11/27/2020

K-median clustering in cluster based SMOTE technique (CBSO)

Final project for COSC 6432, Fall 2020

Garima Singh (1793399) and Ruchi Shah (1800950)
UNIVERSITY OF HOUSTON

Table of Contents

Introduction

Cardiovascular disease is the leading cause of death in the United States. Every 37 seconds a person dies from the deadly disease, that is about 674,000 Americans in one year, and 17 million people globally every year. Heart failure occurs when the heart cannot pump sufficient blood for the survival of the body. In order to detect the symptoms and risk factors, doctors have electronic medical records of patients with symptoms, body features, and results of clinical laboratory tests. This data can be used to derive feature correlations and develop predictive models for detection of a potential cardiovascular disease using machine learning tools. In this project, the dataset contains 299 patients with 13 attributes. The attributes are related to underlying health conditions such as anemia, blood report results, health habits all the way to the death event (target feature).
Correct prediction is important for three reasons:

- Patients can participate in shared decision making,
- Overestimating risks leads to unneeded treatment, testing and increase costs.
- Underestimating risks increases chances of death and loss of early recover.

We use the results of (Chicco and Jurman, 2020) as a benchmark. This dataset has a **class imbalance with 32% positives and 68% negatives examples** which effects the prediction accuracy of the minority class. Therefore, we propose a novel idea to generate synthetic samples in the data preprocessing step to balance the class distribution. We performed preliminary analysis on five different classifiers using scikit learn toolbox. Basis the results, we developed our custom classifier to improve the prediction for imbalanced datasets using our proposed novel idea.

Literature Review

We did a thorough research on the various techniques used by different researchers to deal with class imbalance. Our literature review comprises of 3 main components (1) Benchmarking the predictive model performance on this dataset, (2) Most extensively used SMOTE (**S**ynthetic **M**inority **O**versampling **T**echnique (Chawla et. al., 2002) and its modified extensions in the subsequent years, (3) SMOTE variants using clustering-based methods and identify a research gap to propose a **"k-median based clustering in place of the k-means clustering or a combination of both as suitable"**.

Benchmarking machine learning performance on this dataset

Chicco and Jurman (2020) have applied nine basic machine learning classifiers and one ensemble learning method (Boosting) to predict if a patient with a given set of symptoms is likely to survive or die in very near future. They best metrics they achieved are: Accuracy 0.585, F-score 0.754 and

0.418 MCC (Mathew's Correlation Coefficient). This prediction is better than the state-of-the art clinical prediction done by medical practitioners (Buchal et. al., 2020, Grove MW, 2005) based merely on domain expertise and experience. Clinical predictions are often intuitive, inconsistent and sometimes influenced by a negative bias. There is a huge scope of improvement given the importance of correct prediction.

The dataset has 32% positives and 68% negatives examples. (Chicco and Jurman, 2020) noted that this class imbalance effects the rates of correct predictions for true positives vs true negative, but they do not address the imbalance. Therefore, we take the predictions of Chicco et. al., 2020 as benchmark and attempt to improve the predictions by addressing the class imbalance. We expect that balancing the class will remove the bias of the learner (classification algorithm) for one class over the other and improve the prediction accuracy for the smaller class (true positives).

There are many datasets in the health domain itself such drug development, cancer tumor precision. where we see imbalanced data.

The original SMOTE technique and its extensions

The original SMOTE proposed by (Chawla et al., in 2002) is to pick up a sample randomly from the minority class, find its k- nearest neighbors and generate random samples on the line joining the initially picked up sample and of one the nearest neighbors. There have been several extension and modification over the original SMOTE method. These can be classified into seven categories as mentioned in (Fernández et. Al., 2018) addressing one or more of the following aspects of the baseline SMOTE.

1. Initial sample is selection around which the new samples will be generated.
2. Changing the kind of interpolation is used to generate new data points form the initial selected sample and its nearest neighbor.
3. Using an adaptive generation instead of mere interpolation
4. Variations in integrating the oversampling of the minority class with under-sampling of the majority class.
5. Incorporating dimensionality changes during or after sample generation.
6. Relabeling the dataset after the generation of synthetic samples
7. Some techniques focus on filtering noise from the new samples

Out of these, we found the second class of SMOTE variants (interpolation variations) are the most interesting. Therefore, in this category we reviewed the cluster-based interpolations.

Clustering based SMOTE

The table below summarises the salient features of clustering based variants of SMOTE till date.

S.No.	Algorithm name	Features	Reference
1	AHC	First attempt to use clustering in class imbalance. Uses K-means to cluster the majority class.	(Cohen et. al., 2006)
2	CBSO	Hierarchical Clustering of minority class using k-means and k-nearest neighbours. A minority example is picked up randomly and synthetic samples are generated within the cluster it belongs to.	(Barua et al., 2011)
3	ProWSyn	Proximity weighted synthetic oversampling technique for imbalanced data set learning. Weights for the minority examples is based on sample's proximity information, i.e., distance from boundary	(Barua, Islam, & Murase, 2013)
4	MWSMOTE	The distance of most difficult-to-classify minority samples is analysed and a weights are assigned to them based on the distance of the example from the nearest minority example.	(Barua, Islam, Yao, & Murase, 2014)
5	MOT2LD	Minority Oversampling Technique based on Local Densities in Low-Dimensional Space (or MOT2LD in short). At first, all the training examples are mapped to a low dimension (LD) space, followed by clustering in this space. A weight (dot product of local minority density and local majority count) is assigned to the minority example.	(Xie et al., 2015)
6	DGSMOTE	Solves online data imbalance problem by integrating Extreme Learning Machine (ELM), and SMOTE. If there is a severe imbalance, the granulated division for major-class samples is done according to the samples' distribution characteristic, further the original examples are replaced by the obtained granule core.	(Gong & Gu, 2016)
7	CURE-SMOTE	Noisy samples and outliers are removed while clustering the minority class, generates random points between representative points and centre point.	(Ma & Fan, 2017)
8	SOMO	Utilises self-organising maps to produce two-dimensional representation of examples followed by clustering and artificial data generation	(Douzas & Bacao, 2017)

From the above table, we can clearly see that all most all of the cluster-based techniques perform the average or the L2 norm. It occurred to us that since the given class has a skewed distribution, using an average for clustering would not provide the best results.

SMOTE and SVM

Support Vector Machine (SVM) is a popular machine learning model that is well suited for most classification problems. One of the major challenges SVM faces is to get a correct classification of minority class objects. SVM does not work well for skewed data sets which makes it very difficult to get the optimal solution for the SVM. (Cervantes et. al., 2020) have done a comprehensive survey on impact how the performance of SVM is adversely affected by class imbalance. The survey shows that the classifier was severely affected when applied to imbalanced datasets especially when the ratio between the majority and the minority class was large. In such a case, the hyperplane obtained after training the dataset was biased towards the minority class which adversely affected the accuracy of the model. The survey also highlights that only a fraction (0.9%) of papers were published in 2018 on SVM focused on imbalanced datasets. This underscores two things: (1) poor performance of SVM on imbalanced datasets;(2) overlooked importance of research on the imbalanced datasets. Therefore, this became a good incentive for us to learn more about how we can apply SVM on imbalanced datasets without having to tradeoff on accuracy.

Novel idea

Hypothesis

Our hypothesis to improve the Cluster based SMOTE (CBSO) technique of (Barua et. al, 2011) is to use k**-median instead of k-means to generate synthetic samples**. Using median or mean would depend on the feature distribution and whether the data is normally distributed or skewed.

We call this idea ‘med-CBSO’.

We draw our **inspiration from our class lecture on ‘Unsupervised Learning and Clustering (Slide 5)’** where k-medians were proposed as one of the solutions to make clustering robust to outliers.

Advantages and challenges

Most algorithms use the state-of-the-art k-means to form clusters that minimizes the Euclidean distance (L2 norm) to converge to the cluster centers. This technique works well when the data is uniformly distributed. For an imbalanced dataset like ours, we propose to use k-median instead. This method minimizes the absolute distance (L1 norm) to obtain the clusters centers.

We anticipated it would work better in the cases where the features have a skewed distribution. In such cases the median would be a better representative of the central tendency of the data.

We understand the added computational complexity associated with calculation of L1 norms, but it was worth trying to see if this pre-processing technique could improve the accuracy and prediction on both the classes for any base classifier.

Algorithm

A brief algorithm of the CBSO is below and we highlight the step where we make the change.

1. Compute the number of artificial samples to be generated based on ***percentage*** balancing required (percentage = user specified).
2. Find the ***K-nearest neighbours*** *(K=user specified)* for each minority sample, x_i.
3. Formulate a density distribution function using $\widehat{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i}, where\ r_i = \frac{\Delta_i}{K}, i = 1, 2, 3, m_s = numbers\ of\ minority\ samples$
4. Again, calculate the number of minority points that need to be artificially generated from each sample x_i to match the percentage balance required and call it ,g_i.
5. Perform hierarchical clustering of the minority examples using k-means algorithm[*] as explained in (Barua et. al., 2011)
6. For each minority sample x_i, generate g_i new data points using the following: Select a sample y from i^{th} cluster to which x_i belongs and generate the new point as: $s = x_i + \alpha \times (y - x_i)$, where α is a random number in the range [0, 1].

[*] In **our novel technique** we have clustered the minority samples using the 'k-median' clustering of NumPy instead of the 'k-means' clustering. In the next section, we compare the clusters formed in the two cases and how it impacts the classification performance.

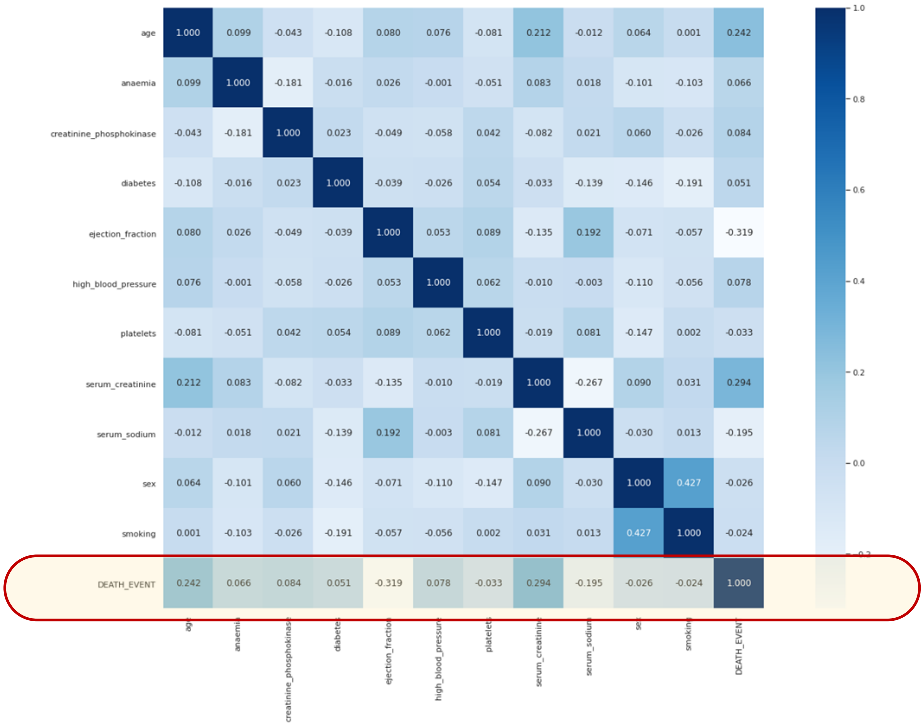
Experiments

Data pre-processing

We have used the standard procedure of data pre-processing. Our initial analysis on the dataset showed no missing values. We plotted the below histograms to study the data distribution and check for outliers. Some features are normally distributed, some highly skewed therefore, median clustering is expected to give better clusters. **There is a clear class imbalance.**

Feature Selection

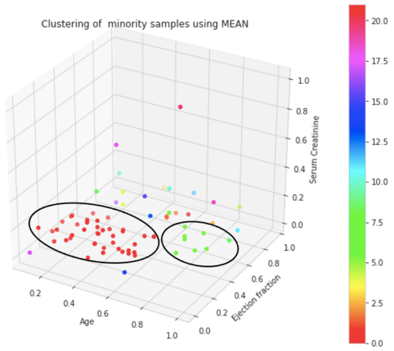
We used Pearson's method to obtain the feature correlation(heat map diagram below). Our ranking of features were same as (Chicco et. al., 2020). We performed the classification using just the top three features. The **top three correlated features** with target (death event) are ejection fraction(-0.32), serum creatinine (0.29) and age (0.24). Sex-smoking (0.43) also had a high correlation. However, in our case, even though feature selection was not a mandate for SVM, we were able to maintain the same prediction accuracy using just small subset of features. Original dataset is very small; therefore, we could not highlight the benefit of feature selection, however, for a larger dataset, having a tall and skinny input would positively impact the performance of the algorithm.



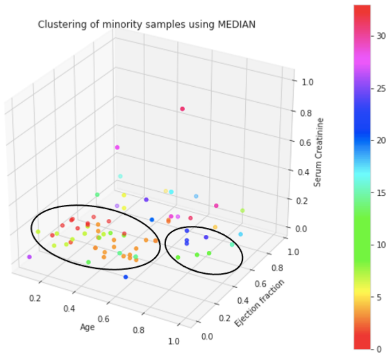
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	DEATH_EVENT
age	1.000	0.099	-0.043	-0.108	0.080	0.076	-0.081	0.212	-0.012	0.064	0.001	0.242
anaemia	0.099	1.000	-0.181	-0.016	0.026	-0.001	-0.051	0.083	0.018	-0.101	-0.103	0.066
creatinine_phosphokinase	-0.043	-0.181	1.000	0.023	-0.049	-0.058	0.042	-0.082	0.021	0.060	-0.026	0.084
diabetes	-0.108	-0.016	0.023	1.000	-0.039	-0.026	0.054	-0.033	-0.139	-0.146	-0.191	0.051
ejection_fraction	0.080	0.026	-0.049	-0.039	1.000	0.053	0.089	-0.135	0.192	-0.071	-0.057	-0.319
high_blood_pressure	0.076	-0.001	-0.058	-0.026	0.053	1.000	0.062	-0.010	-0.003	-0.110	-0.056	0.078
platelets	-0.081	-0.051	0.042	0.054	0.089	0.062	1.000	-0.019	0.081	-0.147	0.002	-0.033
serum_creatinine	0.212	0.083	-0.082	-0.033	-0.135	-0.010	-0.019	1.000	-0.267	0.090	0.031	0.294
serum_sodium	-0.012	0.018	0.021	-0.139	0.192	-0.003	0.081	-0.267	1.000	-0.030	0.013	-0.195
sex	0.064	-0.101	0.060	-0.146	-0.071	-0.110	-0.147	0.090	-0.030	1.000	0.427	-0.026
smoking	0.001	-0.103	-0.026	-0.191	-0.057	-0.056	0.002	0.031	0.013	0.427	1.000	-0.024
DEATH_EVENT	0.242	0.066	0.084	0.051	-0.319	0.078	-0.033	0.294	-0.195	-0.026	-0.024	1.000

Minority sample clustering

We observed that more clusters were formed when median is used as a cluster centre. The 74 minority samples were grouped into 34 and 22 clusters in the median and mean methods respectively as shown in figure below.

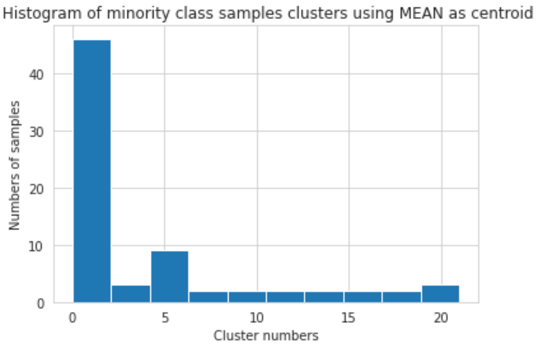


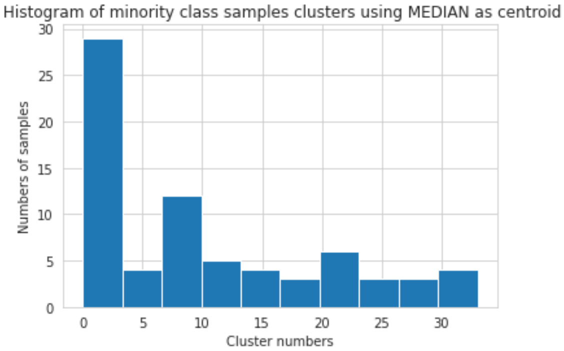
Clustering of minority samples using MEAN



Clustering of minority samples using MEDIAN

As show in the histogram below, median have more uniform allocation of samples of different clusters compared to mean which can be clearly seem from the 3D plots below.





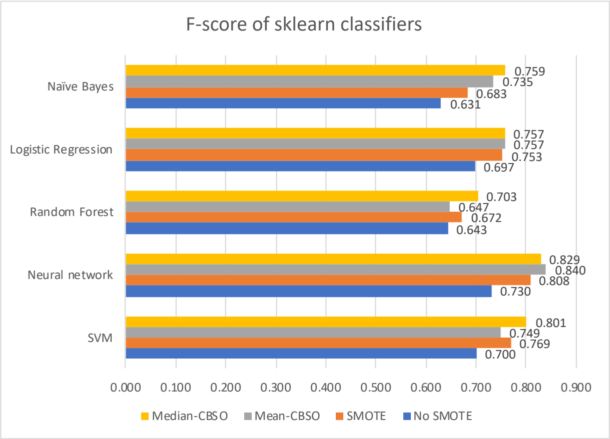
Model training and classification

After performing feature selection we study the effect our novel med-CBSO on five classifiers from the scikit learn library. All the models from scikit-learn were tuned using Grid Search. Out of these SVM showed the highest performance improvement using k-median instead of k-means for pre-processing (~5-7% improvement in prediction). Therefore we further tuned, and optimised the SVM beyond the scikit-learn toolbox.

We have implemented our SVM using Gradient Descent optimization technique. We tried both linear and Kernel SVM to understand the quality of the dataset after the synthetic samples were added to the minority classes, basically to check if the data is a linearly separable data in lower dimensions. Feature selection helped speed up the process without any trade-off in the prediction accuracy. The Kernel SVM gave the **best accuracy at 80%.**

Why SVM?

After changing the mean-based clustering to median-based clustering, we tested the changes on five different classifiers from the scikit-learn library (SVM, NN, Random forest, Logistic Regression and Naive Bayes) and compared their performance using the F-score, Accuracy and MCC as metrics. as in figure below.



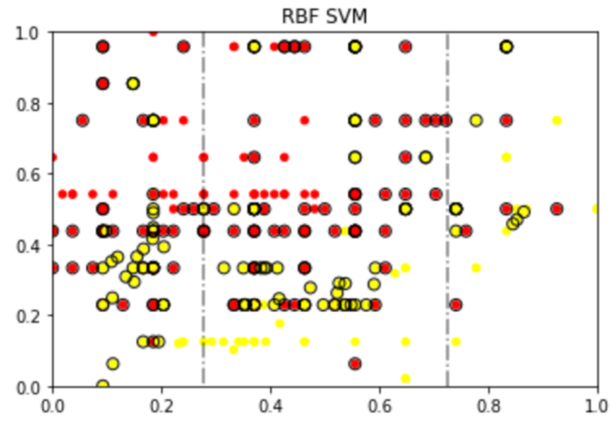
We have following observations:

- SVM showed the maximum improvement after balancing the minority class and proved to be the second-best classifier for this dataset.
- Neural network classifiers did not display any improvement over the SMOTE with balancing the minority classes using mean CBSO or our novel median-CBSO. It seemed like the best performance was driven by feature selection instead of the balancing the classes. NN had the least impact of after balancing the datasets.
- Random forest gave the worst accuracy on this data set and did not show any consistent improvement after class balancing.
- Logistic regression's accuracy did improve a little after adding synthetic samples to the minority class, but the output was is invariant to the type of the class balancing method that was used.
- Naive-Bias also shows improvement with class balancing, but overall ranks 3rd in performance score.

Since the SVM classifier has performance comparable to Neural-Network and displayed a worthy improvement with data balancing (**~78%**), we further decided to use SVM as our base classifier beyond the scikit learn toolbox to see if we can get better scores than the Neural-Net and study the effect of data balancing on SVM in detail.

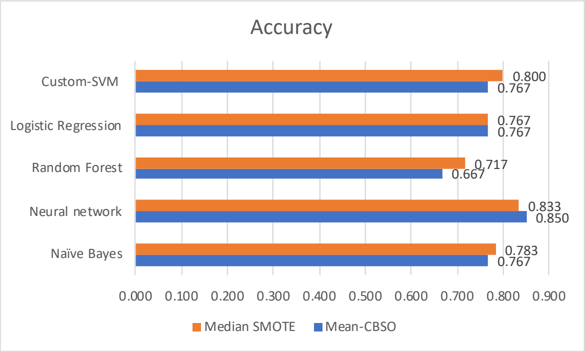
Results

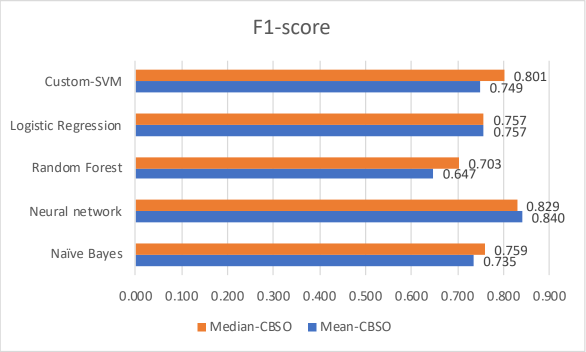
Hyperplane training – We solve the SVM dual problem using Gradient descent optimization technique. The data is imbalanced and non-linearly separable. Therefore, we have used the Kernel trick to map the input to a higher dimension.



We compare the Performance Metrics of SVM with other scikit-learn algorithms for k-mean (CBSO) vs k-median clustering based technique (med-CBSO)

From the below diagrams, we can see that SVM shows the best performance among all the classifiers except Neural networks and changing the balancing method from k-mean to k-median clustering for pre-processing the data gave higher performance in terms of both accuracy and f1-score .This is because SVMs are very susceptible to outliers. Using k-median clustering technique, ensures the clusters are centred around the median regardless of the condition of the dataset.





Conclusion

- For SVM, Random Forest and Naive bias Med-CBSO (our novel technique) is a better class balancing technique when compared to the state-of-the-art SMOTE and mean-CBSO.
- Although we could not beat the scores of NN for this dataset all our custom SVMs have better scores than the scikit-learn’s tuned SVM. Among these median-based SMOTE has the best performance.
- The datasets where SVM is the best classifier our novel technique should give the better. We will explore this in future.
- Mean-CBSO has the best performance for scikit-learn Neural Network classifier.
- Our performance scores better as compared to (Chicco et. al., 2020).
- “No One for all” stands true for data balancing techniques also. The choice of technique not only depends on the dataset, but also the type of classifier we use.
- For the given data set we did find median clustering does not impact computational time.

Future work

- We will compare the performance of our data set with other cluster based methods as summarised in the Literature review section and also larger datasets and higher IR (imbalance ratios).

- Implement the second part of our hypothesis, where mean and median can be used in combination using the feature distribution (skewed or normal)
- In terms of SVM, we could extend this project to
    - Use second order optimization techniques like Newton based methods like Interior Point Method (IPM), Alternating Direction Method of Multipliers (ADMM). Research says these techniques are less dependent on the condition of the Kernel, therefore, we could get a much higher accuracy and linear convergence in lesser iterations.
    - Understand the Kernel. If the Kernel is not full rank, we could approximate the Kernel to reduce it to a lower rank. Solving the approximated kernel would be much more stable with using the second order SVM-optimization techniques.
    - Additionally, we could use mixed precision instead of lower precision on GPUs to get the advantage of a better performance with a lesser trade-off on accuracy. This will be very beneficial to solve big data problems.

References

Buchal et. al., 2020, Physician Judgement vs Model-Predicted Prognosis in Patients with Heart Failure, Canadian Journal of Cardiology,Volume 36, Issue 1, Pp 84-91

Zahid FM, Ramzan S, Faisal S, Hussain I (2019) Gender based survival prediction models for heart failure patients: A case study in Pakistan. PLoS ONE 14(2): e0210602. https://doi.org/10.1371/journal.pone.0210602

Grove WM. Clinical versus statistical prediction: the contribution of Paul E. Meehl. J Clin Psychol 2005;61:1233-43

Barua, S., Islam, M. M., & Murase, K. (2011). A novel synthetic minority oversampling technique for imbalanced data set learning. In Neural Information Processing - 18th International Conference (ICONIP), pp. 735–744.

Sebastian Ruder, Insight Centre for Data Analytics, NUI Galway Aylien Ltd., Dublin: An overview of gradient descent optimization algorithms.

Jair Cervantesa, Farid Garcia-Lamonta, Lisbeth Rodríguez-Mazahuab, Asdrubal Lopez: 2019 - A comprehensive survey on support vector machine classification:Applications, challenges and trends. pp 189-191, 195-198

https://pypi.org/project/smote-variants/

(Used the smote-variants toolbox (version 0.4.0) and made changes in the code for clustering part to implement median based clustering instead of mean and further analysis).

The original code of SMOTE variant library can be found below:

https://smote-variants.readthedocs.io/en/latest/_modules/smote_variants/_smote_variants.html