in the most fruitful outcome.

3. Resampling techniques

Preprocessing of data by resampling techniques, in order to balance the ratio of instances present in majority and minority classes, can be divided in to three categories: oversampling, undersampling, and hybrid solutions (Fig. 1) [9]. Below, we discuss these in detail.

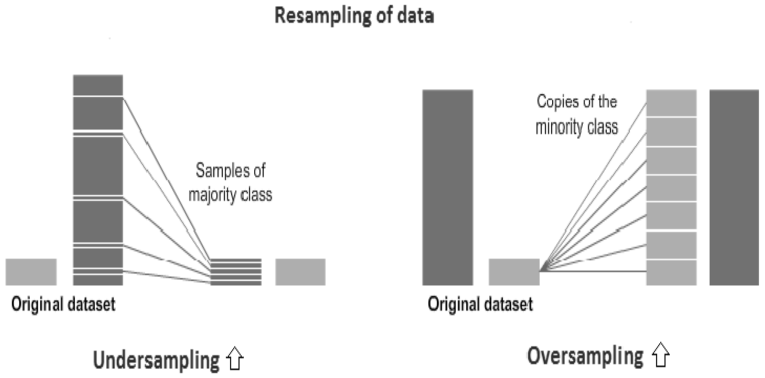

Fig. 1: Preprocessing of Imbalanced Data.

3.1. Under sampling techniques

Random undersampling is the basic and simplest method for resampling of an imbalanced dataset, in which the samples of majority class are randomly eliminated from the class to balance the distribution of classes for learning process (Fig. 2).

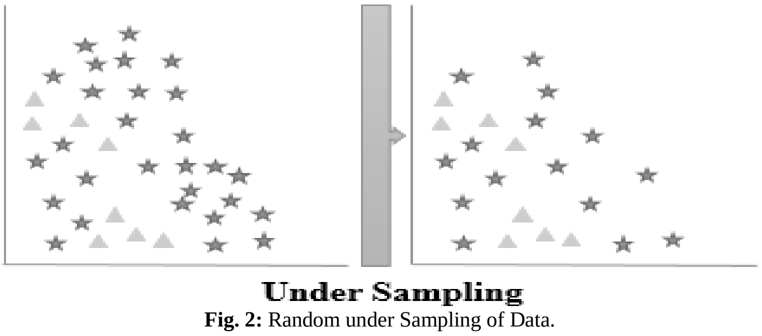

Fig. 2: Random under Sampling of Data.

This method is simple and comparatively less complex than other methods or oversampling of data. However, because of its significant drawback that it losses potential samples that can be useful in learning process, or removing the samples randomly also removes useful data with it. Therefore, the researchers and practitioners usually prefer oversampling techniques to undersampling. Random undersampling is also widely used due to its simplicity and its weakness are overcome when it is used with other methods [12], [13].

Inverse Random Undersampling (IRUS) [21] uses inverse (ratio of unbalance cardinality) approach to alleviate the imbalanced data problem. It creates a number of different training sets by severely undersampling the majority class. Then, decision boundary is discovered for each training set which separates the majority class from minority class. It is also useful for multi-label classification. An ant optimization based undersampling method ACOSampling is proposed in [22], which is essentially used for classification of imbalanced DNA microarray data. In this method, feature selection technique is used initially to remove the noisy genes in data. Then original training data is divided randomly and repeatedly into two groups: training and validation sets. One modified ACO model is managed to filter majority samples having less potential or no information and then search equivalent best training sample subset. The majority samples with high frequencies are combined with all minority samples finally to create balance training set. A new Noise-Filtered Undersampling Scheme is proposed which incorporates a noise filter before resampling [23]. It performs significantly for solving imbalance classification problem.

Clustering based undersampling is proposed in [24], where clustering is used initially. In this method, cluster numbers in majority class are maintained as equal number of data points in minority class. Two strategies are employed by this method: first is the clusters centers present the majority class, and secondly nearest neighbours of the cluster centers is used. Then deletion or addition of five to ten cluster centers in the majority class takes place according to the further study for examining. It is evaluated that nearest neighbours of cluster centers approach used by this method is the rightest choice for the imbalanced data undersampling. A fast clustering-based under-sampling method is proposed to solve the problem of imbalanced data distribution [25]. Fast-CBUS has various significant characteristics, time complexity of this method is confined to the number of samples of minority class. Moreover, each cluster is trained by a specific classifier. It takes into consideration the distinguished problem of undersampling which is loss of information in majority class samples. In all sets of experiments Fast-CBUS is considered to be only method which in on Pareto frontier.

3.2. Oversampling techniques

The simplest technique used for oversampling is random oversampling (Fig. 3). Random oversampling selects the samples randomly and generates new samples in minority class. Although, it increases the number of samples, but new samples are often quite similar to the original samples which may result in overfitting as the generated samples are exact replication of samples [12]. Synthetic Minority Oversampling Technique (SMOTE) is proposed by Chawla et al. in 2002 to overcome this problem [13]. In this technique new samples are generated by linear interpolation of an inferior sample with their randomly selected k-Nearest Neighbors (kNN). This technique generates new samples without examining the majority class samples, which may induce overlapping between majority and minority samples, causing over-generalization along with amplifying the noise. Despite these drawbacks, research community widely adopts SMOTE due to its simplicity [26]. Various extensions and modification of this technique have been proposed to eliminate its weaknesses. Some of the used filtering-based methods for avoiding the noise are SMOTE-TL and SMOTE-EL [27].

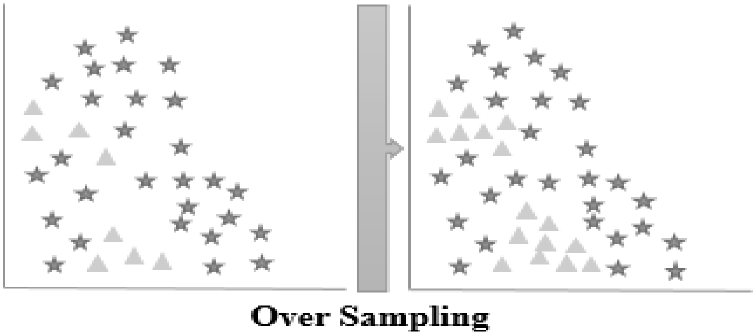

Fig. 3: Random Oversampling of Data.

Over-generalization occurs on account of intense imbalance ratio as the minority class samples sparse into majority class region creates noise and overlapping when generated new samples. Borderline-Smote [16] presents this solution by identifying the borderline among the two classes. It oversamples the minority samples only on the border line. Safe-level SMOTE [14] addresses this issue as generating the synthetic samples into safest level. For each minority sample, it calculates a safe-level value defined as number of minority samples between its kNN and the new samples generated.

ADASYN [28] allots weights to those samples of minority class whose neighbourhood contains more majority class samples because they have more chances of being neglected by the classifier. Therefore, it assigns weights to those samples that have higher chances to be oversampled, hence learnability of classifier is improved. But, Border-line SMOTE and ADASYN fail to find all minority samples on decision boundary. Extension of ADASYN is KernelADASYN [29] which is based on estimation of kernel density of imbalanced data problem. While considering probability destruction of dataset, it generates new synthetic samples. Moreover, it assigns more weights to the samples that are hard-to-learn, which enhance the classifier performance. Learning from minority class becomes challenging in such imbalanced datasets, in which samples of minority class are distributed sparsely in feature space, as compared to a large amount of majority class samples. Another oversampling method is Random-SMOTE [30] proposed for certain problem. Random-SMOTE generates synthetic samples randomly in minority class samples space and reduces the sparseness of the minority samples. Even though, it is simple method yet effective as compared to random oversampling method.

Modified synthetic minority oversampling technique (MSMOTE) [31] is the modification in SMOTE. In this method, the calculation of distances among all samples of minority class is divided into groups: safe, borderline and safe samples. DBSMOTE [26] is density-based method, in which DB-SCAN algorithm is used to find random clusters of different shapes. It generates synthetic samples from each minority class samples along a shortest path to a clusters pseudo-centroid.

Majority Weight Minority Oversampling Technique MWMOTE [32] addresses this problem by proposing a two-step process, finds minority and majority border samples than assigns weights to minority samples based on their Euclidean distances to majority samples. Those which have more weights are more important and higher is the chance to be oversampled. However, the small disjoints which are far from majority samples remain neglected, which may contain useful information for minority class. Similarly, in RWO-Sampling [33], a random walk oversampling generates synthetic samples by random walking from the real data. Here, standard deviation of the generated samples and expected average is equivalent to original minority class data. After generating synthetic samples, it also expands the minority class boundary. However, this method increases the chances of overfitting.

Enhanced Minority Oversampling Technique (EMOTE) [34], balances the dataset to enhance the performance of classifier. The approach used to balance the dataset in this technique is the adjustment of incorrectly classified examples of minority class into correctly classified examples by generating new synthetic examples from their nearest neighbour. G-SMOTE [35] is a synthetic minority oversampling technique based on Gaussian Mixture Model (GMM) for imbalanced learning. This technique determines the outliers from minority class samples by applying GMM, which also eliminates the synthetic samples from majority class. Unlike existing sampling techniques which generate sample in a linear sampling space, G-SMOTE generates new samples in a high dimensional feature space.

Synthetic Informative Minority Oversampling (SIMO) and Weighted (W-SIMO) [36] techniques are combined with SVM in order to improve learning efficiency when employed with an imbalanced dataset. In this approach, informative minority samples are selected, which are close to the decision boundary of SVM. In this approach, selected data points are used to generate synthetic minority samples then a new SVM model is developed for the oversampled dataset. On the other hand, W-SIMO identifies the incorrectly classified informative minority samples and again oversample them with a higher degree. It focuses more on misclassified informative minority samples. However, computational time is the core limitation of this technique when applied to large datasets.

Samples of minority classes that are boundary classes in real world applications usually possess the primary interest, such as in product inspection, quality control and disease diagnosis. Consequently, performance of minority class should be improved in context of ordinal regression. In order to tackle imbalanced ordinal regression problem, a technique based on generation direction aware synthetic minority oversampling is proposed (SMOR) [37]. A selection weight is assigned by SMOR that is used to make synthetic samples for every individual generation direction. This technique reduces the distortion factor in sample structure and improves ordering of boundary classes without damaging the ordering of majority class. In Radial Based Oversampling for Noisy Imbalanced Data Classification [38], Radial-Based Oversampling (RBO) method finds such appropriate areas in minority class for the generation of synthetic samples based on the imbalanced ratio using radial basis functions. As well as, this method removes noise from dataset which considerably affects the classifier performance. Areas which posed difficulty to the classifier are detected by calculation of joint potentials in each area of feature space and new samples are generated in this specific region. This approach tends to be more guided for oversampling process as it indicates the safest places for generating new samples and also avoids class overlapping. Density based approach is also used in this technique which handles noise and makes this method more robust.

3.2.1. Clustering-based oversampling

Aforementioned techniques deal with between-class imbalance, but there are techniques proposed in recent five years which at the same time also reduce within-class imbalance. These are clustering based oversampling techniques, in which input space is partitioned at the first stage and then resampling methods are applied on the dataset in order to balance the ratio of samples in each cluster.

Cluster-SMOTE [39] is the first technique proposed in last decade in which it applies k-means algorithm and then generates synthetic samples by using SMOTE. This technique does not determine the optimal number of clusters, also does not specify the number of samples generated in each cluster. Adaptive Semi-Unsupervised Weighted Oversampling (A-SUWO) [40] belongs to the same category, which uses hierarchical clustering and oversample hard-to-learn instance which are closer to decision boundary by assigning them weights accordingly and also takes into account the small concepts of minority class. Self-Organizing Map Oversampling (SOMO) [41] uses a self-organizing map to converts the input data into a two-dimensional space. Where data is generated in identified effective and safe areas. SMOTE is applied in order to overcome the imbalance issue between as well as within the classes by generating the samples within minority cluster found in lower dimensional space also between neighbouring clusters. k-means SMOTE [42] avoids noise generation and reduces imbalanced ratio effectively between the classes and within-class. This method uses k-means to cluster the data and it focuses more on critical areas of the input space for data generation. Here, class imbalance is reduced by assigning more weights to sparse minority clusters. However, detecting proper value for k is significant for the effectiveness of k-means SMOTE, as it persuades the number of minority clusters.

3.3. Hybrid techniques

Neighbourhood Balanced Bagging (NBBag) is a combination of original sampling techniques and creates an effective resampling technique to handle the imbalanced data [27]. Combination of Practical Swarm Optimization (PSO) and SMOTE including C5 decision tree, logistic regression (LR) are used for resampling of an imbalanced data and this hybrid algorithm PSO+SMOTE+C5 [58] is significantly effective for breast cancer patients 5-year lastingness. Combination of SMOTE+PSO+RBF (aided radial basis function classifier) [50] proposed which have been shown to be efficient technique for solving imbalanced data problem. After the evaluation of this technique by using different matrices fusion of SMOTE+PSO+ RBF can be applied on real world applications concurring imbalanced data. However, it does not produce satisfactory results for highly imbalanced datasets. Extension of SMOTE and Iterative–Partitioning Filter (SMOTE-IPF) handles noises and regular the class boundaries [39]. This hybrid solution eliminates noise from the data. Similarly, IPF is also used with random undersampling techniques like Easy Ensemble (EE), EE-IPF work effectively and removes noise from both majority and minority classes [43]. A new hybrid evolutionary preprocessing method is generally proposed in [44] to resample the imbalanced datasets, which is based on both undersampling and oversampling techniques. In this hybrid method, new samples are generated in minority class on the base of fuzzy logic. An evolutionary computational method consists of cross generational elitist selection, Heterogeneous recombination and Cataclysmic mutation (CHC) is applied to undersample the majority class samples and newly generated minority samples. The main weaknesses of oversampling and undersampling that is over generalization, and removal of useful data respectively is avoided by this hybrid method.

The hybrid of combining oversampling and undersampling as well as integration of resampling techniques with ensemble classifiers appears to be effective and enhance the performance significantly. A new method, Cluster Based Instance Selection (CBIS) [45] uses undersampling approach in which clustering analysis groups the majority class instances into the subclasses in dataset and instance selection remove the unrepresented data instances from each subclass. Another application of undersampling adapted with one class SVM have been recently proposed for data overlapping and imbalanced problem [46]. Tomek-link undersampling is used to eliminate overlapped, redundant and borderline instances from the majority class and overcome the imbalances and overlapped cases. Proposed method enhances the accuracy of classifying minority class and accuracy of majority class by elimination of instances to a significant extent. This method is effective for binary and multiclass datasets.

Synthetic minority oversampling techniques is combined with ensemble-based models, boosting to enhance the performance of SMOTE to propose SMOTEboost [47]. Likewise, oversampling technique is also combined with bagging to propose over-bagging [48], this model is similar to boosting. These models used for oversampling of data generate new samples by considering the minority class. Random Undersampling with boosting, RUSboost [49] is also similar to UnderBagging [50], in which instances from the majority class are removed randomly to balance the distribution of samples in both classes. In each iteration, it takes into account the majority class. Both the methods outperformed their standard techniques. For comparison purpose, Table 1 and Table 2 present pros and cons of the above discussed techniques.

Using the bagging method along with an oversampling technique, a new ensemble method Bagging of Extrapolation Borderline-SMOTE SVM [51] is proposed to integrate borderline information in order to deal with imbalanced data problem. An adaptive sampling method extrapolation Borderline-SMOTE is applied and former imbalanced data set is aggregated by bootstrapping. Bagging enhanced the aptitude of model's generalization and also avoid overfitting.

A recent application of boosting algorithm (adaboost) for imbalanced data classification is, cluster-based undersampling with boosting, CUSBoost [52] which employs clustering on majority class samples and then applies random undersampling. This way, it allows to (Adaboost) boosting algorithm to use samples from all areas of dataset. This algorithm outperforms RUSBoost and SMOTEboost when the dataset possesses higher imbalance ratio. Another hybrid method HUSboost based on ensemble approach with sampling techniques is proposed [53] to deal with imbalanced data problem. Undersampling is used with boosting algorithm in this hybrid method which follows up in three steps. The first step is data cleaning, removal of noisy samples by using Tomel-Links. Second, data balancing in which random undersampling technique is applied on majority class samples to create several balanced subsets. Third step is classification, in which Adaboost with decision tree (CART), Adaboost with SVM and Random Forest (RF) are implemented and soft voting approach is used to get combined results.

4. Research gap

Preprocessing of the data has its own significance in every field of artificial intelligence especially in data mining; such as, many real-world applications endure imbalanced data problems. Research community focuses more on resampling of data in preprocessing although they face many challenges in this area. Researchers have overcome many weaknesses in techniques and approaches used for re-sampling. However, still some drawbacks remain uncovered. In MWMOTE, minority samples that are hard-to-learn, chosen for oversampling which are closer to the majority samples and decision boundary, but small concepts which are far from the majority samples remains unprocessed. This may contain important information for minority class. Adaptive semi-unsupervised learning resolves class imbalance problem by considering small concepts and oversampling, though the model is more complex compared to other, which are parallel to this method. K-means SMOTE is simple and fast method for oversampling, which generates synthetic samples in safe area and also avoids noise and over fitting. However, k-means clustering is used for between the classes and with-in the class's imbalance problem, which need proper value of k for the efficiency of the proposed solution.

4.1. Overview on the advantages and limitations

Section 3 described most of the preprocessing techniques and methods which have been successfully used for imbalanced data problems. Identified pros of the data-level techniques, are briefly described in following points.

- Learning algorithms provide significant results, if preprocessing to balance the datasets is performed with vigilance.
- Pre-processing methods give objective oriented output, because they are problem specific particularly.
- Pre-processing techniques are more versatile and can be applied globally since these are not specific to any classifier.
- Pre-processing techniques are non-effective of run and trial overheads for the cost evaluation approaches.

Hence, pre-processing methods and techniques for imbalanced dataset problem are proven to be effective however some limitations are also identified, major limitations of them are as follows:

- Pre-processing techniques must be repeatedly applied that can cause overfitting and classified poorly by a classifier and result in expensive computational cost overhead to learning processes.
- Some of the useful training instances can be possibly removed from dataset like in undersampling for an effective learning process.

Table 1: Comparison of Oversampling Techniques

Oversampling Technique	Ref.	Pros	Cons
SMOTE	[13]	It generates synthetic samples by kNN.	It may cause over-fitting and noise.
Borderline-SMOTE	[16]	It creates a borderline b/w majority and minority samples.	Does not identify samples near decision boundary.
ADASYN	[28]	Assigns weights to minority samples.	Fails to discover all minority samples on decision boundary.
Safe-level SMOTE	[14]	Generates samples in safe areas.	Far from decision boundary and neglected by classifier.
MSMOTE	[31]	Reduces the noise.	It cannot identify the priorities of important features.
DBSMOTE	[54]	Generates samples based on density.	Not prior to borderline samples.
MWMOTE	[32]	Over-samples the samples which are hard-to-learn.	Neglects small concepts which are far from majority class.
RWO-Sampling	[33]	Size of generated samples remain same to original samples. Increases decision boundary for minority samples.	Likelihood to overfitting.
EMOTE	[55]	Correctly classified minority samples	Overpopulate minority regions
G-SMOTE	[56]	Generate samples in high dimensional feature space	Highly complex in computation
SIMO and W-SIMO	[57]	Identify the informative minority samples for oversampling	Computationally expensive

Kernel-ADASYN	[29]	Chooses hard-to-learn samples for oversampling and avoid noisy.	Cannot overcome the over-lapping of majority minority samples.
Cluster-SMOTE	[39]	Clusters the input space then oversamples.	Does not determine the optimal number of clusters.
A-SUWO	[30]	Considers the small concepts of minority samples.	Complex model.
SOMO	[40]	Effective for high dimensional data.	Computationally expensive
K-Means SMOTE	[42]	Effective for b/w class and within-class imbalance ratio.	Proper value of k should be determined.

4.2. Evaluation metrics

Performance of re-sampling techniques and approaches are measured by some common matrices, but in case when distribution of the class is not uniform, all the metrics are not suitable. However, to cope with imbalanced data specifically some metrics have been developed and employed [60]: Precision (1), Recall (2), $F_{measure}$ (3), G_{mean} (4), and Area Under Receiving operator Characteristics graph (AUC). By using Precision and Recall, we can calculate F-measure, in confusion matrix, while majority instances are referred as negative (N), and minority instances as positive (P).

Precision evaluates the classifier’s exactness, which means the total number of samples that labelled correctly as positive (minority) that are positive in actual. Classifier’s completeness is evaluated by Recall in a way that the number of positive samples are classified as positive correctly. The virtual importance between recall and precision is adjusted for $F_{measure}$ by the parameter β. G_{mean} validates the accuracy considering both classes. Consequently, if the classifier favoured the negative (majority class) samples and ignores the positive (minority class) samples, then a low G_{mean} will be obtained by the classifier.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F_{measure} = \frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall \times Precision} \tag{3}$$

$$G_{mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{4}$$

AUC is suitable for performance evaluation for the class imbalance problem since it is not dedicated to the distribution of two classes in any dataset. AUC is area under the ROC which is acquired by scheming the False Positive Rate (FPR) divided by True Positive Rate (TPR) (5). Defined as following where number of negative (majority) instances are denoted by N_n and number of positive (minority) instances are referred by N_p.

$$TPR = \frac{TP}{N_p}, FPR = \frac{FP}{N_n} \tag{5}$$

Table 2: Comparison of under-Sampling and Hybrid Techniques

Under-sampling Technique	Ref.	Pros	Cons
IRUS	[21]	Improves multi-label classification accuracy	Severely removing majority samples also removes useful data.
ACO-Sampling	[22]	Effectively reduces imbalance problem in DNA microarray data.	Can improve for other microarray datasets.
Clustering Based Undersampling	[24]	Balances the distribution based on clusters.	Removes clusters which may contain potential majority samples.
A Noise Filter Undersampling	[23]	Reduces noise before preprocessing.	Only significant for undersampling.
Fast-CBUS	[25]	Reduce the chance of losing potential data	Optimal value for k could enhance the performance
Hybrid Techniques			
SMOTE + PSO + C5	[48]	Best for five years lastingness for breast cancer patient.	Can be improved for other cancer datasets.
SMOTE + RBFN	[50]	Higher the IR value of dataset tends to better result.	Needs large storage space.
SMOTE + IPF	[59]	Excludes noise from data and borderline problems.	Effective for small size sample.
NBBAG	[31]	Better results from identical bagging oversampling methods.	Impact of high cost
RUS-IPF	[43]	Removes noise from the data.	Sample size impact on the performance.
Borderline-SMOTE SVM	[51]	Enhance model generalization	Computational time is high
CUSBoost	[52]	Higher imbalance ratio higher will be accuracy	Loss of potential data by series of elimination process
CBIS	[45]	Determine unrepresented samples in data and remove them	Complexity is higher that could be reduced

5. Discussion

This research covers the literature for the preprocessing of the data and determined fruitful approaches for learning from imbalanced data and the importance of minority class in real world domains. In particular cases, algorithmic level approaches may be effective, but they have significant drawback of being algorithm specific. It is better to classify datasets, which presents distinct characteristics by different classifiers, though this is relatively difficult to modify the classifier according to a particular imbalance problem. Contrarily, pre-processing techniques or data level approaches are more robust and effective as these are independent of classifiers and flexible. After applying data level techniques, the standard classifiers can learn data more effectively hence their efficiency enhances. It is, therefore, this research purely focused on preprocessing of data.

This study found the ratio of oversampling methods proposed by authors much more than under sampling techniques. As discussed earlier, the undersampling approaches may lead to the loss of potential data which could be harmful in learning process. Although, many researchers have made ample effort in avoiding this loss and proposed remarkable methods. However, oversampling methods are more practiced by researchers due to their simplicity and efficiency recorded in results. The major problems faced in oversampling by researchers are related to overlapping, overfitting of synthetic samples and within-class imbalances which is reduced by using clustering approach. Also, many other solutions regarding these problems are still on research line.

6. Conclusion

This paper covered the literature and found the theoretical concepts of resampling of data in order to achieve balanced distribution. This study presented numerous challenges and hurdles in the way of researchers for handling imbalanced data. It is necessary to balance the imbalanced class with the help of effective methods while considering the cost factor. The right selection of the classifier methods with performance evaluation metrics should be applied in order to accomplish better results. In conclusion, this study found the k-means SMOTE more efficient for between the classes and within-class imbalance problem, if the proper value k can be defined. The future work can be suggested to find better alternative or appropriate value of k for the influence of minority class clusters numbers. Also, more focus can be put on undersampling techniques to improve by hybridizing with any other suitable method.

Acknowledgement

University Tun Hussein Onn Malaysia (UTHM) supports this research; the authors would like to express a special gratitude and thanks to research management center (RMC) for providing Postgraduate Incentive Research Grant, Vote No.H334.

References

[1] Sun, Y., Wong, A. K. C., and Kamel, M. S., "Classification of imbalanced data: a review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 22, no. 4, pp. 687–719, 2009. https://doi.org/10.1142/S0218001409007326.

[2] Batista, Gustavo E. A. P. A., et al., "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *Sigkdd Explorations*, vol. 6, no. 1, pp. 20–29, 2004. https://doi.org/10.1145/1007730.1007735.

[3] Domingos, Pedro M., "MetaCost: A General Method for Making Classifiers Cost-Sensitive," *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155–164, 1999. https://doi.org/10.1145/312129.312220.

[4] Ting, Kai Ming., "An Instance-Weighting Method to Induce Cost-Sensitive Trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, 659–665, 2002. https://doi.org/10.1109/TKDE.2002.1000348.

[5] Raskutti, Bhavani, and Adam Kowalczyk, "Extreme Re-Balancing for SVMs: A Case Study," *Sigkdd Explorations*, vol. 6, no. 1, 60–69, 2004. https://doi.org/10.1145/1007730.1007739.

[6] Wu, Gang, and Edward Y. Chang, "*Class-Boundary Alignment for Imbalanced Dataset Learning*," *ICML 2003 Workshop on learning from imbalanced data sets II, Washington, DC*, 49-56, 2003.

[7] Yan, Rong, et al., "On Predicting Rare Classes with SVM Ensembles in Scene Classification," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*, vol. 3, pp. 21–24, 2003.

[8] Beyan, Cigdem, and Robert B. Fisher, "Classifying Imbalanced Data Sets Using Similarity Based Hierarchical Decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653–1672, 2015. https://doi.org/10.1016/j.patcog.2014.10.032.

[9] Galar, Mikel, et al., "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *Systems Man and Cybernetics*, vol. 42, no. 4, pp. 463–484, 2012. https://doi.org/10.1109/TSMCC.2011.2161285.

[10] Joshi, Mahesh V., et al., "Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements," *Proceedings 2001 IEEE International Conference on Data Mining*, 257–264, 2001.

[11] Ling, Charles X., and Victor S. Sheng, *Cost-Sensitive Learning and the Class Imbalance Problem*, University of Western Ontario, 2008.

[12] Chawla, N., et al., "Special issues on learning from imbalanced data sets," *ACM SigKDD Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004. https://doi.org/10.1145/1007730.1007733.

[13] Chawla, Nitesh V., et al., "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002. https://doi.org/10.1613/jair.953.

[14] Bunkhumpornpat, Chumphol, et al., "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem," *PAKDD '09 Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 475–482, 2009. https://doi.org/10.1007/978-3-642-01307-2_43.

[15] Nickerson, Adam, et al., *Using Unsupervised Learning to Guide Resampling in Imbalanced Data Sets*, AISTATS, 2001.

[16] Han, Hui, et al., "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," *International Conference on Intelligent Computing*, pp. 878–887, 2005. https://doi.org/10.1007/11538059_91.

[17] Budgen, David, and Pearl Brereton, "Performing Systematic Literature Reviews in Software Engineering," *Proceedings of the 28th International Conference on Software Engineering*, pp. 1051–1052, 2006. https://doi.org/10.1145/1134285.1134500.

[18] Petersen, Kai, et al., "Systematic Mapping Studies in Software Engineering," *EASE'08 Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, pp. 68–77, 2008.

[19] Brereton, Pearl, et al., "Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain," *Journal of Systems and Software*, vol. 80, no. 4, pp. 571–583, 2007. https://doi.org/10.1016/j.jss.2006.07.009.

[20] Govindan, M. E.Kannan, and Martin Brandt Jepsen., "ELECTRE: A Comprehensive Literature Review on Methodologies and Applications," *European Journal of Operational Research*, vol. 250, no. 1, pp. 1–29, 2016. https://doi.org/10.1016/j.ejor.2015.07.019.

[21] Tahir, M. A., Kittler, J., & Yan, F., "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognition*, vol. 45, no. 10, pp. 3738–3750, 2012. https://doi.org/10.1016/j.patcog.2012.03.014.

[22] Yu, H., Ni, J., & Zhao, J., "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data," *Neurocomputing*, vol. 101, pp. 309–318, 2013. https://doi.org/10.1016/j.neucom.2012.08.018.

[23] Kang, Q., Chen, X., Li, S., & Zhou, M., "A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 47, no. 12, pp. 4263–4274, 2017. https://doi.org/10.1109/TCYB.2016.2606104.

[24] Lin, W.-C., Tsai, C.-F., Hu, Y.-H., & Jhang, J.-S., "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017. https://doi.org/10.1016/j.ins.2017.05.008.

[25] N. Ofek, L. Rokach, R. Stern, and A. Shabtai, “Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem,” *Neurocomputing*, vol. 243, pp. 88–102, 2017. https://doi.org/10.1016/j.neucom.2017.03.011.

[26] He, Haibo, and Edwardo A. Garcia, “Learning from Imbalanced Data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. https://doi.org/10.1109/TKDE.2008.239.

[27] Błaszczyński, Jerzy, and Jerzy Stefanowski, “Neighbourhood Sampling in Bagging for Imbalanced Data,” *Neurocomputing*, vol. 150, pp. 529–542, 2015. https://doi.org/10.1016/j.neucom.2014.07.064.

[28] He, Haibo, et al., “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328, 2008. https://doi.org/10.1109/IJCNN.2008.4633969.

[29] Tang, Bo, and Haibo He., “KernelADASYN: Kernel Based Adaptive Synthetic Data Generation for Imbalanced Learning,” *2015 IEEE Congress on Evolutionary Computation (CEC)*, pp. 664–671, 2015. https://doi.org/10.1109/CEC.2015.7256954.

[30] Y. Dong and X. Wang, “A new over-sampling approach: Random-SMOTE for learning from imbalanced data sets,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7091 LNAI, pp. 343–352, 2011. https://doi.org/10.1007/978-3-642-25975-3_30.

[31] Hu, S., Liang, Y., Ma, L., & He, Y., “MSMOTE: Improving Classification Performance When Training Data is Imbalanced,” *2009 Second International Workshop on Computer Science and Engineering*, vol. 2, pp. 13–17, 2009. https://doi.org/10.1109/WCSE.2009.756.

[32] Barua, S., Islam, M. M., Yao, X., & Murase, K., “MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 2, pp. 405–425, 2014. https://doi.org/10.1109/TKDE.2012.232.

[33] Zhang, H., & Li, M., “RWO-Sampling: A random walk over-sampling approach to imbalanced data classification,” *Information Fusion*, vol. 20, pp. 99–116, 2014. https://doi.org/10.1016/j.inffus.2013.12.003.

[34] S. Babu and N. R. Ananthanarayanan, “EMOTE: Enhanced Minority Oversampling TEchnique,” *J. Intell. Fuzzy Syst.*, vol. 33, no. 1, pp. 67–78, 2017. https://doi.org/10.3233/JIFS-161114.

[35] Z. Tianlun and Y. Xi, “G-SMOTE: A GMM-BASED SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE for IMBALANCED LEARNING,” *arxiv1810.10363v1, a Prepr.*, 2018.

[36] S. Piri, D. Delen, and T. Liu, “A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets,” *Decis. Support Syst.*, vol. 106, pp. 15–29, 2018. https://doi.org/10.1016/j.dss.2017.11.006.

[37] T. Zhu, Y. Lin, Y. Liu, W. Zhang, and J. Zhang, “Minority oversampling for imbalanced ordinal regression,” *Knowledge-Based Syst.*, vol. 166, pp. 140–155, 2019. https://doi.org/10.1016/j.knosys.2018.12.021

[38] M. Koziarski, B. Krawczyk, and M. Woźniak, “Radial-Based oversampling for noisy imbalanced data classification,” *Neurocomputing*, no. 2019, 2019. https://doi.org/10.1016/j.neucom.2018.04.089.

[39] Cieslak, D. A., Chawla, N. V., & Striegel, A., “Combating imbalance in network intrusion datasets,” *2006 IEEE International Conference on Granular Computing*, pp. 732–737, 2006.

[40] Nekooeimehr, I., & Lai-Yuen, S. K., “Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets,” *Expert Systems with Applications*, pp. 405–416, 2016. https://doi.org/10.1016/j.eswa.2015.10.031.

[41] Douzas, G., & Bacao, F., “Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning,” *Expert Systems With Applications*, vol. 82, pp. 40–52, 2017. https://doi.org/10.1016/j.eswa.2017.03.073.

[42] Douzas, G., & Bacao, F., “Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,” *Information Sciences*, vol. 465, pp. 1–16, 2018. https://doi.org/10.1016/j.ins.2018.06.056.

[43] Chen, X., Kang, Q., Zhou, M., & Wei, Z., “A novel under-sampling algorithm based on Iterative-Partitioning Filters for imbalanced classification,” *IEEE International Conference on Automation Science and Engineering (CASE)*, pp. 490–494, 2016. https://doi.org/10.1109/COASE.2016.7743445.

[44] C. F. Tsai, W. C. Lin, Y. H. Hu, and G. T. Yao, “Under-sampling class imbalanced datasets by combining clustering analysis and instance selection,” *Inf. Sci. (Ny).*, vol. 477, pp. 47–54, 2019. https://doi.org/10.1016/j.ins.2018.10.029.

[45] D. Devi, S. K. Biswas, and B. Purkayastha, “Learning in presence of class imbalance and class overlapping by using one-class SVM and undersampling technique,” *Conn. Sci.*, vol. 31, no. 2, pp. 105–142, 2019. https://doi.org/10.1080/09540091.2018.1560394.

[46] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, “A Novel Ensemble Method for Imbalanced Data Learning,” *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–11, 2017. https://doi.org/10.1155/2017/1827016.

[47] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W., “SMOTEBoost: Improving Prediction of the Minority Class in Boosting,” *European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 107–119, 2003. https://doi.org/10.1007/978-3-540-39804-2_12.

[48] Barandela, R., Valdovinos, R. M., & Sánchez, J. S., “New Applications of Ensembles of Classifiers,” *Pattern Analysis and Applications*, vol. 6, no. 3, pp. 245–256, 2003. https://doi.org/10.1007/s10044-003-0192-z.

[49] Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V., & Napolitano, A., “RUSBoost: A Hybrid Approach to Alleviating Class Imbalance,” *Systems Man and Cybernetics*, vol. 40, no. 1, pp. 185–197, 2010. https://doi.org/10.1109/TSMCA.2009.2029559.

[50] Gao, M., Hong, X., Chen, S., & Harris, C. J., “A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems,” *Neurocomputing*, vol. 74, no. 17, pp. 3456–3466, 2011. https://doi.org/10.1016/j.neucom.2011.06.010.

[51] F. Rayhan, S. Ahmed, A. Mahbub, R. Jani, S. Shatabda, and D. M. Farid, “CUSBoost: Cluster-Based Under-Sampling with Boosting for Imbalanced Classification,” *2nd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solut. CSITSS 2017*, pp. 1–5, 2018. https://doi.org/10.1109/CSITSS.2017.8447534.

[52] M. H. Popel, K. M. Hasib, S. A. Habib, and F. M. Shah, “A Hybrid Under-Sampling Method to Classify Imbalanced Data CANDIDATES ’ DECLARATION,” *2018 21st Int. Conf. Comput. Inf. Technol.*, no. May 2018, pp. 1–7, 2018. https://doi.org/10.1109/ICCITECHN.2018.8631915.

[53] M. H. Popel, K. M. Hasib, S. A. Habib, and F. M. Shah, “A Hybrid Under-Sampling Method to Classify Imbalanced Data CANDIDATES ’ DECLARATION,” *2018 21st Int. Conf. Comput. Inf. Technol.*, no. May 2018, pp. 1–7, 2018. https://doi.org/10.1109/ICCITECHN.2018.8631915.

[54] Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C., “DBSMOTE: Density-Based Synthetic Minority Over-Sampling Technique,” *Applied Intelligence*, vol. 36, no. 3, pp. 664–684, 2012. https://doi.org/10.1007/s10489-011-0287-y.

[55] S. Babu and N. R. Ananthanarayanan, “EMOTE: Enhanced Minority Oversampling TEchnique,” *J. Intell. Fuzzy Syst.*, vol. 33, no. 1, pp. 67–78, 2017. https://doi.org/10.3233/JIFS-161114.

[56] Z. Tianlun and Y. Xi, “G-SMOTE: A GMM-BASED SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE FOR IMBALANCED LEARNING,” *arxiv1810.10363v1, a Prepr.*, 2018.

[57] S. Piri, D. Delen, and T. Liu, “A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets,” *Decis. Support Syst.*, vol. 106, pp. 15–29, 2018. https://doi.org/10.1016/j.dss.2017.11.006.

[58] Wang, K.-J., Makond, B., Chen, K.-H., & Wang, K.-M., “A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients,” *Applied Soft Computing*, vol. 20, pp. 15–24, 2014. https://doi.org/10.1016/j.asoc.2013.09.014.

[59] 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*, vol. 291, pp. 184-203, 2015. https://doi.org/10.1016/j.ins.2014.08.051.

[60] He, H., & Ma, Y., “Assessment Metrics for Imbalanced Learning,” *Imbalanced Learning: Foundations, Algorithms, andApplications*, 2016.