BEÜ Fen Bilimleri Dergisi
10 (4), 1563-1574, 2021
DOI: 10.17798/bitlisfen.985519

BEU Journal of Science
10 (4), 1563-1574, 2021

Araştırma Makalesi / Research Article

Improving Machine Learning Performance of Imbalanced Data by Resampling: DBSCAN and Weighted Arithmetic Mean

Serkan GÜLDAL*

Adiyaman University, Art and Science Faculty, Physics Department, Adiyaman
(ORCID: 0000-0002-4247-0786)

Abstract

Improvement of digital technology has caused the collected data sizes to increase at an accelerating rate. The increase in data size comes with new problems such as imbalanced data. If a dataset is imbalanced, the classes are not equally distributed. Therefore, the classification of the data causes performance losses since the classification algorithms assume the datasets are balanced. While the classification favors the majority class, the minority class is often misclassified. To reduce the imbalanced ratio, various studies have been performed in recent years. In general terms, these studies are undersampling, oversampling, or both to balance the imbalanced datasets. In this study, an oversampling method is proposed employing distance combined with mean based resampling method to produce synthetic samples for the minority class. For the resampling process, the distances between pairs are calculated by the Euclidean distance metric between the minority class members. Based on the calculated distances, the denser zones are identified in the sense of DBSCAN around every datum. The new synthetic samples are formed between the points in the zones and central points by using the Weighted Arithmetic Mean. Thus, in this study, the dataset has been approximated 500 (majority) and 535 (from 268 minority data). Moreover, Random Forest (RF) and Support Vector Machine (SVM) algorithms are used for the classification of raw and balanced datasets. The result showed that the proposed method has the best machine learning performance among all the listed methods.

Keywords: Machine Learning, Random Forest, Support Vector Machine, Synthetic Data, Medical Data

Dengesiz Verilerin Yeniden Örnekleme ile Makine Öğrenimi Performansını İyileştirilme: DBSCAN ve Ağırlıklı Aritmetik Ortalama

Öz

Dijital teknolojinin gelişmesi, toplanan veri boyutlarının artan bir hızla artmasına neden olmuştur. Veri boyutundaki artış, dengesiz veri gibi yeni sorunları da beraberinde getirmektedir. Bir veri kümesi dengesizse, sınıflar eşit olarak dağıtılmamıştır. Bu nedenle, sınıflandırma algoritmaları veri kümeleri dengelenmiş varsayımı ile tasarlandığından, veriler sınıflandırılırken performans kayıplarına neden olur. Sınıflandırma çoğunluk sınıfını desteklerken, azınlık sınıfı genellikle yanlış sınıflandırılır. Veri setlerinin dengesizliklerini azaltmak için son yıllarda çeşitli çalışmalar yapılmıştır. Genel anlamda, bu çalışmalar veri kümelerini dengelemek için yetersiz örnekleme, aşırı örnekleme veya her ikisi şeklindedir. Bu çalışmada, sentetik numuneler üretmek için ortalama ile birleştirilmiş uzaklık tabanlı azınlık sınıfını yeniden örnekleme yönteminin kullanıldığı bir aşırı örnekleme yöntemi önerilmiştir. Yeniden örnekleme işlemi için azınlık sınıfındaki çiftler arasındaki uzaklıklar Öklid uzaklık metriği ile hesaplanır. Hesaplanan mesafeler göz önünde bulundurularak, yoğun bölgeler DBSCAN yöntemi dikkate alınarak her veri noktası etrafında tanımlanır. Yeni sentetik numuneler, Ağırlıklı Aritmetik Ortalama kullanılarak bölgenin içinde kalan noktalar ile merkez noktalar arasında oluşturulur. Böylece bu çalışmada veri seti 500 (çoğunluk) ve 535 (268 azınlık verisinden) olarak yeniden tanımlanmıştır. Ham ve dengeli veri kümelerini Rassal Orman (RF) ve Destek Vektör Makinesi (SVM) algoritmaları ile sınıflandırılmıştır. Sonuçlar önerilen yöntemin listelenen tüm yöntemler arasında en iyi makine öğrenimi performansa sahip olduğunu göstermiştir.

Anahtar Kelimeler: Makine Öğrenimi, Rastgele Orman, Destek Vektör Makinesi, Sentetik Veri, Tıbbi Veri

*Corresponding author: SrknGldl@hotmail.com
Received: 20.08.2021, Accepted: 12.10.2021

1. Introduction

Machine Learning (ML) methods are widely used in data research such as the medical diagnosis of diseases [1]. More specifically, supervised ML methods are trained by the available medical data for the prediction of medical diagnosis, so high-quality data becomes crucial to increase accuracy of the diagnosis. One of the important factors is the balance ratio between labeled data. The nature of the collected medical data is being imbalanced. The imbalanced data has a worsening effect on the performance of predictive ML algorithms on the raw dataset [2, 3]. To increase the accuracy of the prediction, the classes need to be balanced [4, 5]. Several methods have been developed to prevent performance losses such as Random Over Sampling (ROS) [6], Random Under Sampling (RUS) [7, 8], and Synthetic Minority Oversampling Technique (SMOTE) [9, 10].

Imbalanced data indicates that the majority class is more assertive in classification methods, and the minority class is generally ignored [9, 11]. Therefore, increasing misclassified cases reduces the accuracy and other performance measurements of the model [12]. Additionally, ROS, RUS, and SMOTE have been proposed to improve the balance, so the performance of the classifier algorithm [13]. SMOTE, which produces synthetic data, is one of the well-known sampling methods which increases the accuracy rate [9, 14].

Before synthetic data generation, the pairs in the dataset need to be identified based on the data topology. In our study, we use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) since it prioritize the denser areas of minority data distribution. DBSCAN is a clustering algorithm which finds the central data point(s) and expands the classification around it [15, 16]. Thus, it solely adopts the data distribution. DBSCAN is applied in many fields such as anomaly detection, medical imaging, and video processing [17-22]. In our study, we consider every data point a central point, so we identify the closest neighbors for the rest of the data set with the specified point. More details are given in the following sections.

In this study, a synthetically generated sample by means of the Weighted Arithmetic Mean (WAM) approach uses a dataset with an imbalanced distribution of diabetes patients. The study aims to balance the raw data to increase accuracy. The balanced dataset is classified by Random Forest (RF) and Support Vector Machine (SVM) algorithms, and the results are presented. As performance indicators, Accuracy (Acc), Precision (P), Recall (R), F1 score (F1), and Area Under Receiver Operating Characteristic curve (ROC) values are taken into account.

2. Materials and Method

2.1. Dataset Used

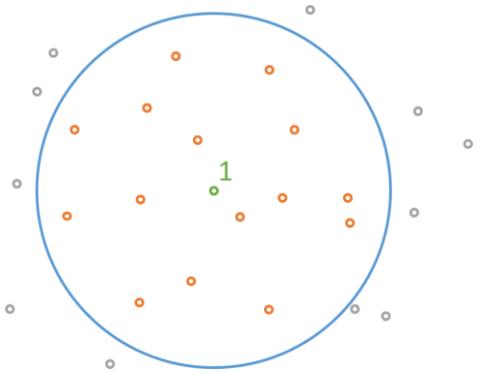
In this study, Pima Indians real diabetes dataset is used from KEEL (Knowledge Extraction based on Evolutionary Learning) opensource software tool site [23]. The dataset purposes to diagnose whether a patient is diabatic based on related diagnostic measurements. In this dataset, a total of 768 patient women were subjected, including 500 of whom not having diabetes and 268 of whom having diabetes. 8 attributes are available, namely Pregnancies, Glucose, Blood pressure (mm Hg), Skin thickness (mm), Insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years), and Outcome (Diabetic = 1 and Non-Diabetic = 0). Therefore, the results of the trained ML model show whether the patient is diabetic.

The imbalance ratio between the majority and minority classes is 53%. In this study, the diabetic patient population in the minority class is resampled and approximated to the majority class to balance the available data.

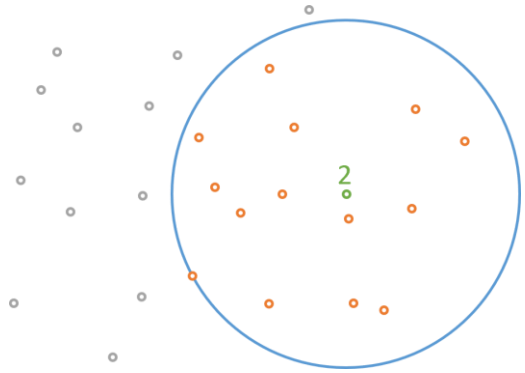
2.2. Proposed Method

In this study, the Euclidean distance metric is used to identify the distance between pairs. DBSCAN algorithm defined the neighbors around the selected datum with a specified range in the minority class. In Figure 1, the DBSCAN method is shown for the same data and 2 different points. In Figure 1.a, datapoint 1 is paired with 15 other datapoints. In Figure 1.b, datapoint 2 is paired with 14 other datapoints. Likewise, pairs are defined for all data points. If the total number of data points does not

equal or more than the missing number of data, the zone radius is increased and then the same process goes over. The zone is expanded until the required number of data points is obtained.



a) Datapoint 1 is shown by the color green. The zone is defined by a blue circle. The paired neighbors are colored by orange



b) Datapoint 2 is shown by the color green. The zone is defined by a blue circle. The paired neighbors are colored by orange

Figure 1. DBSCAN method pairs the central point with the neighbors in the zone.

By using Weighted Arithmetic Mean (WAM), among the pairs, synthetic data is generated in an amount close to the number of data required to balance. The steps of the developed method are listed as follows.

- First, the imbalance ratio is calculated by dividing the number of samples into classes. If the dataset has considerably imbalanced, further steps are applied.

- The Euclidean distances are calculated between minority pairs. If the considered data points x and y, all attributes are included to calculate the distance

$$x = [x_1, x_2, \ldots . x_n]^T \\ y = [y_1, y_2, \ldots . y_n]^T \tag{1}$$

- The Euclidean distance metric is shown by formula 2.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

- For each data point, a zone is determined then the remaining samples within the zone were paired with the selected datum. The radius of the zone is extended until the data is balanced by using brute force.

- We use Weighted Arithmetic Mean (WAM) to generate the synthetic samples between pairs. The WAM for vectors is defined as follows: Let α be a random number from [0,1]. For the given x and y in equation 3, the equation to generate WAM is given in equation 4.

$$x = [x_1, x_2, \ldots . x_n]^T \in R_+^n \\ y = [y_1, y_2, \ldots . y_n]^T \in R_+^n \tag{3}$$

$$x \, \Delta \, y = \begin{bmatrix} x_1(1-\alpha) + \alpha y_1 \\ x_2(1-\alpha) + \alpha y_2 \\ \vdots \\ x_n(1-\alpha) + \alpha y_n \end{bmatrix} \tag{4}$$

- Therefore, for generating synthetic samples, we use the following formula,

$$S_{new} = x \, \Delta \, y \tag{5}$$

- Formula 5 is repeated until all of the selected pairs are consumed.

The introduced method is depicted by a flowchart diagram in Figure 5 in the appendix.

2.3. Classifiers: Random Forest and Support Vector Machine

2.3.1. Random Forest

There exist numerous classification techniques such as Decision Trees, K-Nearest Neighbor, Random Forest, Support Vector Machine. Relatively, its simplicity, comprehensibility, and high predictive efficiency, decision trees are the most preferred method among these. Although decision trees have many advantages over the other classifiers, they have some disadvantages, including inconsistency. The inconsistencies are eliminated by constructing a random forest made of decision trees. Random Forest (RF) algorithm makes predictions based on several classifiers rather than a single classifier to improve the performance. In general terms, RF takes a random subset of variables to obtain a split at each node of the trees. For classification, the input vector is transferred to each tree in the algorithm, and an overall decision is done by each tree casts a vote for one of the classes. The class that has the most votes is chosen by the algorithm [24].

Throughout our analysis, the training dataset makes up 60% of the data and the test dataset makes up 40%. The training dataset is used to train the tree, while the test dataset is used to figure out the generalized error rate of the tree. The analysis is based on 1500 runs of RF to extract the statistically significant result and to prevent any error is done by the selection.

2.3.2. Support Vector Machine

Vapnik suggested the Support Vector Machine (SVM) algorithm as a modern solution for pattern recognition problems [25]. The SVM algorithm maps the data into a high-dimensional feature hyperspace in order to find the best separating hyperplane by optimizing the margin between classes. The SVM algorithm is known as a supervised machine learning algorithm. In the algorithm, new objects can be classified using an SVM classifier by adequate training and testing results. The SVM algorithm is a proven method for various classification problems including medical diagnostics and text characterization. In our study, we used SVM classifiers based on the SVM algorithm [26]. Mathematical representation is shown in equation 6 which indicates the separating hyperplane for the classes from the training set.

$$< w, z > + b = 0 \tag{6}$$

where w is a vector-perpendicular to the separating hyperplane, and b is the shortest distance between the origin and the hyperplane. Also $< w, z >$ is the dot product of w and z. The objects that are nearest to the separating hyperplane are precisely on the margins [26].

2.4. Model Performance Measurements

For binary classification, a confusion matrix is shown in **Table 1** to evaluate the performance of machine learning methods. The columns present predictions and the rows present the actual values. In the confusion matrix, TP, FP, TN, FN represents True Positive, False Positive, True Negative, and False Negative respectively.

Table 1. Confusion matrix for binary classification

	Predicted class Positive (Diabetic = 1)	**Predicted class Negative (Non- Diabetic = 0)**
Actual class Positive (Diabetic = 1)	True Positive (TP)	False Negative (FN)
Actual class Negative (Non-Diabetic = 0)	False Positive (FP)	True Negative (TN)

Based on the confusion matrix, a variety of evaluation metrics are created. In the present study, widely accepted measurements such as Accuracy, Recall, Precision, F1 score, and Area Under Receiver Operating Characteristic curve were utilized as the performance indicators [27].

Accuracy is defined by the ratio of samples correctly classified to the number of all samples.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

The recall is the ratio of samples correctly classified as positive to the total number of positive samples.

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

Precision is the ratio of samples correctly classified as positive by the total number of samples classified as positive.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

F1 score is the harmonic mean of Precision and Recall scores.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

3. Results and Discussions

The Pima Indians Diabetes dataset is used throughout the study. In the study, the performance increase is aimed by means of DBSCAN with the Weighted Arithmetic Mean (WAM) resampling method. The total number of patients in the dataset is 768. 500 of them are non-Diabetic (majority class, 0) and 268 of them are Diabetic (minority class, 1). The minority class is resampled synthetically in order to minimize the imbalance rate and provide a balanced dataset. To obtain perfect balance minority class needs to resample 232 times.

Based on the DBSCAN method, the paired neighbor zone is identified by brute force. In **Figure 2**, the zone for DBSCAN is expanded, and for the specified zone, a total number of paired neighbors are given. The sufficient number of neighbors which 267 are obtained for radius 17. 267 synthetic samples were generated from minority class samples by WAM, yielding a total number of 535 samples for the minority class.

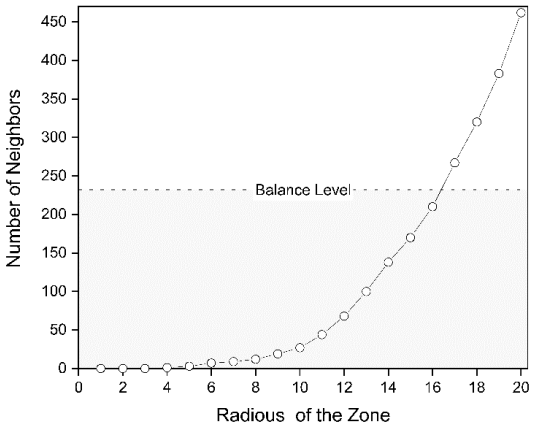


Figure 2. The zone is expanded until 20-unit length radius. For radius 17, 267 pairs which are sufficient to balance the dataset are identified.

The PIMA dataset is presented by Age and Glucose attributes in 2-dimensional graph to illustrate the imbalanced and balanced dataset classes distributions. The original data and balanced data are represented in **Figure 3**and 4, respectively.

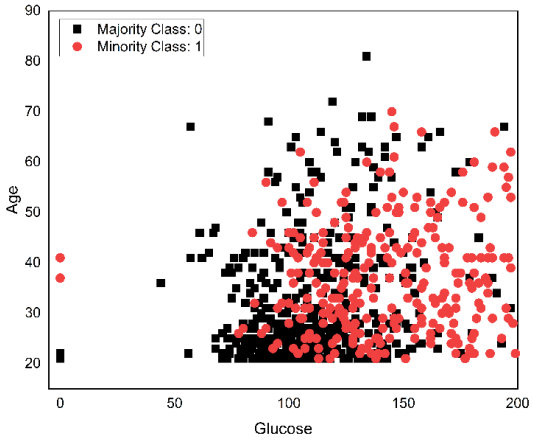


Figure 3. (Color online) The raw dataset is 500 majority and 268 minority classes.

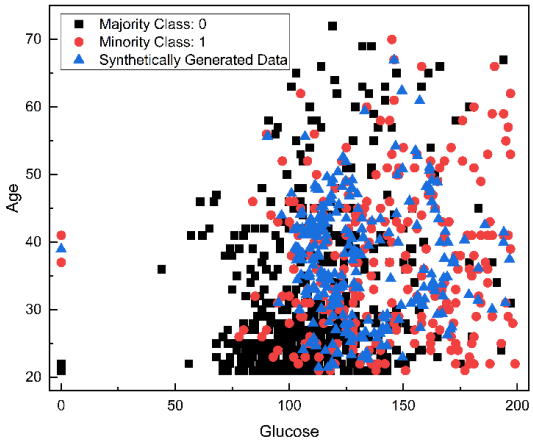
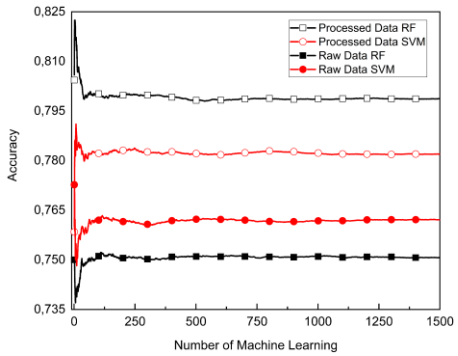


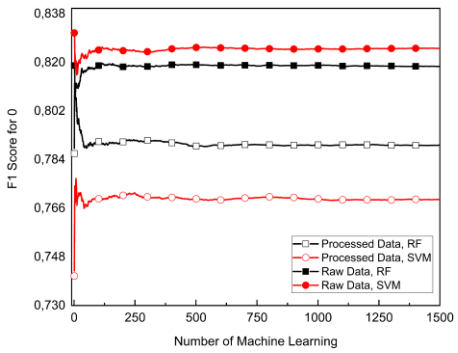
Figure 4. (Color online) Resampled data with the original data. Additional to the original data, 267 synthetic data are generated by DBSCAN.

In this study, we used Random Forest (RF) and Support Vector Machine (SVM) as the classifiers. The data is divided into sections randomly as 60% for training and 40% for testing. The classifications are repeated 1500 times for the same datasets with different training and test sets to prevent any selection bias.

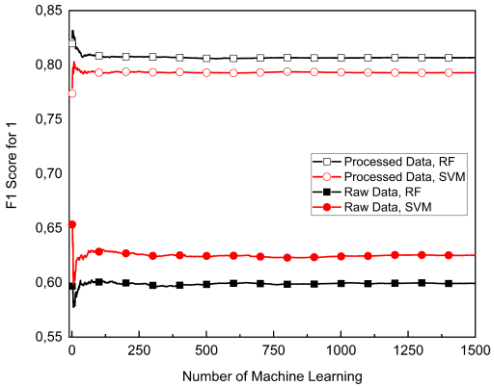
We have used Accuracy, Recall, Precision, F1 score, and Area Under Receiver Operating Characteristic curve (ROC) to measure the performance of ML. **Figure** .a shows the cumulative accuracy of the predictions which are done by ML. The results show that RF has a slight performance improvement, but SVM is significantly benefited from synthetically generated data. In **Figure** .b and c, cumulative F1 scores are shown for minority class, 1, and majority class, 0. While synthetic data improve the F1 score of the minority class, the majority class is negatively affected. The measurements show a similar result for Recall and ROC (See **Figure** .f, g, h, and i), but Precision. The cumulative precision score for the majority class shows an atypical behavior in **Figure** .d. While SVM has improved Precision score, RF has affected otherwise. In the study, both classification algorithms show a similar trend with various amounts. This is not the case for the Precision score of the majority class as in **Figure** .e. Lastly, the measurement results are averaged and summarized in **Table 2** and **Table 3**.



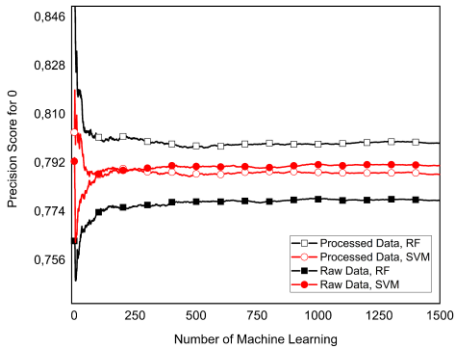
a) Cumulative accuracy scoring shows balanced dataset has better performance.



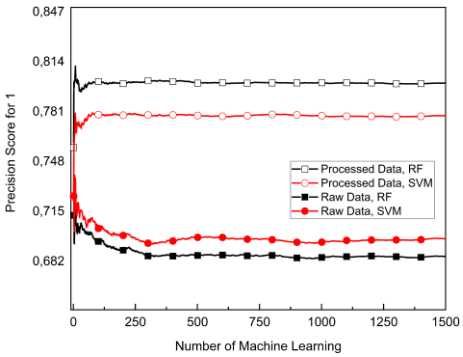
b) 0 which is initially the majority class has a relatively lower cumulative F1 score after the dataset is balanced.



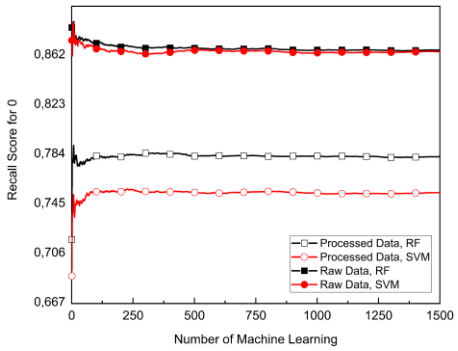
c) 1 which is initially the minority class has a relatively higher cumulative F1 score after the dataset is balanced.



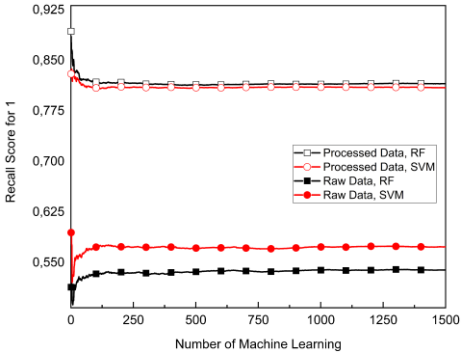
d) 0 which is initially majority class has a relatively lower cumulative Precision after the dataset is balanced for SVM classification. However, there is an increment of the Precision score for RF classification



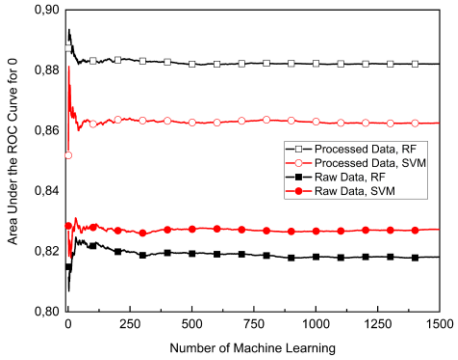
e) 1 which is initially a minority class has a higher cumulative Precision score after the dataset is balanced.



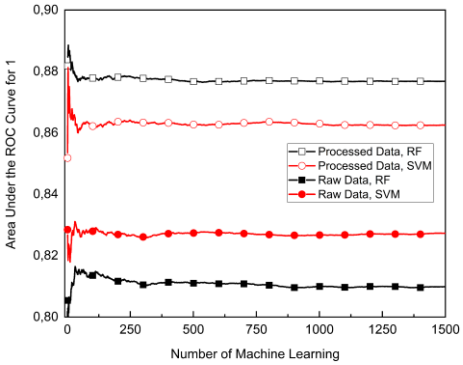
f) 0 which is initially majority class has a lower cumulative Recall score after the dataset is balanced.



g) 1 which is initially minority class has a higher cumulative Recall score after the dataset is balanced.



h) 0 which is initially the majority class has a lower cumulative ROC. After the dataset is balanced, there is performance improvement.



i) 1 which is initially the minority class has a lower cumulative ROC. After the dataset is balanced, there is performance improvement.

Figure 5. Cumulative performance measurements of ML classifications for 1500 repeats

The average of the measurements provides a conclusive result as shown in **Table 2** and **Table 3** for RF and SVM respectively.

In **Table 2**, the original dataset and the synthetically generated datasets (by RUS, ROS, and SMOTE) are trained and tested by RF and SVM. The synthetically generated data provide an improvement on performance scores. The accuracy score of original data increased from 0.751 to 0.737, 0.757, and 0.785 for Raw, RUS, ROS, and SMOTE respectively. However, our proposed, DBSCAN combined with WAM, the method has the best performance, 0.798 in the listed methods. Similarly, our method has the best score for other quality measurements. For example, for minority class, the Precision score is increased from 0.685 to 0.799, the Recall score is increased from 0.539 to 0.815, the F1 score is increased from 0.599 to 0.806, and ROC is increased from 0.809 to 0.876. Additionally, there is an improvement in the Precision (from 0.778 to 0.799) and ROC (from 0.818 to 0.882) value of the majority class (0)

Table 2. Performance values of Pima dataset classification results with Random Forest

Dataset	Class	Accuracy	Precision	Recall	F1 Score	ROC
Raw	0	0.751	0.778	**0.865**	0.819	0.818
	1		0.685	0.539	0.599	0.809
RUS	0	0.737	0.733	0.749	0.738	0.817
	1		0.746	0.726	0.733	0.818
ROS	0	0.757	**0.891**	0.733	**0.827**	0.817
	1		0.505	0.710	0.588	0.817
SMOTE	0	0.785	0.789	0.784	0.784	0.863
	1		0.784	0.786	0.786	0.863
DBSCAN WAM	0	**0.798**	0.799	0.781	0.789	**0.882**
	1		**0.799**	**0.815**	**0.806**	**0.876**

Likewise, all datasets are trained and tested under the same condition by SVM. The summarized results are given in Table 3. Synthetically generated datasets by RUS and ROS shows lower performance than Raw data. The accuracy score is from 0.762 to 0.738 and 0.737 for RUS and ROS respectively. The SMOTE algorithm provides slide improvement of Accuracy. The proposed algorithm has the best Accuracy performance of all by 0.781. Similarly, the other quality factors for minority class are the best between the listed synthetic data generation methods such as 0.778, 0.809, 0.792, and 0.862 for Precision, Recall, F1 Score, and ROC respectively. Additionally, ROC performance for the majority class is the best (0.862) by the proposed method.

To sum up, the classification results are summarized in Table 2 and Table 3, the minority group of the resampled dataset generally produced the best performance results against the raw data in all metrics. Additionally, the comparison of classification algorithms shows that the Random Forest is more successful than the Support Vector Machine result for synthetically generated datasets from the considered dataset.

Table 3. Performance Values of Pima Dataset Classification Results with SVM Algorithm.

Dataset	Class	Accuracy	Precision	Recall	F1 Score	ROC
Raw	0	0.762	0.791	**0.864**	**0.825**	0.827
	1		0.697	0.573	0.625	0.827
RUS	0	0.738	0.741	0.736	0.737	0.830
	1		0.738	0.741	0.738	0.830
ROS	0	0.737	**0.895**	0.738	0.808	0.822
	1		0.481	0.736	0.579	0.822
SMOTE	0	0.764	0.776	0.745	0.759	0.850
	1		0.755	0.784	0.768	0.850
DBSCAN WAM	0	**0.781**	0.787	0.753	0.769	**0.862**
	1		**0.778**	**0.809**	**0.792**	**0.862**

4. Conclusions

Improvement of digital technology causes increasing data collection. In the collected data, the imbalanced dataset is an emerging problem. A high imbalance ratio between the data classes reduces the Machine Learning (ML) performance considerably. To remedy ML performance, a synthetic sample generating method is introduced in this study. The Euclidean distance metric is used to calculate the distance between minority class. The DBSCAN methodology is used to identify the dense zones around every point. The zones are expanded until a sufficient number of pairs are obtained. Therefore, between the points in the zone and the central points, synthetic data is generated by Weighted Arithmetic Mean (WAM). The imbalanced and balanced datasets are classified using the Random Forest (RF) and Support Vector Machine (SVM). When the quality metrics of ML for raw and resampled datasets were compared, the resampled dataset with proposed methods are showed better performance measurement. Additionally, the performance measurements show the proposed method has the best of the listed methods such as ROS, RUS, and SMOTE. The best accuracy performance is 0.798 and ROC performances are 0.882 (for majority class, 0) and 0.876 (for minority class, 1), and the best minority class scores for Precision, Recall, and F1 Score are respectively, 0.799, 0.815, and 0.806 are obtained by the proposed resampling method.

To sum up, the result of the experimental study, the dataset balanced using the proposed method based on DBSCAN combined with WAM is more successful than the raw dataset and the other listed methods.

Acknowledgments

The authors gratefully share their appreciation to İbrahim Halil Gümüş and Mustafa Yavaş. Technical discussion with them, and their encouragement is a key part to mature this research.

The Declaration of Publishing Ethics

The author declares that this study complies with Research and Publication Ethics.

References

[1] Gopinath M., Aarthy S., Manchanda A. 2019 Machine Learning on Medical Dataset. in Information Systems Design and Intelligent Applications, S. C. Satapathy, V. Bhateja, R. Somanah, X.-S. Yang, and R. Senkerik Eds. Singapore: Springer. 133-143.

[2] He H., Garcia E. A. 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21 (9): 1263-1284.

[3] Weiss G. M. 2004. Mining with rarity: a unifying framework. SIGKDD Explorations Newsletter, 6 (1): 7–19.

[4] Mohammed A. J., Hassan M. M., Kadir D. H. 2020. Improving classification performance for a novel imbalanced medical dataset using SMOTE method. International Journal, 9 (3): 3161-3172.

[5] Rahman M. M., Davis D. N. 2013. Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing, 3 (2): 224-228.

[6] Hayati M., Muthmainah S., Ghufran S. 2021. Random and synthetic over-sampling approach to resolve data imbalance in classification. International Journal of Artificial Intelligence Research, 4 (2): 86-94.

[7] Zuech R., Hancock J., Khoshgoftaar T. M. 2021. Detecting web attacks using random undersampling and ensemble learners. Journal of Big Data, 8 (1): 1-20.

[8] Elhassan T., M A., F A.-M., Shoukri M. 2016. Classification of imbalance data using tomek link (T-Link) combined with random under-sampling (RUS) as a data reduction method. Global Journal of Technology and Optimization, 01.

[9] Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P. 2002. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16: 321-357.

[10] Yavaş M., Güran A., Uysal M. 2021. Covid-19 veri kümesinin SMOTE tabanlı örnekleme yöntemi uygulanarak sınıflandırılması. Avrupa Bilim ve Teknoloji Dergisi: 258-264. [Online]. Available: https://dergipark.org.tr/tr/pub/ejosat/issue/56356/779952.

[11] Han H., Wang W.-Y. Mao B.-H. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, In: International Conference on Intelligent Computing: Springer, 878-887.

[12] Chawla N. V., Japkowicz N., and Kotcz A. 2004. Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 6 (1): 1-6.

[13] Kovács G. 2019. Smote-Variants: A python implementation of 85 minority oversampling techniques. Neurocomputing, 366: 352-354.

[14] Hassan G. A. A. M., Yıldırım D., Masoud 2021. Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data, 10, doi: https://dergipark.org.tr/tr/pub/bitlisfen/939733.

[15] Ester M., Kriegel H.-P., Sander J., Xu X. 1996, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, presented at the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon.

[16] Schubert E., Sander J., Ester M., Kriegel H. P., Xu X. 2017. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. ACM Transactions on Database Systems, 42 (3): Article 19.

[17] Bilgin T., Çamurcu Y. 2005. DBSCAN, OPTICS ve K-Means Kümeleme Algoritmalasının Uygulamalı Karşılastırılması.

[18] Dokuz A. S., Çelik M., Ecemis A. 2020. DBSCAN Algoritması Kullanarak Bitcoin Fiyatlarında Anormallik Tespiti.

[19] Yaşar H., Albayrak M. Comparison of serial and parallel programming performance in outlier detection with DBSCAN algorithm. Bilecik Şeyh Edebali Üniversitesi Fen Bilimleri Dergisi, 7 (1): 129-140.

[20] Alhussein I., Ali A. H. 2020. Application of DBSCAN to Anomaly Detection in Airport Terminals. In: 2020 3rd International Conference on Engineering Technology and its Applications (IICETA), 6-7 September, Iraq, 112-116.

[21] Baselice F., Coppolino L., Antonio S. D., Ferraioli G., Sgaglione L. 2015. A DBSCAN Based Approach for Jointly Segment and Classify Brain MR Images, In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 25-29 August, 2993-2996.

[22] Huan Y., Wenhui Z. 2013. DBSCAN Data Clustering Algorithm for Video Stabilizing System, In: Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC), 25-29 August, 1297-1301.

[23] KEEL. "Pima Indians Diabetes Dataset." KEEL. https://sci2s.ugr.es/keel/dataset.php?cod=21 (accessed 12.04.2021.

[24] Liaw A., Wiener M. 2002. Classification and Regression by Random Forest. R news, 2(3): 18-22.

[25] Vapnik V. 2013. The Nature of Statistical Learning Theory, 2nd ed. New York, USA: Springer Science & Business Media.

[26] Demidova L., Klyueva I., Sokolova Y., Stepanov N., Tyart N. 2017. Intellectual approaches to improvement of the classification decisions quality on the base of the SVM classifier. Procedia Computer Science, 103: 222-230.

[27] Fatourechi M., Ward R. K., Mason S. G., Huggins J., Schlögl A., Birch G. E. 2008. Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets, presented at the Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications. [Online]. Available: https://doi.org/10.1109/ICMLA.2008.34.

Appendix

The algorithm flowchart of the proposed resampling method is shown in Figure 5.

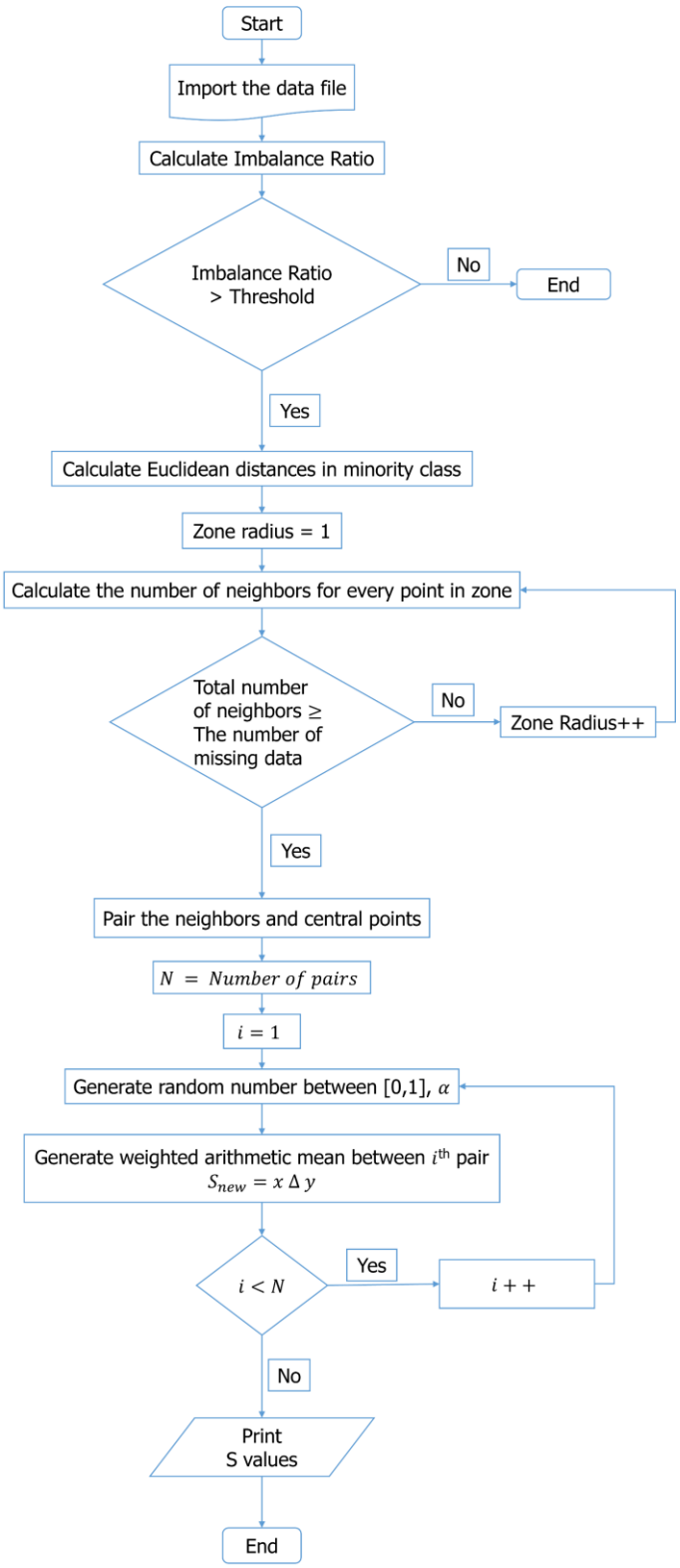


Figure 5. Algorithm flowchart of resampling method by DBSCAN combined with Weighted Arithmetic Mean