Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 13 (2020) pp. 044-052

KDOS: Kernel Density based Over Sampling: - A Solution to Skewed Class Distribution *

Rekha Gillala[0000-0003-2688-2323] [1], V Krishna Reddy [2] and Amit Kumar Tyagi[0000-0003-2657-8700] [3]

[1]Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation,
Hyderabad, Telangana, India-500075
gillala.rekha@klh.edu.in

[2], Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation,
Vaddeswaram, Guntur, Andhra Pradesh, India – 522502
vkrishnareddy@kluniversity.in

[3] School of Computer Science and Engineering, Vellore Institute of Technology,
Chennai Campus, Chennai, 600127, Tamilnadu, India.
amitkrtyagi025@gmail.com

Abstract: **In data science, the problem of skewed response variable is very common. Many real-world applications such as medical diagnosis, credit card fraud detection, system intrusion detection and many others suffer from abnormal behavior of class distribution. Most common approach to combat skewed distribution is through sampling the class either majority or minority to achieve balanced distribution. In this paper, we investigate the performance of kernel density estimator for oversampling the skewed data distribution. We believe that Kernel Density Estimator offers a more natural way of generating synthetic samples for minority class that is less prone to over fitting than other standard sampling techniques. Experimental results show that the KDOS can outperform other standard sampling techniques on 12 real time data sets using precision, recall, F-measure, ROC (AUC) and accuracy. Furthermore, the proposed method outperforms when applied using ensemble classification algorithms.**

Keywords: Kernel, KDE, Class imbalance, Sampling, Oversampling.

I. Introduction

Many real world applications face the challenge of skewed class distribution. It is widespread in the fields of medical diagnostics, fraud detection, network intrusion detection and many others involving rare events [14] [21]. Skewed class distribution problem is a binary classification problem, where the target (class label) variable consists of two classes where samples of one class vastly outnumbered than other class. The class with more number of samples is called as majority/ negative class and the class with lesser number of samples is called as minority/positive class. In such cases, the traditional machine learning algorithms tend to be biased towards majority class.

To address the skewed class distribution problem, researchers have proposed various strategies [3] [9] [8] that are broadly divided into three categories: sampling, cost-sensitive learning and ensemble approaches. Sampling techniques involves balancing the class distribution by either oversampling the minority class or under-sampling the majority class. This is a very well-known approach applied in various scenarios with better performance. However, it suffers from limitations. Oversampling may lead to over fitting of samples while under-sampling may lead to loss of potentially important information. Whereas, cost sensitive learning works on penalizing the minority class instances each time it is misclassified. The main objective is to minimize the overall cost by putting more emphasis on minority class instances [13]. Ensemble techniques combine two or more techniques and popular for increasing the accuracy by combining several classifiers. It has been successfully applied for skewed data distribution [19] [16] [20].The combination of ensemble learning with sampling techniques for handling class imbalance problem have been proposed in the literature with better results [5] [22].

In this paper, we propose a kernel density based oversampling approach to deal with skewed class distribution. Kernel density estimation is a non-parametric method for estimating the probability density distribution based on the given sample [23] [25]. It estimates the unknown density function by considering set of homogeneous kernel functions centered at each sample point. We can generate new samples based on the density function. The proposed technique offers an effective approach for generating synthetic instances based on non-parametric estimations. Numerical experiments of our methods show better results than existing re-sampling techniques such as Random Over-Sampling (ROS) [4], Synthetic Minority Oversampling TEchnique (SMOTE) [2], ADAptive SYNthetic sampling approach (ADASYN) [10].

The paper is organized as follows. In Section 2, we give an overview of the relevant literature for our study. In Section 3, we describe the methodology used in the study. We present our numerical experiments and results in Section 4, and Section 5 concludes the paper.

II. Literature Survey

The problem of skewed class distribution arises in a number of real-life applications and various solutions to address this problem have been proposed by research community [18]. Krawczyk [14] presents a good survey on this problem with challenges and future directions. One of the common and popular techniques to deal with skewed class distribution is through sampling whereby oversampling the minority class and under-sampling the majority class. In the former, the minority class is repeatedly sampled to generate the synthetic data in size proportion to majority class. Similarly, in the later approach the majority samples are discarded to achieve a balanced class distribution.

One of the most popular oversampling techniques called SMOTE proposed by Chawla et al [2]. In their approach, new synthetic samples are generated by linear interpolation between the existing minority samples. Many variant of SMOTE have been proposed in the literature, Adaptive synthetic sampling approach called ADASYN [10] is one of the popular technique among them. It generates synthetic data in the boundary neighborhood between the minority classes. There also exists nonlinear technique such as KernelSVM, where the author interpolates the points between the feature space.

Among under-sampling techniques, the most popular techniques are Tomeklink [1], Random Under-Sampling (RUS), and NearMiss [17]. In NearMiss approach the minority samples are selected based on the average distance between the negative samples to the k closest samples of the positive class is the smallest. Yen et al [27], proposed cluster based under-sampling technique and presented a good performance results. Author [11] argued that combination of oversampling and under-sampling techniques may also improve the performance of the classifiers. Author [7] proposed cluster-based under-sampling based on farthest neighbors.

As this paper deals with data-level techniques, a brief introduction to various data-level techniques are described as follow. In data-level approach, the sample dataset is modified to balance the class distribution. The foremost aim is to maintain equality in the class distribution for the datasets using sampling methods such as over-sampling, under-sampling and combination of both. The oversampling and under-sampling techniques are the two popular techniques in sampling-based classification to address the imbalance problem. In the oversampling technique, some samples are added to the minority class to make it balanced when very less information is available for minority class samples. In the under-sampling technique, some samples of the majority class are eliminated to make the dataset balanced. Apart from above, the hybrid techniques usually come with a combination of both over and under-sampling methods. Figure 1 presents the different approaches applied at data-level to address the class imbalance problem.

The density distribution based sampling technique proposed in this paper samples the minority class instances based on underlying probability distribution.

In general, probability density estimation techniques can be classified into parametric and nonparametric. In parametric methods a fixed density function is assumed and its parameters are then estimated to obtain current samples based on maximum likelihood method. The main drawback of parametric estimation is venerable to over fitting and may raise bias, whereas, non parametric methods estimate the probability density distribution directly from the given data. Among the non-parametric methods, the most popular approach proposed in the literature is Kernel Density Estimation (KDE). It is well known technique widely used both in machine learning and statistics [25]. KDE provided best results and successfully used in wide range of applications including diagnosis of breast cancer [24], Automated image annotation [26], and Outlier detection in high dimensional data [12]. In [6] the author proposed KDE sampling approach for oversampling the minority samples and trained using radial basis function classifier. Our paper differs from [6] wherein we perform a systematic study of KDE with different classification algorithms (ensemble techniques). We compare the performance of our approach with state-of-the-art methods using large number of data samples.

Hence in this section, we reviewed the article proposed for solving class imbalance problem. Next section presents the kernel density sampling technique.

III. KDE sampling

Kernel Density Estimation (KDE) is a non-parametric technique to estimate probability density function on a finite sample data. It is an important statistical tool for data analysis. The resulting density function can be used to investigate the variable properties of a given sample. Let x_1, x_2, x_3,..., x_n be a independent and identically distributed sample drawn from some distribution with an unknown density *f*. Then the kernel density estimate of f is given by

$$\widehat{f(X)} = \frac{1}{n}\sum_{i=1}^{n} k_h\left(x - x_i\right) \tag{1}$$

Where *K* is the kernel function, A kernel with subscript *h* is called the scaled kernel h is the bandwidth parameter and $K_h(t) = \frac{1}{h}k(\frac{1}{h})$.

The value of f(X) is estimated as the average distance from x to the sample point x_i using kernel function K(t). There are number of kernel functions that can be used such as linear kernel, Polynomial Kernel, Gaussian Kernel, Exponential Kernel, Laplacian Kernel, ANOVA Kernel, Hyperbolic Tangent (Sigmoid) Kernel, Quadratic Kernel. In this work, Gaussian kernel is used as it is the most popular function represented as

$$K\left(x, y\right) = exp\left(\frac{\left\|x - y\right\|^2}{2\sigma^2}\right) \tag{2}$$

The Gaussian kernel computes the similarity between the data samples in higher dimensional space.

The difference between KDOS sampling and other standard sampling methods is illustrated in figure 2. The data points in the figure are uniformly distributed with radius of 2 from the centre totally 100 samples points have been generated. From figure 2, it has been observed that KDOS creates new samples by moving around existing minority sample space. Hence in this section, we discussed about the kernel density

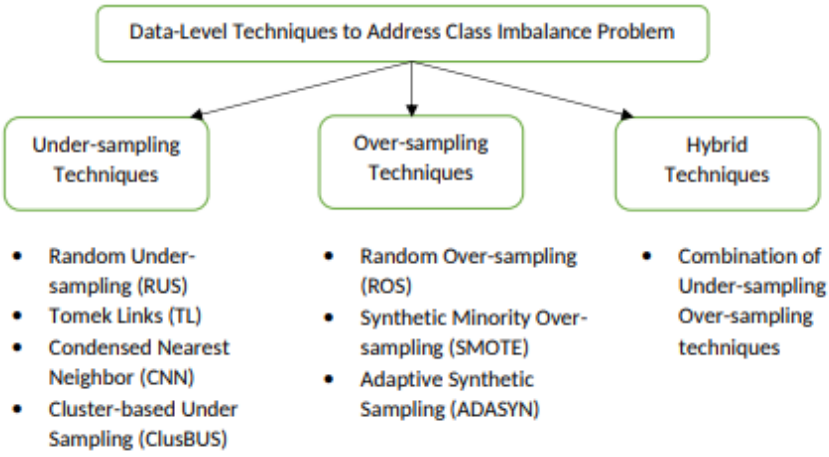

Figure. 1: Different Data-level Techniques Proposed for Handling Class Imbalance Problem [22]

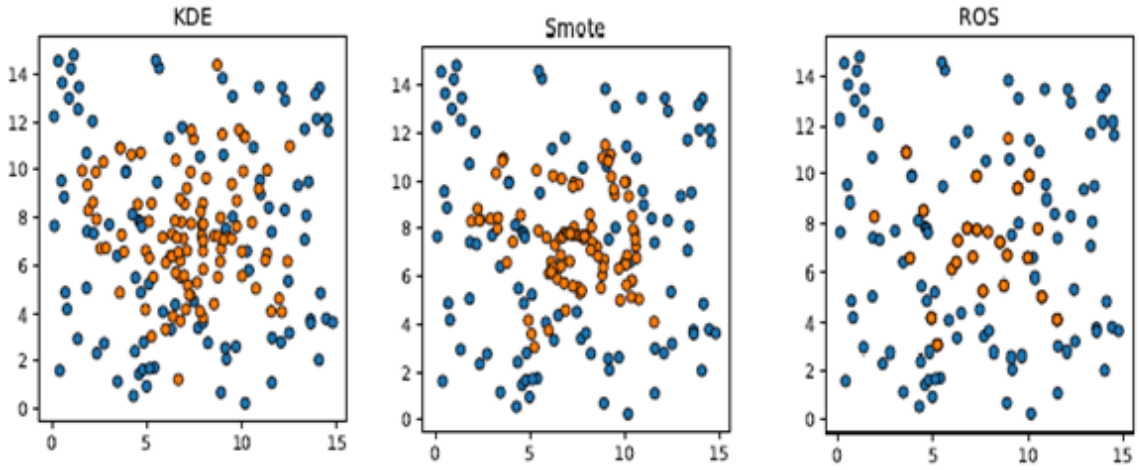

Figure. 2: presents difference between the standard sampling methods and KDOS technique.

oversampling technique and next section presents the numerical experiments and results of the proposed method.

IV. Numerical Experiments and Results

In this section, we carry out a number of experiments by comparing KDOS to three standard sampling techniques very often used in the literature like Random Oversampling, SMOTE and ADASYN. The implementations of the sampling techniques are done in Python using imblearn library[15]. In particular, the default parameters for multivariate Gaussian is determined by cross validation. For each dataset, the following procedure has been performed. The dataset is divided into 10 parts randomly, wherein first one is considered as the test sample and the remaining is taken as training sample. The performance of the classification algorithm is measured using precision, recall, F-Measure, Area Under Curve (AUC) and accuracy. We also considered accuracy but as per the literature, it is not suitable to measure the performance of imbalanced datasets. The formula for calculating F-Measure is shown in equation 3 whereas AUC is computed as shown in equation 4

$$F - Measure = 2 * (Precision * Recall) / (Precision + Recall) \quad (3)$$

$$AUC = \frac{1}{2} \times \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (4)$$

where TP is the True Positive, FN is the False Negative, TN represents the True Negative and FP is the False Positive.

A. Data Sets Used

We used 12 real time data sets from UCI repository with various class imbalance ratios. Table1 shows the list of data sets with its imbalance ratio used in the experiments. Each sampling method is tested on one single classifier (C4.5), and one ensemble classifier (AdaBoost).

During the experiments the data was split into training set and test set. The classifier is trained using training data and results are reported based on test set. During the experiments the data was split into training and testing parts. The results based on the testing part are calculated and reported in the study. Furthermore, each experiment was run twice using different training/testing splits. The results of the experiments on single classifier are presented in table 2 whereas the results pertaining to ensemble classifier are presented in table 3. Figure 3 and 4 shows the Precision, Recall, F-Measure, and ROC results obtained on the imbalanced datasets using C4.5 and ensemble classifier (AdaBoost

Table 1: Datasets used with its IR value

DataSet	Imbalance Ratio (IR)
Bank	7.6:1
Customer Churn	7.5.1
Diabetic	1.86:1
Ecoli	8.6:!
Haberman	2.78:1
Hmeq	3.15:1
Ionosphere	1.87:1
Pima	1.87:1
Satimage	9.3.1
Shuttle	6.02:1
Spambase	3.15:1
Vehicle	3.25:1

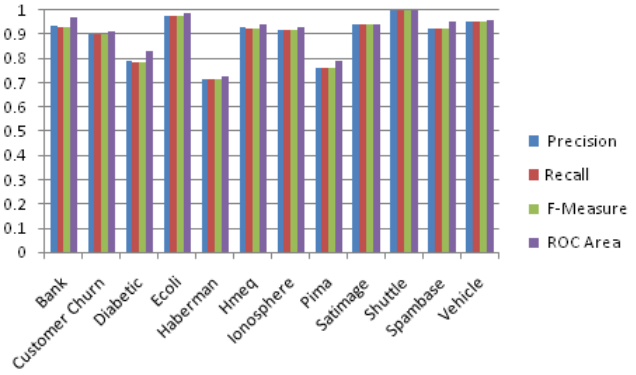

Figure. 3: shows the Precision, Recall, F-Measure, and ROC using C4.5

using C4.5). For the result we conclude that the sampling technique combined with ensemble classification yields better performance than trained on single classifier. We also compared our results with state-of-the-art techniques such as Random Over-Sampling (ROS) [4], Synthetic Minority Oversampling TEchnique (SMOTE) [2], ADAptive SYNthetic sampling approach (ADASYN) [10]. Table 4-7 shows the results and we state that the proposed method (KDOS) performed better than the existing techniques on mostly of the datasets. Graphical presentation of the results are presented in Figure 5-8.

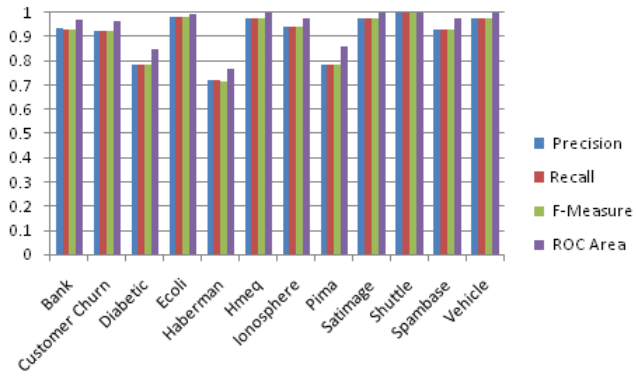

Figure. 4: shows the Precision, Recall, F-Measure, and ROC using AdaBoost

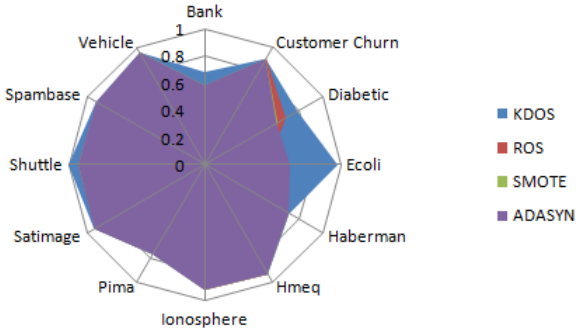

Figure. 5: Comparision of KDOS with state-of-the-art methods using precision based on ensemble classifier (AbaBoost)

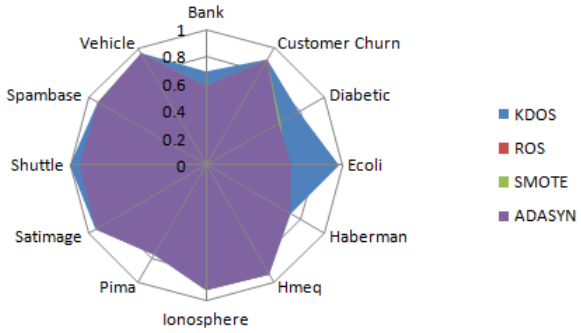

Figure. 6: Comparision of KDOS with state-of-the-art methods using recall based on ensemble classifier (AbaBoost)

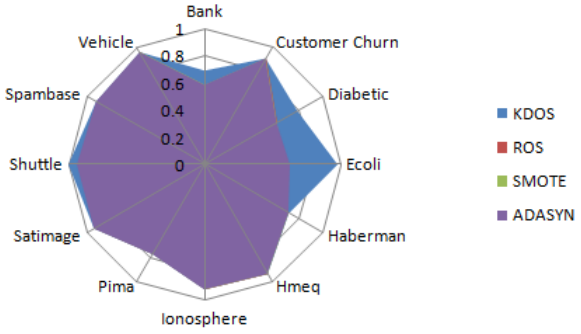

Figure. 7: Comparision of KDOS with state-of-the-art methods using F-measure based on ensemble classifier (AbaBoost)

Table 2: Precision, Recall, F-Measure, and ROC results based on C4.5 classifier

DataSet	Precision	Recall	F-Measure	ROC Area
Bank	0.933	0.932	0.932	0.97
Customer Churn	0.898	0.898	0.898	0.91
Diabetic	0.788	0.787	0.787	0.833
Ecoli	0.978	0.978	0.978	0.988
Haberman	0.717	0.716	0.715	0.726
Hmeq	0.927	0.926	0.926	0.939
Ionosphere	0.917	0.916	0.915	0.93
Pima	0.759	0.759	0.759	0.792
Satimage	0.938	0.938	0.938	0.941
Shuttle	1	1	1	1
Spambase	0.922	0.921	0.921	0.953
Vehicle	0.952	0.951	0.951	0.959

Table 3: Precision, Recall, F-Measure, and ROC results based on ensemble classifier (AbaBoost)

DataSet	Precision	Recall	F-Measure	ROC Area
Bank	0.933	0.932	0.932	0.97
Customer Churn	0.921	0.921	0.921	0.966
Diabetic	0.784	0.783	0.783	0.849
Ecoli	0.982	0.982	0.982	0.995
Haberman	0.719	0.718	0.717	0.767
Hmeq	0.977	0.977	0.977	0.996
Ionosphere	0.943	0.942	0.942	0.978
Pima	0.784	0.784	0.784	0.858
Satimage	0.977	0.977	0.977	0.997
Shuttle	1	1	1	1
Spambase	0.93	0.93	0.929	0.975
Vehicle	0.977	0.977	0.977	0.996

Table 4: Comparision of KDOS with state-of-the-art methods using precision based on ensemble classifier (AbaBoost)

DataSet	KDOS	ROS	SMOTE	ADASYN
Bank	0.681	0.393	0.584	0.584
Customer Churn	0.898	0.898	0.898	0.898
Diabetic	0.788	0.685	0.623	0.613
Ecoli	0.978	0.457	0.519	0.631
Haberman	0.717	0.352	0.717	0.717
Hmeq	0.927	0.927	0.927	0.927
Ionosphere	0.917	0.917	0.917	0.917
Pima	0.759	0.759	0.759	0.759
Satimage	0.938	0.938	0.938	0.938
Shuttle	1	0.936	0.936	0.931
Spambase	0.922	0.922	0.922	0.922
Vehicle	0.952	0.952	0.952	0.952

Table 5: Comparision of KDOS with state-of-the-art methods using recall based on ensemble classifier (AbaBoost)

DataSet	KDOS	ROS	SMOTE	ADASYN
Bank	0.689	0.384	0.584	0.584
Customer Churn	0.898	0.898	0.898	0.898
Diabetic	0.787	0.613	0.634	0.62
Ecoli	0.978	0.471	0.593	0.631
Haberman	0.716	0.354	0.716	0.716
Hmeq	0.926	0.926	0.926	0.926
Ionosphere	0.916	0.916	0.916	0.916
Pima	0.759	0.759	0.759	0.759
Satimage	0.938	0.938	0.938	0.938
Shuttle	1	0.89	0.89	0.931
Spambase	0.921	0.921	0.921	0.921
Vehicle	0.951	0.951	0.951	0.951

V. Conclusion

In this paper, we studied KDOS oversampling technique based on kernel distribution. We consider that KDOS provides a statistically approach to generate synthetic samples in an imbalanced dataset. It creates new instance with no or minimal over fitting. One advantage using KDOS is it can be customize by choosing different kernel functions. Apart, it is well established concept in statistical foundation and has variety of libraries implemented in R, and Python. We carried out a study of KDOS approach on 12 real data sets. The proposed method was compared based on single classifier and ensemble classifier. The results show a better performance

Table 6: Comparision of KDOS with state-of-the-art methods using F-Measure based on ensemble classifier (AbaBoost)

DataSet	KDOS	ROS	SMOTE	ADASYN
Bank	0.689	0.384	0.584	0.584
Customer Churn	0.898	0.898	0.898	0.898
Diabetic	0.787	0.613	0.614	0.614
Ecoli	0.978	0.471	0.593	0.631
Haberman	0.715	0.345	0.715	0.715
Hmeq	0.926	0.926	0.926	0.926
Ionosphere	0.915	0.915	0.915	0.915
Pima	0.759	0.759	0.759	0.759
Satimage	0.938	0.938	0.938	0.938
Shuttle	1	0.915	0.915	0.938
Spambase	0.921	0.921	0.921	0.921
Vehicle	0.951	0.951	0.951	0.951

Table 7: Comparision of KDOS with state-of-the-art methods using ROC based on ensemble classifier (AbaBoost)

DataSet	KDOS	ROS	SMOTE	ADASYN
Bank	0.691	0.375	0.575	0.575
Customer Churn	0.91	0.91	0.91	0.91
Diabetic	0.833	0.633	0.622	0.614
Ecoli	0.988	0.457	0.593	0.611
Haberman	0.726	0.352	0.726	0.726
Hmeq	0.939	0.939	0.939	0.939
Ionosphere	0.93	0.93	0.93	0.93
Pima	0.792	0.792	0.792	0.792
Satimage	0.941	0.941	0.941	0.941
Shuttle	1	0.961	0.961	0.938
Spambase	0.953	0.953	0.953	0.953
Vehicle	0.959	0.959	0.959	0.959

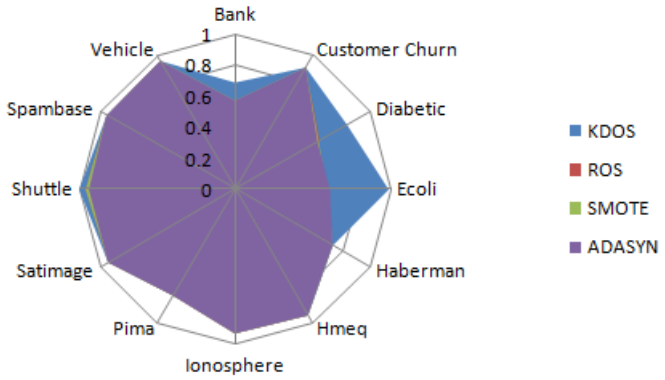

Figure. 8: Comparision of KDOS with state-of-the-art methods using ROC based on ensemble classifier (AbaBoost)

of KDOS when trained using ensemble classification algorithms.

Acknowledgments

This research is funded by the Anumit Academy's Research and Innovation Network (AARIN), India. The authors would like to thank AARIN India, an education foundation body and a research network for supporting the project through its financial assistance.

References

[1] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] Swagatam Das, Shounak Datta, and Bidyut B Chaudhuri. Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81:674–693, 2018.

[4] David J Dittman, Taghi M Khoshgoftaar, Randall Wald, and Amri Napolitano. Comparison of data sampling approaches for imbalanced bioinformatics data. In *The twenty-seventh international FLAIRS conference*, 2014.

[5] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

[6] Ming Gao, Xia Hong, Sheng Chen, Chris J Harris, and Emad Khalaf. Pdfos: Pdf estimation based over-sampling for imbalanced two-class problems. *Neurocomputing*, 138:248–259, 2014.

[7] Amit Kumar Gillala Rekha, Tyagi. Cluster-based under-sampling using farthest neighbour technique for imbalanced datasets. In *10th International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA 2019)*. Springer, 2019.

[8] Qiong Gu, Zhihua Cai, Li Zhu, and Bo Huang. Data mining on imbalanced data sets. In *2008 International Conference on Advanced Computer Theory and Engineering*, pages 1020–1024. IEEE, 2008.

[9] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[10] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. IEEE, 2008.

[11] Chuanxia Jian, Jian Gao, and Yinhui Ao. A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*, 193:115–122, 2016.

[12] Firuz Kamalov and Ho Hon Leung. Outlier detection in high dimensional data. *arXiv preprint arXiv:1909.03681*, 2019.

[13] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

[14] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[15] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

[16] Tian-Yu Liu. Easyensemble and feature selection for imbalance data sets. In *2009 international joint conference on bioinformatics, systems biology and intelligent computing*, pages 517–520. IEEE, 2009.

[17] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, 2003.

[18] Ronaldo C Prati, Gustavo EAPA Batista, and Maria Carolina Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. In *Mexican international conference on artificial intelligence*, pages 312–321. Springer, 2004.

[19] Yun Qian, Yanchun Liang, Mu Li, Guoxiang Feng, and Xiaohu Shi. A resampling ensemble algorithm for classification of imbalance problems. *Neurocomputing*, 143:57–67, 2014.

[20] G Rekha and Amit Kumar Tyagi. Necessary information to know to solve class imbalance problem: From a user's perspective. In *Proceedings of ICRIC 2019*, pages 645–658. Springer, 2020.

[21] G Rekha, Amit Kumar Tyagi, and V Krishna Reddy. Solving class imbalance problem using bagging, boosting techniques, with and without using noise filtering method. *International Journal of Hybrid Intelligent Systems*, (Preprint):1–10, 2019.

[22] G Rekha, Amit Kumar Tyagi, and V Krishna Reddy. A wide scale classification of class imbalance problem and its solutions: A systematic literature review. *Journal of Computer Science*, 15:886–929, 2019.

[23] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[24] Razieh Sheikhpour, Mehdi Agha Sarram, and Robab Sheikhpour. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. *Applied Soft Computing*, 40:113–131, 2016.

[25] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[26] Alexei Yavlinsky, Edward Schofield, and Stefan Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *International Conference on Image and Video Retrieval*, pages 507–517. Springer, 2005.

[27] Show-Jane Yen and Yue-Shi Lee. Cluster-based undersampling approaches for imbalanced data distributions. *Expert Systems with Applications*, 36(3):5718–5727, 2009.