Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2021;13(2):138-46

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2020-80268

Comparison of Unbalanced Data Methods for Support Vector Machines

Destek Vektör Makineleri İçin Dengesiz Veri Yöntemlerinin Karşılaştırılması

Pelin AKIN[a], Yüksel TERZİ[b]

[a]Department of Statistics, Çankırı Karatekin University Faculty of Science, Çankırı, TURKEY
[b]Department of Statistics, Ondokuz Mayıs University Faculty of Science, Samsun, TURKEY

This study was presented as a oral International Conference on Data Science, Machine Learning and Statistics (DMS-2019), June 26-29, 2019, Van, Turkey.

ABSTRACT Objective: The biggest problem we encounter when applying classification algorithms is that the classification categories are not equally distributed. Eight different re-sampling methods were used for balancing the dataset. **Material and Methods:** Support Vector Machines (SVM) were used to compare these methods. SVM are supervised learning models with associated learning algorithms that analyze data used for categorization and regression analysis. The main function of the algorithm is to find the best line, or hyperplane, which divides the data into two classes. SVM is basically a linear classifier that classifies linearly separable data, but, in general, the feature vectors might not be linearly separable. To overcome this issue, what is now called kernel trick was used. **Results:** This article presents a comparative study of different kernel functions (linear, radial, and sigmoid) for unbalanced data. The myocardial infarction dataset which was taken from the Github were classified by 10-fold cross validation to increase the performance. Accuracy, sensitivity, specificity, precision, g-mean and F score were used for comparing the methods. The analysis was carried out by R software. **Conclusion:** As a conclusion, the results of performance metrics for the original data increased through random over sampling examples re-sampling methods for linear and sigmoid kernel functions. Smote method performed better in the case of radial kernel. In general, the unbalance in the data in classification algorithms gives biased results and this should be eliminated.

Keywords: Support vector machine; unbalanced dataset; data sampling methods

ÖZET Amaç: Sınıflandırma algoritmalarını uygularken karşılaştığımız en büyük problem, sınıflandırma kategorilerinin eşit dağılmamasıdır. Veri kümesini dengelemek için 8 farklı yeniden örnekleme yöntemi kullanılır. **Gereç ve Yöntemler:** Bu yöntemleri karşılaştırmak için destek vektör makineleri [support vector machines (SVM)] kullanıldı. SVM, sınıflandırma ve regresyon analizi için kullanılan verileri analiz eden ilişkili öğrenme algoritmalarına sahip denetimli öğrenme modellerindendir. Algoritmanın ana görevi, verileri 2 sınıfa ayıran en doğru hattı veya hiper düzlemi bulmaktır. SVM, temelde doğrusal olarak ayrılabilir verileri sınıflandıran doğrusal bir sınıflandırıcıdır, ancak genel olarak özellik vektörleri doğrusal olarak ayrılamayabilir. Bu sorunun üstesinden gelmek için çekirdek hilesi kullanılır. **Bulgular:** Bu makalede, dengesiz veriler için farklı çekirdek işlevlerinin (doğrusal, Radyal ve Sigmoid) karşılaştırmalı bir çalışması verildi. Github'dan alınan miyokardiyal enfarktüs veri seti, performansı artırmak için 10 kat çapraz doğrulama kullanıldı. Yöntemlerin karşılaştırılmasında doğruluk, duyarlılık, özgüllük, kesinlik, Gmean ve F ölçüsü kullanıldı. Analiz, R yazılımı tarafından gerçekleştirildi. **Sonuç:** Sonuç olarak, doğrusal ve Sigmoid çekirdek fonksiyonları için "random over sampling examples" yeniden örnekleme yöntemi orijinal veriye göre performans ölçütlerinin sonuçlarını artırmıştır. Radyal çekirdek için Smote yönteminin performansı artmıştır. Sınıflandırma algoritmalarında verilerdeki dengesizlik yanlı sonuçlar verir ve bu problem ortadan kaldırılmalıdır.

Anahtar kelimeler: Destek vektör makinesi; dengesiz veri seti; veri örnekleme yöntemleri

In machine learning, as long as the accuracy rate is high, can we say that the model was successful? The biggest issue with classification algorithms is that classes are not equal. Therefore, even if the accuracy rate is high, the classier may be heavily biased toward the majority class. Majority class is char-

Correspondence: Pelin AKIN
Department of Statistics, Çankırı Karatekin University Faculty of Science, Çankırı, TURKEY/TÜRKİYE
E-mail: pelinakin@karatekin.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 27 Nov 2020 ***Received in revised form:*** 20 Feb 2021 ***Accepted:*** 22 Feb 2021 ***Available online:*** 15 Mar 2021

acterized by a large quantity of samples, while the minority class is the class that has fewer examples in the predictive modelling problem. Generally, we encounter this problem when applying classification algorithms in the field of health. To overcome this problem, we need to resample by applying unbalanced data methods. We applied eight different methods in order to increase the model's performance.

After applying these methods to equalize the classes, the dataset was compared using the support vector model. Support vector machines (SVM) are beginning to see increased popularity compared to other algorithms in health care lately. The basis of this algorithm is to separate classes from each other by using a hyperplane. SVM is used for both linearly or non-linearly separable data. If the data are not separated linearly, kernel functions are used to move it to a higher dimension. In this study, sigmoid and radial kernel functions were used along with linear SVM.

Poggio and Girosi and Wahba discuss the use of kernels.[1,2] Zhang et al. proposed kernel-based scaling with SVM classification algorithm for unbalanced data.[3] Jian et al. suggested a new sampling method for classifying unbalanced data.[4] Fan et al. used three unbalanced learning algorithms, Synthetic Minority over sampling Technique (SMOTE), Borderline-SMOTE, and Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN), oversample a minority class, and finally, a training set of class-balanced is obtained. Then used SVM random forest to predict the voice pathology detection model. As a result, they showed that the RF models achieved perfect performance.[5] Zhang et al. proposed a hybrid of the Random Oversampling Sample, K-means, and Support vector machine (RK-SVM) model based on sample selection to overcome the unbalanced classification problem with two classes in breast cancer diagnosis.[6] The Matthew team proposed a weighted kernel-based oversampling algorithm, WK-SMOTE is proposed to balance the class distribution in an SVM classifier.[7]

With this study, we aimed to examine the classification success rate of the SVM algorithm on unbalanced data sets and to determine the most successful method. In this study, different from the studies in the literature, different unbalanced data methods were compared. The rest of our study consists of three main sections. In the second chapter, we explained the unbalanced data methods and SVM algorithms, also known as the most appropriate performance criterion that can be used for unbalanced data sets. In the third chapter, we listed the experiment results obtained in the fourth chapter, and the comments are in the last chapter.

MATERIAL AND METHODS

SUPPORT VECTOR MACHINES

SVMs are used for classification and regression, and are among the supervised learning methods.[8] Based on statistical learning theory, main principle of SVMs is to minimize structural risk. SVMs were developed by Vapnik and Alexei Chervonenkis in 1960. In contrast, however, the first paper on SVMs was published by Vladimir Vapnik and his colleagues Bernhard Boser and Isabelle Guyon in 1992. This method is used in many areas, which include recognition of hand written digits and recognition of objects, as well as speaker identification.[9] SVM was founded on the concept of identifying a hyperplane that better separates a dataset into two classes. SVM can be categorized under two groups: (a) **Linearly separable** data, (b) **Non-linearly separable** data.[10]

The Linearly Separable Case

In this case, there is a line (hyperplane) that separates the data into two distinct classes. There are many hyperplanes that can distinguish between two classes of data. We tried to find a plane that has the maximum margin, that is, the maximum distance between both classes' data points.

A separating hyperline can be shown as:

$$wx + b = 0$$

w is weight vector, and b is bias scaler.

There is a set of training tuples with associated class labels y_i. y_i can take one of two values, either +1 or -1.[9] It can be shown as:

$wx + b \geq +1$ y=+1

$wx + b \geq +1$ y=-1

The maximal margin is$\frac{2}{|w|}$.[11] We need to solve the following equations;

$$min\frac{(\|w\|)^2}{2}$$

$$y_i(wx + b) \geq 1 y_i \epsilon(-1, +1)$$

They are called support vectors. When this optimization problem is solved with the Lagrange multipliers method, we get:

$$L_P(w, b, a) = \frac{1}{2}\|w\| - \sum_{i=1}^{l} \alpha_i [\, y_i(wx + b) - 1]$$

Then, the classification function becomes:

$f(x) = \text{sgn}\big((wx_i) + b\big) = \text{sign}(\sum_{i=1}^{l} y_i \alpha_i (x_i x_j))$.

The Non-linearly Separable Case

A large portion of issues encountered in the real world involve non-separable data, where there is no one hyperline that can successfully separate the training set. One of the solutions to this inseparability problem involves mapping the data to a dimensional space that is relatively higher, and then defining a hyperline there; at which case we call this dimensional space that is relatively higher the feature space, which is different from the input space submerged in the training instances.[12] Under the mapping solution to the SVM has the form

$$f(x) = sgn(\sum_{i=1}^{l} y_i \alpha_i \, \varphi(x_i)\varphi\big(x_j\big) + b)$$

In SVM, the only quantities that need computation are scalar products$(x_i)\varphi\big(x_j\big)$, which are their key properties. This is called kernel function.[9] Using this quantity solution, SVM can be shown as:

$$f(x) = sgn(\sum_{i=1}^{l} y_i \alpha_i \, K(x_i, x_j) + b)$$

Two popular choices for kernel function in the this article are

$$-\|X_i, X_j\|^2 \Big/ 2\sigma^2$$

Radial Basis :$K\big(X_i, X_j\big) = e$

Sigmoid Kerneli: $K\big(X_i, X_j\big) = \tanh\,(kX_i, X_j - \delta)$.[10]

MODEL PERFORMANCE METRICS FOR CLASSIFICATION

To measure the success of the model, the data set is divided into a training set (train set) and a test set (test set). The data in the training set is used to train the model. It uses the data from the test set to measure the performance of the model. Sometimes a validation set is used to fine-tune the parameters of the algorithm. The difference between the test set and the validation set is the data in the test set that the system has not seen before.[13] In this study, the k-fold cross-validation method is used. In the cross-validation method, the data set is divided into k equal parts. While k-1 number of clusters is used in training, the remaining part is used as test data. This process is repeated k times and each time the accuracy value of the model is found by averaging the calculated accuracy values. In this example, the data set is divided into two as training and test sets at a certain rate. This ratio is generally taken as 80% -20%, 70% -30%, 60% -40%. This study examined the results of these three ratios. It was continued according to the results of the 60-40% separation rate, which shows the best performance result.

A confusion matrix is used to describe the performance of a classification algorithm. When we have an unequal number of observations in all classes, or when we have more than two classes in the dataset, we cannot depend on the classification accuracy alone as it can be misleading more often than not. If "no" is too high in the data set, the specifity is expected to be high, while the sensivity is expected to be low. In this case, when the resampling method is applied, the sensivity and precision is expected to increase and "no" is required to be estimated correctly. In these studies, it is more accurate to compare with Fscore and gmean, not the accuracy rate. Because these values are calculated with precision, sensitivity and specifity values. Therefore, we can get a better idea of how our classification model performs and what kind of errors it makes if we calculate a confusion matrix. Confusion matrix is given in Table 1.

TABLE 1: Confusion matrix.

		Actual	
		Yes	**No**
Predict	Yes	True positives (TP)	False negatives (FN)
	No	False positives (FP)	True negatives (TN)

Some performance criteria are given to determine the classification performance.

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

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$F_{score} = \frac{2 * \text{Sensitivity} * \text{Precision}}{\text{Sensitivity} + \text{Precision}}$$

$$G_{mean} = \sqrt{\text{Sensitivity} * \text{Specificity}}$$

METHODS FOR DEALING WITH UNBALANCED DATA

One of the common issues when attempting classification with machine learning is the problem of class imbalance. If the classification categories are in equally distributed, we can say the dataset is unbalanced.

Therefore, it is safe to say that, when that happens, the classifier can be heavily biased toward the majority class.[14] To solve this problem, under-sampling and over sampling methods are used. Under-sampling refers to the reducing of the majority class by removing random observations until the dataset is balanced.[15] Over-sampling methods produce additional minority class subjects based on the data observed.[16] In this study, we assessed eight different methods of under-sampling and over-sampling to equalize the class distribution of the training data.

Synthetic Minority Over Sampling Technique

SMOTE is an over-sampling method developed by Chawla.[17] In SMOTE, we over-sample the minority class by picking out a sample from each minority class and then creating synthetic examples. These examples sit along the line segments of all the nearest neighbour soft the k minority class. Synthetic samples are producing the following ways: Firstly, the algorithm computes the distance between the feature vectors and their nearest neighbours. After that, we multiply the difference by a random number between (0, 1), and add it back in the feature. Following that, we under-sampling the majority class by randomly extracting samples from the majority class before the minority class reaches a certain percentage of the majority class.[17] The **DMwR** package was used for SMOTE. The parameters of algorithms ***perc.over*** and ***perc.under*** control the quantity of over-sampling seen in the minority class and under-sampling seen in the majority classes, respectively. In this study, we used three different values of perc.over and perc.under.

Over Sampling Algorithms Based on Synthetic Minority Over Sampling Technique

SMOTE is a pioneering algorithm, and it was used in the creation of various other algorithms.

Density-based synthetic minority over-sampling technique

Based on a density-based concept of clusters, Density-based Synthetic Minority Over-sampling Technique (DBSMOTE) method is mainly used in oversampling an arbitrarily shaped cluster identified by Density-based spatial clustering of applications with noise. DBSMOTE functions as a way to produce synthetic instances along a shortest path from each positive instance to a pseudo-centroid of a minority-class cluster.[18]

Safe level synthetic minority over-sampling technique

SMOTE merges the minority instances along a line randomly, connecting a minority instance and it's selected nearest neighbours while ignoring nearby majority instances. Safe Level Synthetic Minority Over-Sampling Technique (SLSMOTE) samples minority instances very thoroughly along the same line with a different weight degree, called safe level. The safe level carries out the computing by using nearest neighbour minority instances. Through taking the safe level ratio of all the cases, t-generation of each synthetic instance takes places in a safe position.[19]

Random Over Sampling Examples

Random over sampling examples (ROSE) (random on sampling samples) is based on a smoothed out bootstrap form of resampling from data and is based on the kernel method.

Random Under-sampling

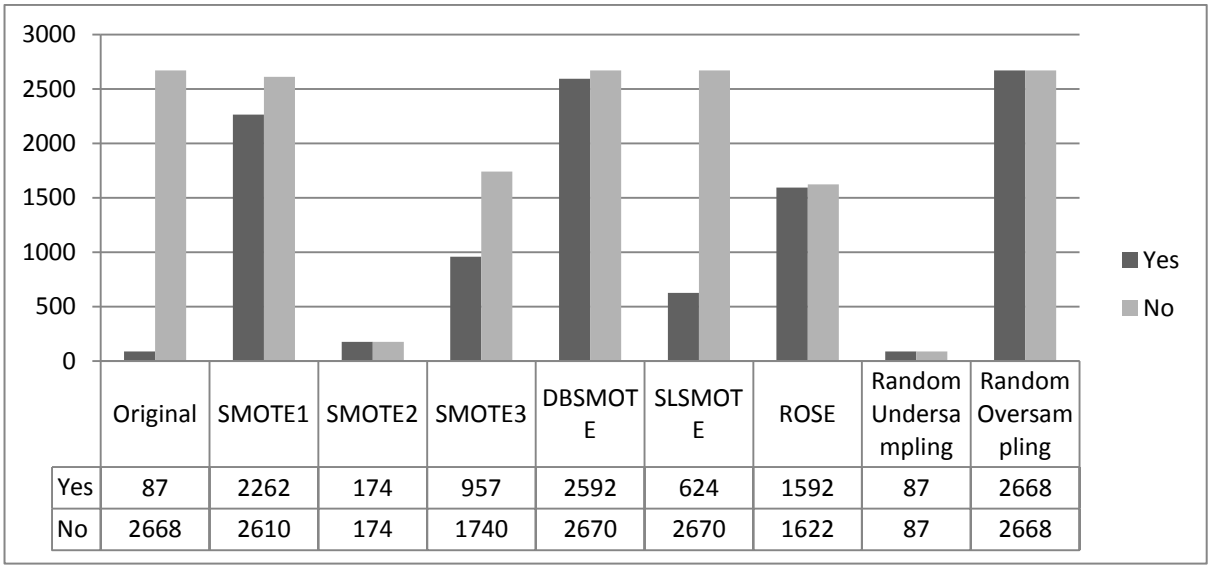
Random undersampling techniques randomly remove samples of majority class in order to establish a distribution.[20]

Random Over-sampling

The random oversampling copies a certain quantity of minority class samples in a random fashion, and then adds them to the dataset.[20]

RESULTS

The dataset used in this study, the myocardial infarction dataset, was taken from Github. It includes 4,590 patients and 27 different attributes. The data set was seperated by 60% training and 40% test data. Each resampling method was applied to the training data. SVM models were established on the balanced training data. Then, using the test data, the performance of the models was examined. The mean of each performance measure over 100 runs is given, and the standard deviation is shown in parentheses. Figure 1 shows an overview of the dataset used. The portion of the minority class in the original data increased from 3% to approximately 50% after the aforementioned methods were applied.



	Original	SMOTE1	SMOTE2	SMOTE3	DBSMOTE	SLSMOTE	ROSE	Random Undersampling	Random Oversampling
Yes	87	2262	174	957	2592	624	1592	87	2668
No	2668	2610	174	1740	2670	2670	1622	87	2668

FIGURE 1: Overview of datasets used.

SMOTE: Synthetic minority over sampling technique; DBSMOTE: Density-based synthetic minority over-sampling technique; SLSMOTE: Safe level synthetic minority over-sampling technique; ROSE: Random over sampling examples.

Linear SVM was applied after separating the data set by 60% training and 40% test data. We compared the prediction success of the algorithms on the test data. Table 2 shows the results of the linear SVM analysis after the resampling method is applied. When the F score and g-mean values are examined, the best model is ROSE. In the original data, the estimated percentage of patients who had a heart attack increased from zero to 65%. The F score of this model was 0.22 and the g-mean value was 0.75.

TABLE 2: Result of linear support vector machines.

Model	Sensitivity	Specificity	Precision	F_{score}	Accuracy	Gmean
Original	0.000 (0)	0.9963 (0)	NA NA	NA NA	0.9689 (0)	0.000 (0)
SMOTE1	0.6137 (0.0659)	0.7390 (0.0134)	0.1065 (0.0097)	0.1814 (0.0162)	0.8278 (0.0122)	0.7145 (0.0363)
SMOTE2	0.7653 (0.0683)	0.6895 (0.0259)	0.0872 (0.0073)	0.1548 (0.0122)	0.7653 (0.0238)	0.7262 (0.0259)
SMOTE3	0.5121 (0.0633)	0.8279 (0.0104)	0.1275 (0.0137)	0.2039 (0.0213)	0.8757 (0.0095)	0.6727 (0.0403)
DBSMOTE	0.5486 (0.0694)	0.8033 (0.0164)	0.1172 (0.0175)	0.1924 (0.0253)	0.8555 (0.0148)	0.6864 (0.0165)
SLSMOTE	0.2783 (0.0681)	0.9135 (0.0109)	0.1958 (0.0478)	0.2254 (0.0462)	0.9403 (0.0097)	0.5094 (0.0237)
Random under-sampling	0.7002 (0.0158)	0.7323 (0.0053)	0.087 (0.0030)	0.1554 (0.0052)	0.7635 (0.0057)	0.7321 (0.0111)
Random over-sampling	0.7544 (0)	0.7074 (0)	0.1012 (0)	0.1784 (0)	0.7841 (0)	0.7696 (0)
ROSE	0.6473 (0.0175)	0.7912 (0.0039)	0.1346 (0.0004)	0.2228 (0.0008)	0.8597 (0.0032)	0.7489 (0.0095)

SMOTE: Synthetic minority over sampling technique; DBSMOTE: Density-based synthetic minority over-sampling technique; SLSMOTE: Safe level synthetic minority over-sampling technique; ROSE: Random over sampling examples.

SVM classifier with sigmoid kernel was applied and compared the prediction success of the algorithms on the test data. Table 3 shows the results of the SVM classifier using sigmoid kernel analysis for each re-sampling method. When the F score and g-mean values are examined, the best model is ROSE. In the original data, the estimated percentage of patients who had a heart attack increased from 0 to 72%. The F score of this model was 0.18 and the g-mean value was 0.75.

TABLE 3: Support vector machines classifier using sigmoid kernel.

Model	Sensitivity	Specificity	Precision	F_{score}	Accuracy	Gmean
Original	0.0057 (0.0086)	0.969 (0.0012)	0.0456 (0.0748)	0.0317 (0.0059)	0.9656 (0.0012)	0.0429 (0.0630)
SMOTE1	0.6519 (0.0649)	0.8347 (0.0245)	0.0743 (0.0059)	0.1334 (0.0103)	0.7364 (0.0226)	0.6928 (0.0291)
SMOTE2	0.7336 (0.0238)	0.7677 (0.0257)	0.0869 (0.0066)	0.1553 (0.0109)	0.7513 (0.0238)	0.7417 (0.0248)
SMOTE3	0.5773 (0.0737)	0.8874 (0.0156)	0.0971 (0.0092)	0.1661 (0.0158)	0.8200 (0.0139)	0.6897 (0.0403)
DBSMOTE	0.7973 (0.0631)	0.8655 (0.0208)	0.0922 (0.0117)	0.1599 (0.0183)	0.7973 (0.0190)	0.7009 (0.0083)
SLSMOTE	0.4106 (0.0096)	0.9618 (0.0207)	0.1351 (0.0264)	0.2004 (0.0334)	0.8977 (0.0178)	0.6047 (0.0203)
Random under-sampling	0.7712 (0.0070)	0.7655 (0.0003)	0.0845 (0.0006)	0.1524 (0.0012)	0.7335 (0)	0.7515 (0.0034)
Random over-sampling	0.7893 (0.0017)	0.7851 (0.0001)	0.0796 (0.0037)	0.1446 (0.0024)	0.7099 (0.0002)	0.7472 (0.0002)
ROSE	0.7175 (0.0076)	0.8665 (0.015)	0.100 (0.0023)	0.1751 (0.0026)	0.7890 (0.0076)	0.7534 (0.010)

SMOTE: Synthetic minority over sampling technique; DBSMOTE: Density-based synthetic minority over-sampling technique; SLSMOTE: Safe level synthetic minority over-sampling technique; ROSE: Random over sampling examples.

SVM classifier using radial kernel was applied and calculated the prediction success of the algorithms on the test data. As indicated by the outcomes in Table 4, if we consider the F score and g-mean, SMOTE performed best. In the original data, the estimated percentage of patients who had a heart attack increased from 0% to 65%. The F score of this model was 0.16 and the g-mean value was 0.72.

TABLE 4: Support vector machines classifier using radial kernel.

Model	Sensitivity	Specificity	Precision	F_{score}	Accuracy	Gmean
Original	0.0559 (0.0249)	0.9984 (0.0012)	0.5817 (0.0748)	0.1028 (0.0059)	0.9691 (0.0011)	0.2277 (0.0630)
SMOTE1	0.3409 (0.0602)	0.9203 (0.0079)	0.1206 (0.0178)	0.1779 (0.0266)	0.9023 (0.0072)	0.5578 (0.0491)
SMOTE2	0.6551 (0.0605)	0.7968 (0.0257)	0.3941 (0.0066)	0.1644 (0.0109)	0.7924 (0.0238)	0.7211 (0.0248)
SMOTE3	0.3495 (0.0567)	0.9237 (0.0078)	0.1283 (0.0181)	0.1873 (0.0263)	0.9059 (0.0072)	0.5662 (0.0463)
DBSMOTE	0.1221 (0.0492)	0.9811 (0.0036)	0.1706 (0.0543)	0.1402 (0.0481)	0.9541 (0.0031)	0.3336 (0.0330)
SLSMOTE	0.1054 (0.0488)	0.9914 (0.0041)	0.2966 (0.1196)	0.1513 (0.0589)	0.9636 (0.0041)	0.2922 (0.0894)
Random under-sampling	0.6830 (0.0123)	0.7643 (0.0059)	0.0851 (0.0028)	0.1513 (0.0047)	0.7617 (0.0061)	0.7225 (0.0094)
Random over-sampling	0.1752 (0.0017)	0.9837 (0.0002)	0.2560 (0.0037)	0.2081 (0.0025)	0.9586 (0.0002)	0.4152 (0.0022)
ROSE	0.4380 (0.0053)	0.9304 (0.0017)	0.1680 (0.0023)	0.2428 (0.0013)	0.9151 (0.0015)	0.6384 (0.0033)

SMOTE: Synthetic minority over sampling technique; DBSMOTE: Density-based synthetic minority over-sampling technique; SLSMOTE: Safe level synthetic minority over-sampling technique; ROSE: Random over sampling examples.

DISCUSSION

In this study, we applied support vector methods to determine the patients that had a heart attack. When applying classification algorithms, the biggest problem in health data proved to be that the data is unbalanced. Because the number of patients who have had a heart attack is low, models have trouble predicting who had a heart attack. In order to eliminate this problem, we used eight different methods and compared the results with the original data. ROSE method performed best in Linear SVM and SVM with sigmoid kernel, while smote was the best method in SVM with radial kernel. Looking at the comparison results of these three SVM methods, the best method proved to be SVM with sigmoid kernel. Because 72% of patients who had a heart attack predicted correctly.As in similar studies, unbalanced data methods increased prediction success in this study.[3,5]

CONCLUSION

The biggest problem we encounter when implementing classification algorithms is that the data is unbalanced. This problem reflects as high accuracy rate on paper. If we make predictions with this model with high accuracy, our results will be biased. When implementing classification algorithms, the analysis should be carried out after the removal of unbalanced data. After solving the unbalanced data problem, a SVM suitable for the data structure was chosen. SVM varies according to the style of the data. When applying SVM in a study, not only linear but other kernel methods should be tried. In this study, it is suggested that we can use the SVM model to predict whether it will have a heart attack or not. After eliminating the unbalanced data issue, the estimation rate of the patients that had a heart attack increased from 0% to 72%.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Pelin Akın;* ***Design:*** *Pelin Akın;* ***Control/Supervision:*** *Yüksel Terzi;* ***Analysis and/or Interpretation:*** *Pelin Akın;* ***Literature Review:*** *Pelin Akın;* ***Writing the Article:*** *Pelin Akın;* ***Critical Review:*** *Yüksel Terzi.*

REFERENCES

1. Poggio T, Girosi F. Networks for Approximation and Learning. Proceedings of the Ieee. 1990;78(9):1481-97. [Crossref]
2. Wahba G. Spline Models for Observational Data. 1st ed. Philadelphia: SIAM; 1990. [Crossref]
3. Zhang Y, Fu PP, Liu WZ, Chen GL. Imbalanced data classification based on scaling kernel-based support vector machine. Neural Computing & Applications. 2014;25(3-4):927-35. [Crossref]
4. Jian CX, Gao J, Ao YH. A new sampling method for classifying imbalanced data based on support vector machine ensemble. Neurocomputing. 2016;193:115-22. [Crossref]
5. Fan Z, Qian J, Sun B, Wu D, Xu Y, Tao Z, editors. Modeling Voice Pathology Detection Using Imbalanced Learning. 2020 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD); 2020. p.330-4. IEEE. [Crossref]
6. Zhang J, Chen L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. Comput Assist Surg (Abingdon). 2019;24(sup2):62-72. [Crossref] [PubMed]
7. Mathew J, Pang CK, Luo M, Leong WH. Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines. IEEE Trans Neural Netw Learn Syst. 2018;29(9):4065-76. [Crossref] [PubMed]
8. Jakkula V. Tutorial on support vector machine (svm). School of EECS, Washington State University. 2006;37. [Link]
9. Han J, Kamber M. Data Mining: Concepts and Techniques. 2nd edition. San Francisco, CA, USA: Morgan Kaufmann Publishers; 2006.
10. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. London: Springer; 2013. [Link]
11. Alpaydin E. Introduction to Machine Learning (Adaptive Computation and Machine Learning). 2nd ed. The MIT Press; 2010.
12. Maglogiannis IG. Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies. 1st. Netherlands: Los Press; 2007. [Link]
13. Uğuz S. Makine Öğrenmesi Teorik Yönleri ve Python Uygulamaları ile Bir Yapay Zeka Ekolü. 1. Baskı. Ankara: Nobel; 2020. [Link]
14. He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications. 1st. New Jersey: John Wiley & Sons; 2013. [Crossref]
15. Dal Pozzolo A, Caelen O, Bontempi G, eds. When is undersampling effective in unbalanced classification tasks? Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2015: Springer. [Crossref]
16. Blagus R, Lusa L. Joint use of over-and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics. 2015;16(1):363. [Crossref]
17. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002;16:321-57. [Crossref]
18. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. DBSMOTE: density-based synthetic minority over-sampling technique. Applied Intelligence. 2012;36(3):664-84. [Crossref]
19. Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C, eds. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. Pacific-Asia conference on knowledge discovery and data mining; 2009: Springer. [Crossref]
20. Wang X, Pedrycz W, Chan P, He Q, SpringerLink (Online service). Machine Learning and Cybernetics 13th International Conference, Lanzhou, China, July 13-16, 2014. Proceedings. Available from: [Crossref]