Peertechz


ENGINEERING GROUP

ANNALS OF Robotics and Automation

DOI: https://dx.doi.org/10.17352/ara

Research Article

# Enhancing Imbalanced Dataset by Utilizing (*K-NN* Based SMOTE_3D Algorithm)

**Khaldoon Alshouiliy*, Sujan Ray, Ali AlGhamdi and Dharma P Agrawal**

Center for Distributed and Mobile Computing, EECS, University of Cincinnati Cincinnati, OH, 45221-0030, USA

**Received:** 11 March, 2020
**Accepted:** 24 April, 2020
**Published:** 25 April, 2020

***Corresponding author:** Khaldoon Alshouiliy, Center for Distributed and Mobile Computing, EECS, University of Cincinnati Cincinnati, OH, 45221-0030, USA, E-mail: alshoukr@mail.uc.edu

**Keywords:** Imbalanced learning; Machine learning; Oversampling sample subset optimization; SMOTE; AzureML; Kaggle; Synthetic points

https://www.peertechz.com



## Abstract

Big data is currently a huge industry that has grown significantly every year. Big data is being used by machine learning and deep learning algorithm to study, analyze and parse big data and then drive useful and beneficial results. However, most of the real datasets are collected through different organizations and social media and mainly fall under the category of Big Data applications. One of the biggest and most drawbacks of such datasets is an imbalance representation of samples from different categories. In such case, the classifiers and deep learning techniques are not capable of handling issues like these. A majority of existing works tend to overlook these issues. Typical data balancing methods in the literature resort to data resampling whether it is under sampling a majority class samples or oversampling the minority class of samples. In this work, we focus on the minority sample and ignore the majority ones. Many researchers have done many works as most of the work suffers from over sampling or form the generated noise in the dataset. Additionally, works are either suitable for either big data or small data. Moreover, some other work suffers from a long processing time as complicated algorithms are used with many steps to fix the imbalance problem. Therefore, we introduce a new algorithm that deals with all these issues. We have created a short example to explain briefly how the SMOTE works and why we need to enhance the SMOTE and we have done this by using a very well-known imbalance dataset that we downloaded from the Kaggle website. We collect the results by using Azure machine learning platform. Then, we compare the results to see that the model is functional just good with SMOTE and way better than without it.

## Introduction

In the current era, with a huge growth of cell phones and IoT applications, a huge amount of data is being generated every single day. However, most of the collected datasets suffer from an imbalanced re-presentation of samples in different categories. Even though a number of solutions have been introduced, they do have a shortcoming in some aspects of the other.

Even a transient application is a real example of imbalanced datasets as noticed through websites and platforms like Facebook, Twitter, Instagram, and LinkedIn where many people (accounts) have thousands of friends or followers while some others have just a handful few. The big challenges from such imbalanced data requires identifying some good measures to address preprocessing in providing for instance solution to some special issues that might exist in the training dataset that could restore the balance between the majority and the minority classes prior to the machine learning and deep learning phase [1,2].

One of the main reasons that the classification algorithm doesn't work properly is impaired by the class imbalance. In machine learning algorithms, the aim is to maximize classification accuracy, and measure based on the majority class. It could be a reason that the classifier outcomes show high classification accuracy, even when it does not predict a single minority class instance correctly due to imbalance datasets. Take a note that a trivial classifier outcome has classification accuracy of 99.9% assuming that 0.1% transactions are fraudulent to score all credit card transactions as legit will score. However, in this case, all fraud cases remain undetected. To address this problem of imbalanced datasets, most of the researchers use Synthetic Minority Oversampling Technique (SMOTE) [3,4]. Yet this algorithm suffers from drawbacks, like, while generating synthetic examples, SMOTE does not take into consideration neighboring examples from other classes.



**Citation:** Alshouiliy K, Ray S, AlGhamdi A, Agrawal DP (2020) Enhancing Imbalanced Dataset by Utilizing (*K-NN* Based SMOTE_3D Algorithm). Ann Robot Automation 4(1): 001-006. DOI: https://dx.doi.org/10.17352/ara.000002

This can increase overlapping of classes and can introduce additional noise and it's not efficient for high dimensional datasets. In this paper, we develop and improve the SMOTE work by mixing it with another algorithm known as *K*-Nearest Neighbor (*K*-NN) [5].

## Literature review

Most applications in real world data mining include learning from imbalanced data sets. Typically, learning from data sets containing very few minority (or interesting) class instances produces skewed classifiers with a higher predictive accuracy over the majority class(s), but a lower predictive accuracy over the minority class. SMOTE (Synthetic Minority Over-sampling Technique) is explicitly designed to learn from imbalanced data sets. This work has a novel method based on a variation of the SMOTE algorithm and the boosting technique to learn from imbalanced data sets. Unlike normal boosting where all misclassified examples are given equal weights, SMOTEBoost generates synthetic examples from the uncommon or minority class, thus implicitly altering the updating weights and compensating for distorted distributions. SMOTEBoost applied to many extremely and moderately imbalanced data sets reveals improved performance of predictions for the minority class and improved overall F-values [20].

One of the greatest problems of machine learning is supervised learning of unequaled domains. These tasks differ from normal learning tasks by assuming a biased distribution of target variables, and disproportionate user domain to underrepresented instances. Most work has focused on tasks of imbalanced classification, where a wide range of approaches have been tested. However, no research was performed on tasks related to imbalanced regression. In this [21] authors present an adaptation of the SMOTEBoost approach to the imbalanced regression issue. Originally developed for classification activities, it incorporates the methods of boosting and the SMOTE resampling technique. They have presented four variants of SMOTEBoost and provide an experimental evaluation using 30 datasets with comprehensive results review to test the ability of SMOTEBoost methods to predict extreme target values and their predictive trade-off on methods of boosting baselines. SMOTEBoost is available commercially via a software kit [21].

Douzas, G et al. introduced a simple and efficient method of over-sampling based on clustering of k-means and SMOTE, which prevents noise generation and effectively overcomes imbalances between and within groups. The method achieves these characteristics by clustering data using k-means, enabling data generation to be centered on critical areas of input space. A high ratio of findings from minorities is seen as an indication that a cluster is a protected place. Only secure cluster over-sampling allows for k-means SMOTE to prevent noise generation. In addition, the average distance between minority samples of a cluster is used to locate sparse areas. More synthetic samples are allocated to sparse minority clusters which mitigate imbalances within the class. Eventually, it discourages overfitting by producing truly new findings using SMOTE instead of replicating existing ones.

Empirical findings from comprehensive studies with 90 datasets show that classification outcomes are enhanced by training data oversampled with the proposed process [16].

The researchers implemented two under sampling strategies that use a clustering technique during the preprocessing phase of the data. Specifically, the number of clusters in the majority class is calculated to be equal to that of the minority class data points. The first strategy uses the cluster centers to represent the largest class, while the second strategy uses cluster center nearest neighbors. A further analysis was carried out to analyze the impact of the addition or deletion of 5 to 10 cluster centers in the majority class on results. The experimental results obtained using 44 small-scale and 2 large-scale datasets showed that five state of the art strategies were outperformed by the clustering-based under sampling method with the second strategy. This method coupled with a single multilayer perceptron classifier and C4.5 decision tree classifier ensembles precisely provided optimum performance in both small and large data sets [22].

Researchers have widely adopted SMOTE, perhaps due to its versatility and added value with respect to random oversampling. Numerous modifications and extensions have been proposed to the technique which aim to remove its disadvantages. These amendments usually fix one of the flaws in the initial method [23]. Due to their stated target, they can be divided into algorithms aimed at highlighting some minority class areas, aiming to counter imbalances between classes, or attempting to prevent noise generation.

Borderline-SMOTE1 belongs to the group of methods stressing class regions, concentrating its attention on the boundary of the decision. This is the only algorithm discussed here that does not use a clustering technique, and because of its popularity, is included. The methodology replaces the random collection of findings from SMOTE with a specified set of instances near the class boundary. The mark of the nearest neighbors *k* of a sample is used to determine if it is discarded as noise, selected for its assumed proximity to the class line, or omitted because it is far from the boundary. Borderline-SMOTE2 extends this method to allow interpolation of a minority instance and one of its neighbors of the majority class, setting the interpolation weight to below 0.5 so that the sample produced is closer to the minority sample [24].

Cluster-SMOTE, another approach in the techniques group that emphasizes those class regions, uses *k*-means to cluster the minority class before applying SMOTE within the clusters found. The stated objective of this approach is to improve class regions through the formation of samples within naturally occurring minority class clusters. It is not defined how many instances are generated in each cluster, nor how to determine the optimum number of clusters [25]. Although the approach can mitigate the problem of inequality between groups, removing small disjuncts does not help.

Previously, the Borderline-SMOTE (BSMOTE) is proposed to reduce the number of synthetic instances created by generating these instances based on a borderline between the



**Citation:** Alshouiliy K, Ray S, AlGhamdi A, Agrawal DP (2020) Enhancing Imbalanced Dataset by Utilizing (*K-NN* Based SMOTE_3D Algorithm). Ann Robot Automation 4(1): 001-006. DOI: https://dx.doi.org/10.17352/ara.000002

majority and minority groups. Unfortunately, BSMOTE was unable to provide huge savings in terms of the number of instances produced, trading to the accuracy of classification. In order to improve the accuracy of BSMOTE, H. A. Majzoub and I. Elgedawy introduced an Affinitive Borderline SMOTE (AB-SMOTE) that leverages the BSMOTE and improves the consistency of the synthetic data produced by taking into account the borderline instances affinity. The results of the experiments show that the AB-SOMTE managed to yield the most reliable results in most of the test cases adopted in our sample when compared with BSMOTE [26].

### Synthetic Minority Oversampling Technique (SMOTE)

SMOTE was introduced by Nitesh Chawla et al. in 2002 [6]. Their objective was to resolve an imbalanced dataset in order to obtain trustworthy decisions using machine learning. There had been a simpler way to overcome the minority class dilemma like duplicating samples from the minority class in the training set prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model [7]. SMOTE algorithm starts by first selecting a minority class instance at a random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the *k* nearest neighbors *b* at random and connecting *a* and *b* to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances *a* and *b* according to Haibo He et al.in their 2013 book entitled "Imbalanced Learning: Foundations, Algorithms, and Applications" [8]. This mechanism of creating synthetic creation can be repeated as many times as needed until a targeted balanced percentage is reached. However, some shortcomings of SMOTE are inevitable as it synthetizes new instances without taking into consideration of the majority class which might lead to fuzzy boundaries between the positive and negative classes [9].

### The steps of the algorithm described as [10]:

**Step 1:** Setting the minority class set *A*, for each $x \in A$,, the k-nearest neighbors of *x* are obtained by calculating the Euclidean distance between x and every other sample in set *A*.

**Step 2:** The sampling rate *N* is set according to the imbalanced proportion. For each $x \in A$, , *N* examples (i.e. $x_1$, $x_2$, ...$x_n$) are randomly selected from its *k*-nearest neighbors, and these construct the set $A_1$.

**Step 3:** For each example $x_k \in A_1$ (*k*=1, 2, 3...*N*), the following formula is used to generate a new example: $x' = x + rand(0, 1) * |x - x\_k|$ in which *rand(0, 1)* represents the random number between 0 and 1.

### K-Nearest Neighbor (*K-NN*)

*K*-NN algorithm is a method for classifying objects based on closest training examples in the feature space. *K*-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification [11]. *K*-NN is a technique used in various fields such as pattern recognition, text categorization, moving object recognition, etc. [12]. The *K*-nearest neighbor classifier is a conventional nonparametric classifier that provides good performance for optimal values of *K*. In the *K*-nearest neighbor rule, a test sample is assigned a class most frequently represented among *K* nearest training samples. If two or more such classes exist, then the test sample is assigned the class with minimum average distance to it. It can be shown that the *K*-nearest neighbor rule becomes the Bayes optimal decision rule as *K* goes to infinity [13].

The algorithm can be described as [14]:

1. Load the data,
2. Initialize *K* to your chosen number of neighbors,
3. For each example in the data:
   a. Calculate the distance between the query example and the current example from the data.
   b. Add the distance and the index of the example to an ordered collection.
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances,
5. Pick the first *K* entries from the sorted collection,
6. Get the labels of the selected *K* entries,
7. If regression, return the mean of the *K* labels, and
8. If classification, return the mode of the *K* labels.

## Methodology

In our methodology, we introduce a new idea about how to improve the SMOTE by SMOTE_3D. Firstly, let us understand how the SMOTE works. Figure 1 shows us how SMOTE generates synthetic examples by creating synthetic points around the minor class by selecting the distance between two instances and then generate points between them.

SMOTE is well explained in [1,2]. However, we explain our technique in a new way without compromising fair distribution of the dataset and consequently avoid any noise augmentation or decrease the execution speed, and therefore generate optimal instances. In our proposed method shown in Figure 2, we first ignore the majority group and focus solely on the minority group. In the minority group, we divide the minority instances into multiple subgroups and we calculate the average distance between all the instances of each individual

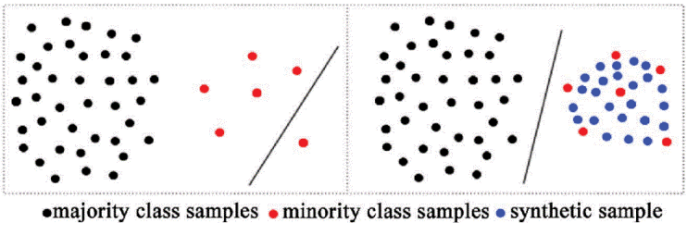


**Figure 1**: SMOTE Generates New Instances [15].



**Citation:** Alshouiliy K, Ray S, AlGhamdi A, Agrawal DP (2020) Enhancing Imbalanced Dataset by Utilizing (*K-NN* Based SMOTE_3D Algorithm). Ann Robot Automation 4(1): 001-006. DOI: https://dx.doi.org/10.17352/ara.000002

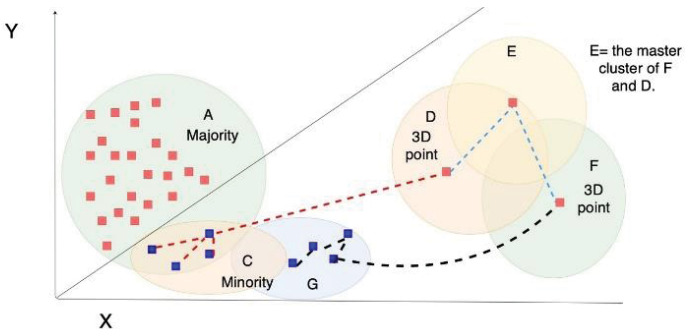


**Figure 2**: Proposed Technique.

group. The average value serves as a new point and works as the *3D* instances by connecting it to all instances in the same group. Then, we create a master cluster of all those *3D* points of every individual group. This solves the problem of the noise. Additionally, when there are many subgroups, it takes time to process and generate new instances. To deal with this issue, our technique works to expedite generation of the new instances by assigning each subgroup to a parallel processor system in order to process all subgroups simultaneously. It is worth mentioning that our method works only with supervised dataset and not unsupervised dataset. Some researchers used the *K*-means algorithm which is suitable for unsupervised dataset as used in [16]. However, some other papers [17] used *K*-NN with supervised datasets.

## SMOTE in a practical way

To understand the SMOTE in a practical way, we used one well-known dataset named "Credit Card Fraud Detection". This dataset is available on Kaggle repository [18]. We first download the dataset file into our local machine, after that we uploaded it to the Azure Machine Learning (AzureML) [19]. Azure is a cloud platform, provided and supported by Microsoft and, it is accessible for students and we used our academic account to run this work. Azure come up with many good features like, no need to be a programmer because it is a drag and drop platform, can deal with big data size that can be saved on Microsoft Azure cloud, and can run the work remotely through the internet which saves us from paying special machine to run the work. Moreover, researchers can pre-process data, analyze and reduce, extract features, train, and test datasets. Azure support most of the dataset types like arff, csv, tsv, and OData as shown in Figure 3.

After we uploaded the dataset (creditecard.csv) into AzureML, we visualized the dataset and we found out that the dataset has 284807 rows and 31 columns. Additionally, the dataset is imbalanced that we need to take care of it as we can see that in Figure 4, when the column named "class" has 0 for who paid and 1 for fraud. So, in this case when we run the machine learning algorithm to predict whether the costumer is going to pay or go fraud, the outcome of the accuracy is almost 100% and that is because of the imbalanced dataset. Then, we apply the SMOTE technique as it is available in this platform and we kept everything as default parameter except the SMOTE percentage which we increase to control the imbalance dataset

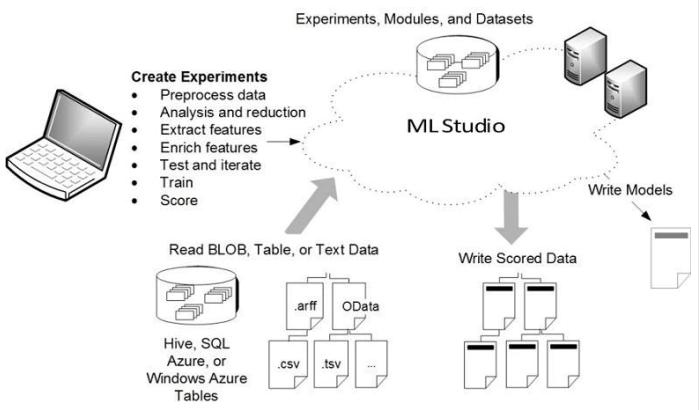


**Figure 3**: Azure Machin Learning Platform [19].

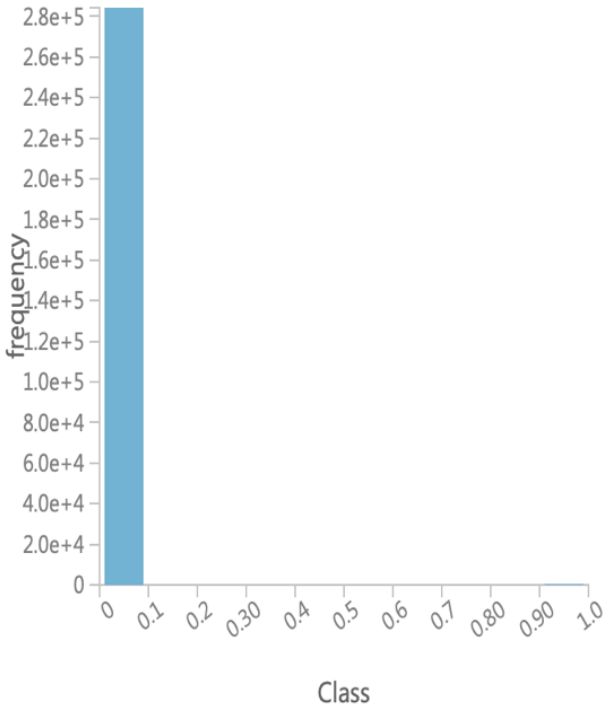


**Figure 4**: Imbalance dataset.

by adding more rows to minority class. After that, we visualize the dataset again and we find out that the number of rows extended to 407807 with 13 columns. We can see the final result in Figure 5.

After we see how the SMOTE works, we tried to implement small example to see what the difference is when we use the SMOTE in the model that has imbalanced dataset and without it. As shown in Figure 6, we used two different model on the same dataset and same algorithm which is "Two-class logistic regression". The results show that the accuracy comes out with 99.9% when we did not use the SMOTE and was above 95% with SMOTE, which is a good result since the dataset became big compared to the original one but still has high accuracy and that is because we did not remove and clean the dataset. We implement it as it is Figure 7, shows the Area Under the Curve (AUC) result. Even SMOTE has improved the work and



**Citation:** Alshouiliy K, Ray S, AlGhamdi A, Agrawal DP (2020) Enhancing Imbalanced Dataset by Utilizing (*K-NN* Based SMOTE_3D Algorithm). Ann Robot Automation 4(1): 001-006. DOI: https://dx.doi.org/10.17352/ara.000002

it makes more sense, but it was still suffering from the noise and overlapping problems and that is also one of the reasons that our results was above the 95%. Because of that we will implement our idea in the future to cover these issues.

## Conclusion

The world's big companies and organizations rely on big data analysis to help them in their work by either classifying or predicting algorithms to yield a better machine learning process with acceptable accuracies. This, of course, cannot be achieved without robust and reliable error-free datasets. One common problem among almost all the datasets is either missing values or imbalanced sampling as many of the datasets are collected through social media such as Twitter, Facebook, or LinkedIn.

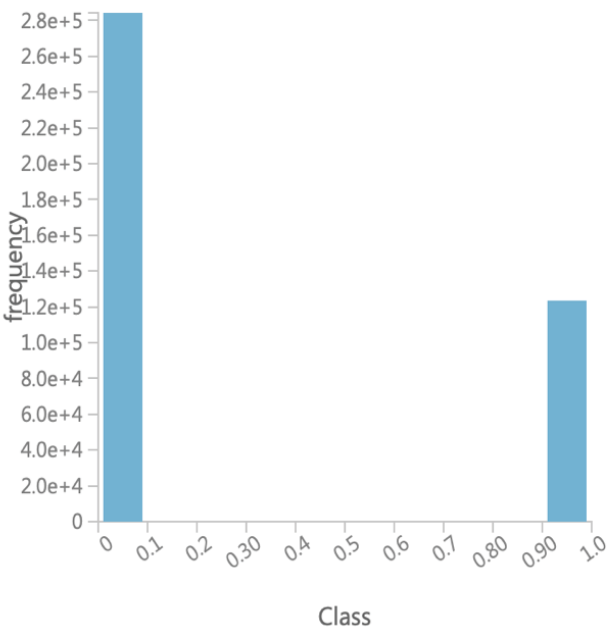


**Figure 5**: Balance Dataset.

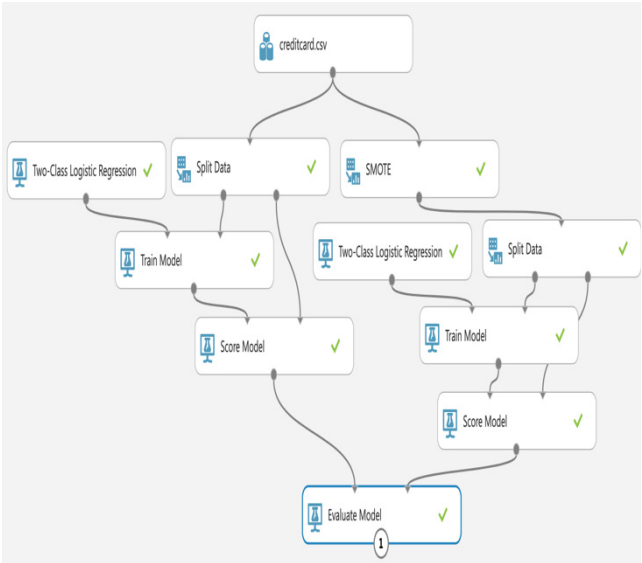


**Figure 6**: SMOTE and Non-SMOTE Models.

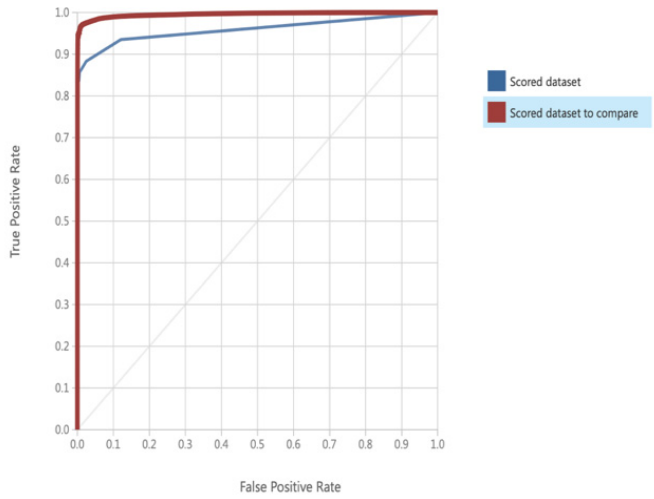


**Figure 7**: AUC Results.

In this case study, we focus on the imbalance dataset and how to improve the procedure of the existing algorithm. There are many algorithms to deal with an imbalanced dataset. A famous one is SMOTE but with the disadvantage of being suitable for big data. In this case, we introduce an idea of SMOTE_3D to fix most of the issues. We expect our algorithm to work properly to handle imbalanced dataset from many aspects. Through the model we created on the AzureML, we find out how SMOTE works and why it is important in each imbalance dataset to use such algorithm that can control the imbalance dataset. We briefly explain the results and then when we need to develop the works of the SMOTE by fixing the issues that the algorithm suffers from it and, that gives our idea SMOTE_3D, a good and strong and promising that can do better than the original SMOTE.

## References

1. Haibo He, Yunqian Ma (2013) Imbalanced Learning: Foundations,Algorithms, and Applications. John Wiley and Sons 216. **Link:** https://bit.ly/2yF46J3

2. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic Minority Over-sampling Technique. JAIR 16. **Link:** https://bit.ly/3au15Za

3. Kotsiantis S, Pintelas P, Anyfantis D, Karagiannopoulos M (2007) Robustness of learning techniques in handling class noise in imbalanced datasets. 247: 21-28. **Link:** https://bit.ly/3bLEBnK

4. Provost F (2000) Machine learning from imbalanced data sets 101. In Proceedings of the AAAI'2000 workshop on imbalanced data sets. AAAI Press 68. **Link:** https://bit.ly/3cKMt9q

5. Cunningham P, Delany SJ (2007) k-Nearest neighbour classifiers. Multiple Classifier Systems 34: 1-17.

6. Chawla NV, Lazarevic A, Hall LO, Bowyer KW (2003) SMOTE Boost: Improving Prediction of the Minority Class in Boosting. Knowledge Discovery in Databases: PKDD 2003 Lecture Notes in Computer Science 107-119. **Link:** https://bit.ly/2KvBzZ0

7. Moniz N, Ribeiro R, Cerqueira V, Chawla N (2018) SMOTEBoost for Regression: Improving the Prediction of Extreme Values. 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). **Link:** https://bit.ly/3cEsmJK



**Citation:** Alshouiliy K, Ray S, AlGhamdi A, Agrawal DP (2020) Enhancing Imbalanced Dataset by Utilizing (*K-NN* Based SMOTE_3D Algorithm). Ann Robot Automation 4(1): 001-006. DOI: https://dx.doi.org/10.17352/ara.000002

8. Douzas G, Bacao F, Last F (2018) Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences 465: 1-20. Link: https://bit.ly/2x4vfVj

9. Lin WC, Tsai CF, Hu YH Jhang JS (2017) Clustering-based under sampling in class-imbalanced data. Information Sciences 409-410: 17-26. Link: https://bit.ly/3cH0qVM

10. Vanhoeyveld J, Martens D (2017) Imbalanced classification in sparse and large behaviour datasets. Data Mining and Knowledge Discovery 32: 25-82. Link: https://bit.ly/3ay1n1k

11. Han H, Wang WY, Mao BH (2005) Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Lecture Notes in Computer Science Advances in Intelligent Computing 878-887. Link: https://bit.ly/3aCgab4

12. Cieslak D, Chawla N, Striegel A (2006) Combating imbalance in network intrusion datasets. 2006 IEEE International Conference on Granular Computing. Link: https://bit.ly/2xY6af9

13. Majzoub HA, Elgedawy I (2020) AB-SMOTE: An Affinitive Borderline SMOTE Approach for Imbalanced Data Binary Classification. International Journal of Machine Learning and Computing 10: 31-37. Link: https://bit.ly/2xY21Yq

14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) "SMOTE: synthetic minority over-sampling technique." J Artif Intell Res 16: 321-357. Link: https://bit.ly/2zjnj35

15. "SMOTE Oversampling for Imbalanced Classification with Python". Link: https://bit.ly/2S1RKS7

16. Weiss GM (2013) Foundations of Imbalanced Learning. Imbalanced Learning. Link: https://bit.ly/351Wmg9

17. Ma L, Fan S (2017) CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinformatics 18: 169. Link: https://bit.ly/2Y5BeUP

18. Handling Imbalanced Data with SMOTE. Link: https://bit.ly/2S5FZdf

19. Imandoust SB, Bolandraftar M (2013) Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. International Journal of Engineering Research and Applications 3: 605-610. Link: https://bit.ly/2Y11GPy

20. Kalaivani P, Shunmuganathan KL (2014) "An improved K-nearest-neighbor algorithm using genetic algorithm for sentiment classification. International Conference on Circuits, Power and Computing Technologies. Link: https://bit.ly/3cKdcmt

21. Kataria A, Singh M (2013) A Review of Data Classification Using K-Nearest Neighbour Algorithm. Link: https://bit.ly/2VzRUC4

22. Machine Learning Basics with the K-Nearest Neighbors Algorithm. Link: https://bit.ly/3aDru6J

23. Dang XT, Hirose O, Saethang T, Tran VA, Nguyen LA, et al. (2013) A novel over-sampling method and its application to miRNA prediction. Journal of Biomedical Science and Engineering 6: 236-248. Link: https://bit.ly/3bB5hrk

24. Beckmann M, Ebecken NF, de Lima BSP (2015) A KNN undersampling approach for data balancing. Journal of Intelligent Learning Systems and Applications 7: 104. Link: https://bit.ly/3aCfhiK

25. Credit Card Fraud Detection. Link: https://bit.ly/3eK2xdb

26. Azure Machine Learning Studio.








**Citation:** Alshouiliy K, Ray S, AlGhamdi A, Agrawal DP (2020) Enhancing Imbalanced Dataset by Utilizing (*K-NN* Based SMOTE_3D Algorithm). Ann Robot Automation 4(1): 001-006. DOI: https://dx.doi.org/10.17352/ara.000002