MDPI

Article

Class Imbalance in Crop/Weed Classification

Brian E McGinnis [1,†,‡]

[1] Affiliation 1; evanmc@arizona.edu
* Correspondence: evanmc@arizona.edu
† Current address: University of Arizona, Tucson, AZ 85721.
‡ These authors contributed equally to this work.

Abstract: This paper presents an overview of solutions to an all too common problem in the classification of vegetation in agricultural images: there are many more crop plants than there are weeds. There are two solutions to this problem examined in this paper: over-sampling the minority class and a combination of over-sampling the minority and under-sampling the majority. Oversampling the minority class involves the generation of synthetic data (SMOTE, Borderline SMOTE, ADASYN, KMeans, and SVM), while combined approaches include SMOTE+Tomek and SMOTE+ENN. Both approaches are evaluated in the context of the classification of crops and weeds using several approaches: Decision Tree, KNN, Logistic Regression, Multi-layer Perceptron, Random Forest, and SVM. The analysis shows that several under-sampling techniques have the greatest positive improvement on Decision Tree classification, but had an almost entirely negative impact on Random Forest classification. Decision Tree was also the greatest beneficiary of the combined approach, but the greatest negative impact was seen in both SMOTE+ENN correction for Logistic Regression and SMOTE+Tomek correction for Random Forest.

Keywords: imbalance, SMOTE, SVM, KNN, LDA, ENN, Tomek Links, over-sampling, under-sampling

1. Introduction

Both precision agriculture and non-invasive assessments of a crop depend on accurate identification of undesired vegetation, and Agricultural images of crops taken from an overhead perspective typically contain only two things: crop and weeds. Fortunately for the grower, but unfortunately for training classification models, these images often contain a much higher portion of crop than they do weeds, particularly if a pre-emergent herbicide has been applied or other active weed control measures are taken in advance of an image collection used for training a machine-learning model. Having 10 weeds out of 100 plants in an image set might not be much of a problem, but having only three might be, especially taking into account a train/test split. The result may be that a weed is represented by only one or two samples. Collecting a sufficient number of images such that the is a sufficient number of weeds present in the data set may be seen as the ideal solution, but is not always a realistic option in all circumstances. Imbalanced correction techniques are crucial for improving machine learning model performance when dealing with imbalanced datasets. This paper focuses two general approaches to address this, generate new data for the minority class (over-sampling), and combine generation of new data and considering a subset of the majority class (under-sampling) This is where the term *over-sampling* may lead readers to the wrong conclusion. The minority class is not over-sampled without modification simply by repeating the same data, but synthetic data is generated such that

Received:
Revised:
Accepted:
Published:

Citation: McGinnis, B. Class Imbalance in Crop/Weed Classification. *AgriEngineering* **2025**, *1*, 0. https://doi.org/

the ratio between the two classes does not exhibit an imbalance. A combined approach of selecting a subset of the majority class (crop) and over-sampling the minority may also improve classification performance.

2. Methodology

Images were taken at the University of Arizona's Maricopa Agricultural Center (MAC) located in Maricopa, AZ of a cantaloupe planting in May 2024 (N 33.061813 W 111.965606). Images were obtained using an Apple iPhone 14, using ambient lighting. These images were color corrected using values obtained from images taken under identical lighting conditions of a Datacolor Spyder correction chart using Adobe Lightroom. All processing software was written in Python 3 and data analysis written in R. The processing software depended heavily on three libraries: OpenCV (image processing), imbalanced-learn (imbalance correction), and scikit-learn (machine learning). The image sets examined were segmented using the Combined Indices 2 approach described by Guerrero et al. and manually assigned classification using software written by the author [1]. Present in the image sets are two types of vegetation: crop and weeds. While the ratio of crop to weeds will most assuredly deviate from the 1:1 ideal, it is not particularly important what the specific ratio is, as the data will be trimmed to reflect the various ratios. That is, while a dataset may have a relatively low imbalance of 10:8, samples of the minority class are randomly discarded to achieve a much higher imbalance, say 50:1. Not covered in this breakdown, but quite important, is the topics of when to use imbalance correction techniques and when to perform parameter selection. This analysis will use a fairly commonly used train/test split of the data to form, and then test models. It is on this train set that the imbalance is corrected – the objective is not to classify samples that are the result of correction, but to build a model that will produce results that exceed doing no correction. Parameter selection has similar concerns, but has some complications that merit mention here. Each plant in the image has color, texture, and shape attributes that are extracted and then the subset identified by PCA is used.[1] The details of the extracted parameters is beyond the scope of this document, but what is not is when the parameter analysis takes place. Parameter analysis takes place prior to any imbalance correction. As section 4 details, new, synthetic entries can be constructed as part of imbalance correction, but the corrections are not considered in parameter selection, just the original data. That is, the parameters are selected from the training set prior to the generation of synthetic data for the minority class or removal of samples from the majority.

3. Class Imbalance

Plantings often exhibit a somewhat inconvenient feature: weeds and crop do not appear in the same proportion. While having a low weed count is probably desirable for crop production, it is not for classification when considering those images as a training set. While this can present itself as a relatively mild imbalance of nine weed plants for every 10 crop plants (certainly a bad situation for the grower) or a more extreme ratio of 1000 crop plants for every weed. In these extreme cases of imbalance, only a few samples of the minority class are used in classification if the overall size of the population is not particularly large – and even more unfortunate case is that some weed species may not be present at all in the set used for training, as they are put into a testing class. Fernández et al., in a book discussing imbalanced datasets, detail several over-sampling correction algorithms, among them SMOTE, ADASYN, Borderline SMOTE, Kmeans, and SVM, as well as combined approaches SMOTE+Tomek and SMOTE+ENN [2]. These algorithms seek to address the imbalance by creating synthetic data from the minority (under-sampling) or combining this with selecting a subset of the majority class, both in an attempt to correct

the class ratio to 1:1. In most cases, of course, the solution is to simply collect sufficient data such that a severe imbalance does not exist, but this may not be a practical solution. This leaves two approaches: under-sample the majority or over-sample the minority. In cases where weeds (the minority) are relatively few, providing synthetic data to restore data to a 1:1 class ratio may be effective. There are various approaches that all have the same basic approach: generate data from the minority class that is similar – but not identical – to the data already present. That is, the minority class is effectively over-sampled, as elements of the minority class are used as the basis for new data. Alternatively, the majority class can be under-sampled to achieve a 1:1 ratio, typically in two ways: randomly, and mathematical techniques that are more sensitive to the data relationships in the majority class.

4. Over-Sampling

Over-sampling is a bit more complex than the term implies – minority samples are not simply sampled until the numbers are equal to a majority class. Just duplicating entries can correct the imbalance ratio back to the more desired 1:1, but does not add new data to the models being built than can be used to classify novel instances. Rather, new samples of the minority class are created, but the key question is *which* of those minority instances form the basis of those synthetic samples? Are they chosen at random? Are they selected somehow? In general, over-sampling tends to be preferred over under-sampling, as the latter may remove interesting data points from consideration. This question of which sample points in the minority set are used will be a topic revisited often as various techniques are described, as what points are used. Before details of various approaches are discussed, perhaps it is worthwhile to discuss why highly accurate predictions on highly imbalanced data are quite useless. Achieving a high accuracy in the case where the numbers of crop greatly outnumber the weeds is quite simple: predict everything is a crop. Consider a simplistic case of a dataset containing 100 crop plants and a single weed. Predicting each plant is a crop yields a model that is 99% accurate, a satisfying, but useless, number.

4.1. SMOTE

The *Synthetic Minority Oversampling TEchnique* is the basic technique used for re-balancing the data set – other techniques in this section are extensions to that approach [3]. Synthetic data is generated by selecting examples that are close in terms of the k nearest neighbors, and selecting new points along the line connected those peers. This technique is not without drawbacks worthy of consideration: it generates data in the same direction, a complication for classification in that the decision surface is distorted, and tends to generate noisy data. The synthetic data points do not exhibit the same variation of the underlying real data points, leading to an introduction of over-fitting.

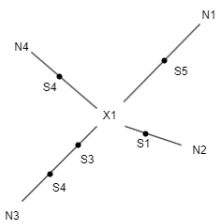

Figure 1. Generating synthetic data points using the SMOTE algorithm considers the k nearest neighbors ($N_{1..4}$) to a point under consideration (X_1). The SMOTE algorithm identifies points along the lines connecting a point to its neighbors ($S_{1..5}$). These points are the synthetic samples used to correct the imbalance.

4.2. ADASYN

The Adaptive Synthetic Sampling approach considers the distribution of the minority points, giving more emphasis to points that a harder to learn. [4] Points that are harder to learn are those that are close to a class border, and in that sense, this approach is close to the borderline proposal. Contrast this with SMOTE. While sharing the same basic approach (considering the k nearest neighbors), SMOTE samples the points uniformly, leading to an oversampling of dense areas, whereas ADASYN has no fixed ratio, but is based on learning difficulty. ADASYN will generate more points for these samples with high learning when processing the same dataset. The term *harder to learn* is a bit imprecise, and warrants some further discussion. ADASYN creates as difficulty ratio for each point, representing the imbalance level in the local neighborhood of the point. This is done by comparing the number of majority class instances (non-minority) to the number of minority class instances within a specified radius around a point. Minority points with a large set of majority class neighbors are said to be harder to learn than those. A disadvantage of using ADASYN is that outlier points tend to have greater representation in the resulting dataset.

4.3. Borderline

In the standard SMOTE approach all members of the minority class are considered in synthetic data generation. In this variant first proposed by Han et al., only those points far from the class border are considered. The rationale behind this approach is that points close to the border contribute little to distinguishing one class from the other, and should not form the basis for new data [5]. In this scheme, points are considered noise if all of their neighbors are of the majority class. To be eligible for resampling, however, a point must have both majority and minority class neighbors. The rationale here is that samples close to a class border tend to be misclassified, and that by limiting the resampling to those points with the smallest risk of misclassification, the overall correct classification rate would be improved.

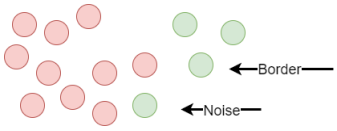

Figure 2. In the borderline variant of SMOTE, points close to the border having only majority class neighbors are considered noise, and are not considered as candidates for selection.

4.4. KMeans

This variant, like many of the others described here, addresses a weakness in basic SMOTE: points are selected randomly for oversampling consideration. A further downside of the base algorithm is the noise it introduces. This is not so much a SMOTE variant as it is a SMOTE supplement. As Last et al. states in a document introducing the algorithm:

> *Another major concern is that SMOTE may further amplify noise present in the data. This is likely to happen when linearly interpolating a noisy minority sample, which is located among majority class instances, and its nearest minority neighbor. [6]*

This approach differs from algorithms such as *borderline* in that this approach views the class to be adjusted in terms of cluster membership. In this approach, the clusters are formed with the *kmeans* approach and SMOTE is applied to those clusters with a high portion of the minority class.

4.5. Support Vector Machine

SVM SMOTE increases the points for the minority class along the decision boundary by generating new points along the lines connecting the support vectors and nearest

neighbors [7]. In contrast with the KMeans approach, but in the same spirit as the borderline approach, this approach considers those points more important for estimating the best decision boundary, and therefor the best candidates for synthetic data generation. This approach first uses SMOTE to create new minority class samples and then uses the new minority class to train SVM. The samples that are identified as the most difficult to learn are then candidates for oversampling.

5. Combined Under-sampling and Over-sampling

Correcting the imbalance by under-sampling the majority class can have an unfortunate side-effect: information loss. Under-sampling, in its most basic form, involves discarding data to achieve the desired balance. Before the details of under-sampling are covered, perhaps this is a topic best addressed by a simple example. Consider the case where the majority class has 1000 samples, but only a few of those samples are representative without regard to the relationship between them (random discard), the discard may eliminate enough of those samples to have an appreciable effect when presented with a novel dataset. In weed/crop classification, however, members of the majority class (crop) are not likely to have small sets representative of a subset, especially if the sample set is taken from a single crop. First proposed in 1976, Tomek described a mechanism to choose discard samples based on their relationship to members of the minority class [8]. Unlike the random selection of candidates, samples must have these characteristics to be considered a *Tomek link*:

1. Sample a's nearest neighbor is b.
2. Sample b's nearest neighbor is a.
3. Sample a and b belong to a different class.

Links with the lowest euclidean distance to minority class are eliminated.

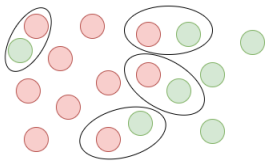

Figure 3. Tomek links are identified by the proximity of minority and majority class samples. Items with the lowest euclidean distance are eliminated.

A second common processing technique is using *edited nearest neighbors* (ENN). The fundamental principle underlying ENN is to remove instances that differ from their nearest neighbors, with the assumption that such instances are likely to be mislabeled (differ from their neighbors) or noisy. The subsequent clarification of the decision boundary may contribute to better outcomes when the dataset is used for training.

1. Compute the k nearest neighbors for each observation in the dataset.
2. Compare the class label (crop or weed) of each instance with the class label of its k nearest neighbors.
3. Remove instances from the majority class that differ from their neighbors

Typically, both under-sampling and oversampling techniques are used to address the imbalance, as Batista et al. describes [9].

6. Results

To assess the efficacy of over-sampling the minority and a combined over- and under-sampling approach, random instances in a dataset were first dropped to achieve five imbalance ratios and then restored to a 1:1 ratio between the two classes using the over-

sampling techniques discussed. Models using the imbalanced data and the artificially balanced data were then compared in terms of the improvement (or degradation) of the Area Under the Curve (AUC) using nine different techniques: Random Forest, Extra Trees, Gradient Boosting, KNN, Linear Discriminant Analysis, Logistic Regression, Multi-Layer Perceptron, Random Forest, and Support Vector Machine. The Receiver Operating Characteristic (ROC) curve represents the probability that the model, if given a randomly chosen positive and negative example, will rank the positive higher than the negative. The AUC is the area underneath this curve – the closer this area is to 1.0, the better a model is said to be.

As Figure 4 shows, substituting synthetic data improves the AUC of various classification techniques in most cases, but some of the impact is quite trivial and – in cases – detrimental (note the case of using a *Random Forest* approach to classification was almost always made worse by these approaches).

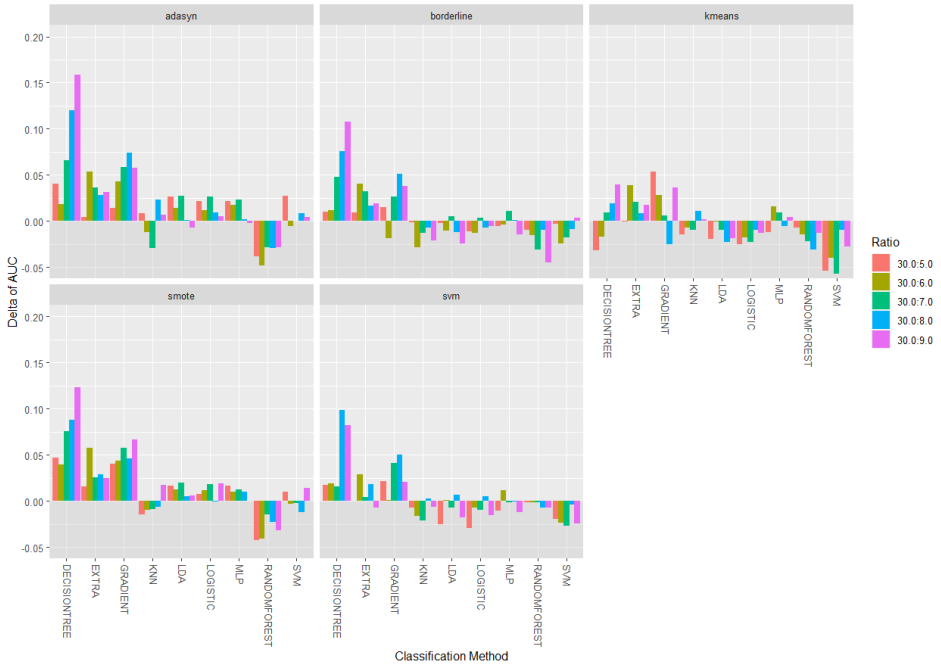

Figure 4. The impact on AUC of class over-sampling imbalance correction techniques for various ratios and classification strategies can be seen in this visualization. Each facet of this display shows the effect of a different over-sampling technique. While many classification techniques are improved, the effect on Decision Tree classification is most pronounced. This dataset was first manipulated to achieve the desired ratio before the algorithms were applied. That is, to achieve a 30:1 imbalance, rows in the minority class were randomly dropped. The change in the AUC scores is shown in this plot for various ML techniques and for each of the correction approaches discussed.

While the difference in the AUC achieved with classification using MLP and SVM (some clarification may be in order: SVM refers to a classification, and SVM-SMOTE refers to the correction technique.) stands out, almost all classification techniques were improved by the correction with a few exceptions. Correcting low-imbalance sets (30:5) yielded worse results in several instances. The postive impacts, however, can be characterized as minimal, however. Consider the ROC curves shown in Figure 5, showing the ROC curve achieved before and after correction using ADASYN.

Combining over- and under-sampling techniques yielded worse results in many instances, but produced better results in most. As Figure 6 shows, the impact tended to be more positive than negative.

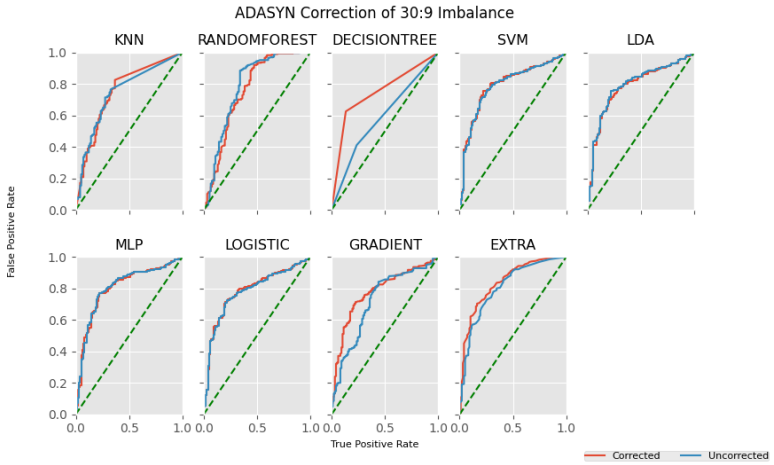

Figure 5. Before and after correction by oversampling the minority class with the ADASYN approach. Note that in many instances the effect of the correction on the ROC curve is quite modest. The dashed green line appearing in each of the graphs is representative of a random classifier, and appears only as a reference. A positive impact on classification using a gradient boosted technique is present, and that technique can certainly benefit from correction.

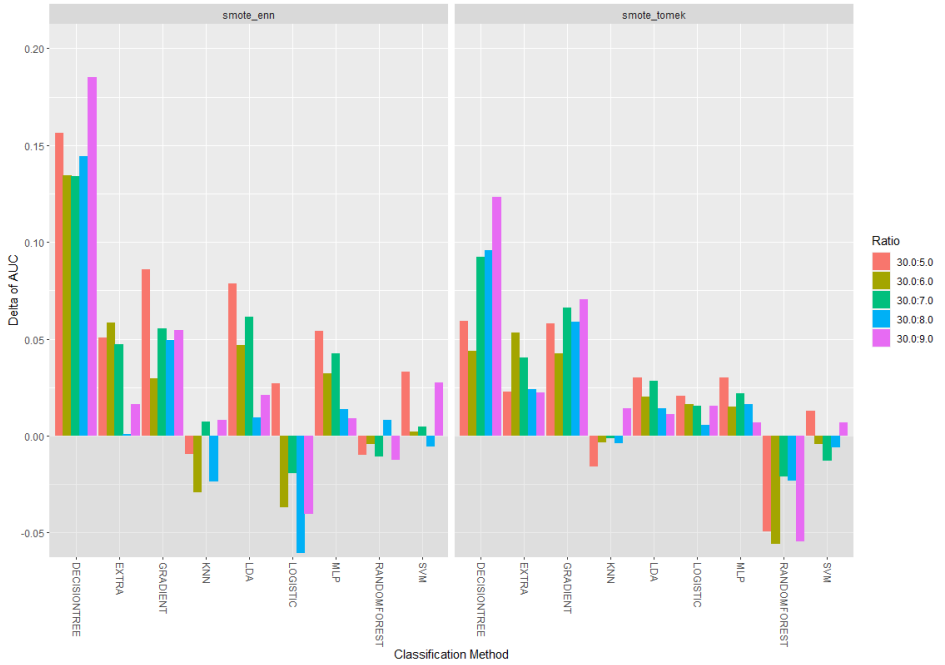

Figure 6. The impact on AUC of using a combined over- and under-sampling approach to class imbalance correction techniques for various ratios and classification strategies can be seen in this visualization. The strategy of using SMOTE+ENN has a particularly negative effect for many of the classification techniques, but the improvement in Decision Tree classification is a clear standout for both techniques.

As with over-sampling the minority class, the effect was still modest, however. As seen in Figure 7, the ROC curves of uncorrected and corrected are, in general, quite close. A notable exception was the difference in the curves shown by *Gradient Boosting*. That classification technique visibly benefited from the combined approach of over- and under-sampling.

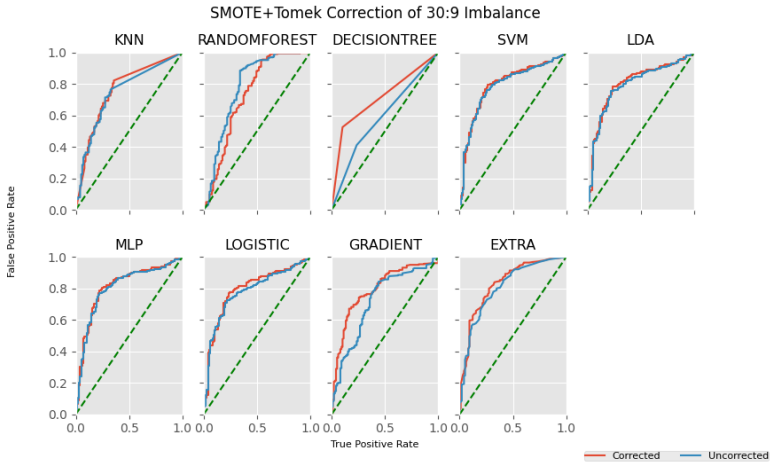

Figure 7. Before and after correction by a combined approach of over-sampling the minority class and under-sampling the majority class using Tomek links. Note that in many instances the effect of the correction is quite modest. A positive impact on classification using a gradient boosted technique is present, and that technique can certainly benefit from correction.

7. Discussion

Imbalance correction is not without cost, both computationally, and accuracy. Some classification techniques (Decision Tree, and SVM) are clear beneficiaries of correction, as both Figures 4 & 6 show. Class imbalance correction is to overcome the bias seen in attempting to learn from situations where one class is significantly larger than another. Introducing a technique that does not improve this bias is not helpful; indeed its outcomes are worse in some situations. While the results reported here involve a single location (MAC) and on dates that were close together (late April and early May 2024), the findings warrant further investigation. Additionally, the correction implementation algorithms are not the product of the author [2]. While it is possible that errors in the implementation used affected the outcomes seen, the imbalance correction techniques examined in this research may not yield results that are acceptable for all situations. Indeed, for many of the techniques, the impact was quite modest overall, and the increase in accuracy may not be justified by the computational cost.

Funding: This research received no external funding.

Data Availability Statement: All code used in the preparation of this paper is available from GitHub. Raw data is available from Figshare.

Acknowledgments: Thanks to Drs. Diaa Elshika and Said Atallah of the University of Arizona for access to the fields where images were acquired.

Conflicts of Interest: None.

Notes

1 Details of the extracted features are beyond the scope of this document, but three features were identified as significant for each plant: GLCM dissimilarity average (shape), saturation (color), and solidity (shape).

2 All processing was performed using the Python imbalanced-learn library that can be obtained from this website: https://imbalanced-learn.org/stable/#

References

1. Guerrero, J.M.; Pajares, G.; Montalvo, M.; Romeo, J.; Guijarro, M. Support Vector Machines for crop/weeds identification in maize fields. *Expert Syst. Appl.* **2012**, *39*, 11149–11155. https://doi.org/10.1016/j.eswa.2012.03.040.

2. Fernández, A.; García, S.; Galar, M.; Prati, R.C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-98074-4.
3. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.
4. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, 2008, pp. 1322–1328. https://doi.org/10.1109/ijcnn.2008.4633969.
5. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Lecture Notes in Computer Science*; Lecture notes in computer science, Springer Berlin Heidelberg: Berlin, Heidelberg, 2005; pp. 878–887. https://doi.org/10.1007/11538059_91.
6. Last, F.; Douzas, G.; Bacao, F. Oversampling for imbalanced learning based on K-means and SMOTE. *arXiv [cs.LG]* **2017**, [arXiv:cs.LG/1711.00837]. https://doi.org/10.1016/j.ins.2018.06.056.
7. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms* **2011**. https://doi.org/10.1504/IJKESDP.2011.039875.
8. Tomek, I. Two modifications of CNN. *IEEE Trans. Syst. Man Cybern.* **1976**, *SMC-6*, 769–772. https://doi.org/10.1109/tsmc.1976.4309452.
9. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor.* **2004**, *6*, 20–29. https://doi.org/10.1145/1007730.1007735.