Available online at www.medicinescience.org

ORIGINAL RESEARCH

Medicine Science International Medical Journal

Medicine Science 2021;10(4):1510-5

Classification of stroke with gradient boosting tree using smote-based oversampling method

Fatma Hilal Yagin, Ipek Balikci Cicek, Zeynep Kucukakcali

Inonu University, Faculty of Medicine, Biostatistics and Medical Informatics, Malatya, Turkey

Received 29 September 2021; Accepted 04 November 2021
Available online 01.12.2021 with doi: 10.5455/medscience.2021.09.322

Abstract

The aim of this study is to classify the disease with the gradient increasing tree classification method in an open access dataset containing data from patients with and without stroke disease. In addition, it is aimed to compare the results by balancing the data with the oversampling method Synthetic Minority Over-sampling Technique (SMOTE) which is one of the data balancing methods in the study. In this study, a dataset containing information about patients with and without stroke disease obtained from the address "https://www.kaggle.com/asaumya/healthcare-problem-prediction-stroke-patients" was used. In the study, SMOTE was used as the data balancing method, and the gradient boosting tree method was used in the modeling. The performance of the model was evaluated by Specificity, sensitivity, accuracy, positive predictive value and negative predictive values. Specificity, sensitivity, accuracy, positive predictive value and negative predictive values were obtained as 0.0887, 0.9772, 0.9339, 0.9544 and 0.1679, respectively, according to the modeling result using the gardient boosting tree method using the original version of the dataset. Specificity, sensitivity, accuracy, positive predictive value and negative predictive values were obtained as 0.0887, 0.9772, 0.9339, 0.9544 and 0.1679, respectively, according to the modeling result using the gardient boosting tree method using the SMOTE applied version of the dataset. When the results obtained from the study were examined, the modeling results made with the SMOTE applied dataset were obtained more consistently and realistically. As a result, it is suggested that researchers use dataset balancing methods to acquire more accurate results whenever they come across an unbalanced dataset problem.

Keywords: Stroke, classification, gradient boosting tree, unbalanced data, SMOTE

Introduction

Stroke causes cognitive, motor, speech, language, and swallowing changes and is one of the most important causes of disability in adults [1]. According to the definition of the World Health Organization, stroke is a condition that occurs with dysfunction in that region of the brain as a result of occlusion or rupture of the vascular structures of the brain and is characterized by clinical manifestations ranging from motor loss, sensory impairment, balance disorder, visual impairment, speech and cognitive dysfunction to coma [2]. Stroke is one of the most common causes of death and is the main cause of permanent and acquired disability in adults worldwide. Considering demographic changes, further increases in stroke rates are expected. The World Health Organization refers to stroke as the impending epidemic of the 21st century. Stroke is the second most common cause of death and the second most common cause of disability worldwide [3, 4]. Stroke is a leading cause of mortality and disability. Although stroke mortality is decreasing, the prevalence of people living with the effects of stroke has increased because of the growing and aging population [5]. Stroke, which has an important place in hospital admission, can be fatal, as well as cause various bodily dysfunctions that require rehabilitation and care more often and cause individual, social and economic problems [6]. It is important to determine the prevalence and risk factors of a disease such as stroke with a high mortality and disability rate and to take measures accordingly, to reduce the death rate, to prevent loss of workforce, and to reduce the burden of the country's economy. If the modifiable risk factors can be controlled with the measures to be taken, a decrease in the incidence of stroke and related death can be observed.

***Corresponding Author:** Fatma Hilal Yagin, Inonu University, Faculty of Medicine, Biostatistics and Medical Informatics, Malatya, Turkey
E-mail: hilal.yagin@inonu.edu.tr

Data mining is the discovery of implicit, previously unknown and potentially useful information about data. Data mining techniques are used to analyze and select data using complex algorithms to discover unknown patterns. Researchers use data mining techniques to diagnose many diseases such as heart disease, diabetes and cancer, and thus data mining in healthcare is becoming more and more popular [7, 8]. Today, the need for data mining in the health and medical sector is more than the needs of other sectors. In this context, using data mining methods, valuable information can be extracted from large databases, which can help healthcare professionals make decisions and improve healthcare services [9]. It is desirable to have almost the same number of samples from each class in a dataset, that is, to have a balanced distribution. However, in cases where it is the opposite, that is, in unbalanced datasets, classifiers tend to tend towards the class setting where the number of samples is in the majority during training. Essentially, this is undesirable.

Therefore, it is important to balance the dataset so that the results are consistent and more unbiased. Therefore, data balancing methods are often used. The dataset used in this study is unbalanced and SMOTE, an oversampling method, was used. [10].

Ensemble methods are frequently preferred today because they combine classifiers and give better results than other machine learning methods. Boosting, one of the Ensemble methods is a machine learning method that combines low-performance classifiers into a more powerful classifier. In Boosting methods, the estimations are sequential, and this estimation process continues as each subsequent estimator learns from the errors of the previous estimators. Gradient Boosting Trees (GBT) is a widely used and highly successful algorithm in the Boosting category [11].

The aim of this study is to classify the disease with the gradient boosting tree classification method in an open-access dataset containing data from patients with and without stroke disease. In addition, the dataset used in the study is unbalanced and it is aimed to compare the results by balancing the data with SMOTE, which is an oversampling method, one of the unbalanced data problem methods.

Materials and Methods

Research design and dataset

In this study, a dataset containing information about patients with and without stroke disease obtained from the address "https://www.kaggle.com/asaumya/healthcare-problem-prediction-stroke-patients" was used. Since the dataset used in this study was published as open access, it does not require ethics committee approval. In the dataset used in this study, there are 4861 (95.1%) patients diagnosed with no stroke and 249 (4.9%) patients diagnosed with suffered stroke.

Since there is a class imbalance problem in the dataset (No stroke: 4861, Suffered stroke: 249), the SMOTE Upsampling operator was used in Rapidminer. Then, two different GBT models were created using 10-fold cross validation on the raw and balanced dataset.

Data Preprocessng

Smote

It is desirable to have almost the same number of samples from each class in a dataset, that is, to have a balanced distribution. However, in cases where it is the opposite, that is, in unbalanced datasets, the classifiers tend to tend towards the class set where the number of samples is in the majority while training. In other words, since the classifier will be trained with the class label with a large number of samples, a bias occurs in the system. In addition, since the classifier with the minority class label cannot train itself sufficiently, it cannot classify successfully in the following stages. Essentially, this is undesirable [10] By using resampling methods, unbalanced datasets can be made more balanced. To do this, the first method is to increase the minority class's data with various methods to obtain classes with an equal number of data (oversampling) [12].

The other method is to obtain a balanced dataset (undersampling) by removing the data belonging to the weighted class from the data set. However, if the studied dataset is not large enough, other methods should be preferred since this method will cause information loss [13]. Along with these, balance can be achieved with methods such as SMOTEENN[14] and SMOTETomek[15], which are a combination of oversampling and undersampling.

Cross Validation

In the study, the performance of the GBT classifier was tested by cross validation. There are various cross-validation methods in the literature. Holdout, K-Fold Cross-Validation, Stratified K-Fold Cross-Validation and Leave-P-Out Cross-Validation are some of them. 10-Fold cross validation was used in the study. In this method, the dataset is divided into 10 parts, one part is the test set and the remaining nine parts are the training set. After the classification process is made according to this partition, each remaining piece becomes the test set. The classification accuracy rate is calculated for each part, and the classification success is found after it is taken into the general environment [16].

Machine Learning Approaches

Gradient Boosting Trees (GBT)

Machine learning (ML) methods require more than just building models and making predictions to improve accuracy. Although machine learning methods mostly have high accuracy performance in classification processes, they cannot give the desired performance level in some data sets. There are many reasons why the desired level of performance is not achieved. Different solutions have been proposed to improve these performance losses experienced while learning is taking place. One of these methods is community learning methods. Ensemble learning methods, unlike the classification of a single machine learning algorithm, provide a common classification result from the predictions of each classifier by classifying the data separately by multiple machine learning algorithms. Thus, joint estimation results provide more accurate, more reliable and higher performance. Today, Ensemble methods are frequently used because of their better performance in industry or competitions. Boosting, one of the Ensemble methods, is an ML

method that combines poor performance classifiers into a stronger classifier. In boosting methods, the estimations are sequential, and this estimation process continues, with each subsequent estimator learning from the errors of the previous estimators. Gradient Boosting Trees (GBT) is a widely used and very successful algorithm in the Boosting category. This algorithm can be used for regression task in addition to classification. GBTs are pretty robust for scaling differences in variables in training data. Categorical and numerical variables also work easily and one of the advantages of the algorithm is that it minimizes differentiable loss functions. The number of trees, learning rate, maximum depth parameters are very important to improve model performance in GBT[17-19]. The pseudo code of the GBT Algorithm is as in Figure 1.

Input: A training dataset $S = \{(x_1, y_1), ..., (x_n, y_n)\}$, a loss function $L(y, F(x))$, the number of iterations M.
Output: A classification model F_M.

procedure ALGORITHMGBT(S, L, M)
  Initialize the model:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

  for $m \in \{1, ..., M\}$ **do**
    for $i \in \{1, ..., N\}$ **do**

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

    end for
    Get the terminal regions R_{jm} by fitting a regression tree to the r_{im} targets, where $j \in \{1, 2, ..., J_m\}$
    for $j \in \{1, ..., J_m\}$ **do**

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)$$

    end for

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

  end for
  Return F_M
end procedure

Figure1. The pseudo code of the GBT Algorithm

Data Analysis

Statistical analyzes were carried out in IBM SPSS 26.0 and modeling analyzes in Rapidminer 9.9 software. Quantitative data are summarized by median (minimum-maximum) and qualitative variables are given by number and percentage. Normal distribution was evaluated with the Kolmogorov-Smirnov test. In order to show whether there is a statistically significant difference between the categories of the dependent variable in terms of input variables, Mann-Whitney U test was used in the analysis of quantitative data and Pearson chi-square test was used in the analysis of qualitative data. $p<0.05$ values were considered statistically significant. In all analyzes, IBM SPSS Statistics 26.0 for the Windows package program was used.

Since there is a class imbalance problem in the dataset (No stroke: 4861, Suffered stroke: 249), the SMOTE Upsampling operator was used in Rapidminer. Then, two different GBT models were created using 10-fold cross validation on the raw and balanced dataset. The methods used in the study are described below.

Results

The distribution of the dependent variable in the dataset utilized in this investigation is shown in the table below.

Table 1. The distribution of the dependent variable

No stroke		Suffered stroke	
Count	Percentage (%)	Count	Percentage (%)
4861	95.1	249	4.9

Table 2 shows descriptive statistics for the independent variables in this study. According to this table, in terms of age, avg glucose level and BMI factors, there is a statistically significant difference between the groups of the dependent variable (stroke) ($p<0.05$).

Table 2. Descriptive statistics table of quantitative independent variables

Variables	Stroke		p-value*
	No stroke	Suffered stroke	
	Median (min-max)	Median (min-max)	
Age	48 (0.08-82)	71 (1.32-82)	<0.001
Avg Glucose Level	91.47 (55.12-267.76)	105.22 (56.11-271.74)	<0.001
BMI	28.0 (10.3-97.6)	29.7 (16.9-56.6)	<0.001

*: Mann Whitney U test

Table 3 shows that; there is a statistically significant relationship between hypertension, ever married, work type, smoking status and heart disease variables the dependent variable (stroke) groups ($p<0.05$).

Table 4 shows the classification matrix for the gradient boosting tree model built using the original dataset.

Table 5 shows the results of the classification performance criterion for the gradient boosting tree model. Specificity, sensitivity, accuracy, positive predictive value and negative predictive values obtained from the model were 0.0887, 0.9772, 0.9339, 0.9544 and 0.1679 respectively.

Table 6 shows the classification matrix of the gradient boosting tree model applied to classify stroke disease in the dataset balanced using the SMOTE method.

Table 4. The gradient boosting tree model's classification matrix in the original dataset

Characteristics	Reference		
Prediction	**Suffered stroke**	**No stroke**	**Total**
Suffered stroke	22	111	133
No stroke	227	4750	4977
Total	249	4861	5110

Table 3. Descriptive statistics for qualitative independent variables

Variables	Categories of Variables	Stroke		p-value*
		No stroke	Suffered stroke	
		Number (%)	Number (%)	
Age group	Male	2007 (41.3)	108 (43.4)	0.516
	Female	2853 (58.7)	141 (56.6)	
Marital status	No Hypertension	4429 (91.1)	183 (73.5)	**<0.001**
	Suffering from Hypertension	432 (8.9)	66 (26.5)	
	No	1728 (35.5)	29 (11.6)	**<0.001**
	Yes	3133 (64.5)	220 (88.4)	
Education Income level	Private	2776 (57.1)	149 (59.8)	**<0.001**
	Self Employed	754 (15.5)	65 (26.1)	
	Govt Job	624 (12.8)	33 (13.3)	
	Children	685 (14.1)	2 (0.8)	
	Never Worked	22 (0.5)	0 (0.0)	
	Urban	2461 (50.6)	135 (54.2)	0.269
	Rural	2400 (49.4)	114 (45.8)	
Service time Working status	Never Smoked	1802 (37.1)	90 (36.1)	**<0.001**
	Formerly Smoked	815 (16.8)	70 (28.1)	
	Unknown	1497 (30.8)	47 (18.9)	
	Smokes	747 (15.4)	42 (16.9)	
Occupational group	No heart disease	4632 (95.3)	202 (81.1)	**<0.001**
	Suffering from heart disease	229 (4.7)	47 (18.9)	

*: Pearson chi-square test

Table 5. The model's classification performance criteria's values in the original dataset

Metric	Value
Specificity	0.9772
Sensitivity	0.0887
Accuracy	0.9339
Positive predictive value	0.1679
Negative predictive value	0.9544

Table 6. Classification matrix of Gradient boost tree model on data with SMOTE method applied

Prediction	Reference		
	Suffered stroke	No stroke	Total
Suffered stroke	4536	785	5321
No stroke	325	4076	4401
Total	4861	4861	9722

Table 7 shows the results of the classification performance criterion for the gradient boosting tree model on SMOTE applied data. Specificity, sensitivity, accuracy, positive predictive value and negative predictive values obtained from the model were 0.9331, 0.8385, 0.8858, 0.9269 and 0.8536 respectively.

Table 7. Classification performance criteria values of the model in the SMOTE applied dataset

Metric	Value
Specificity	0.8385
Sensitivity	0.9331
Accuracy	0.8858
Positive predictive value	0.8536
Negative predictive value	0.9269

Discussion

Stroke is a sudden onset, a focal neurological syndrome that develops due to cerebrovascular disease. In addition, stroke is a

major medical and economic problem that causes severe disability and ranks 3rd among the causes of death after heart diseases and cancer [20, 21]. Although the incidence of stroke varies between regions, it also differs among people in the same country according to race and settlement. In the studies conducted in the last 20 years, the incidence of stroke was found to be 1-3/1000 and the prevalence to be 6/1000 [21]. Stroke is a leading cause of mortality and disability. Despite the fact that stroke mortality is falling, the number of people living with the effects of a stroke has increased as the population grows and ages.

Although stroke mortality is decreasing, the prevalence of people living with the effects of stroke has increased because of the growing and ageing population [5]. The global burden of stroke has risen significantly in recent decades as a result of rising population numbers and aging, as well as an increase in the incidence of modifiable stroke risk factors, particularly in low- and middle-income nations. The number of patients who will need care by clinicians with expertise in neurological conditions will continue to grow in the coming decades [22]. For this reason, it is very important to determine the factors associated with the disease and to ensure the preventability of the disease in terms of reducing the economic burden.

In the literature, there are some studies on unbalanced datasets and various suggested methods. In a study, it has been suggested how the most appropriate learning method should be for unbalanced datasets [23]. In this study, it was emphasized that classical classifier methods naturally tended towards class types with high distribution, and to overcome this problem, known methods were combined and used.

A comprehensive research in this area is included in an article named "Learning from Imbalanced Data" [24]. In this study, the vital areas of unbalanced data were examined, it was determined how knowledge discovery should be made in such cases, and suggestions were presented about the stages of the learning process. In addition, the performance analysis methods are included.

In another study related to the subject, the Borderline-SMOTE method was proposed. In the known SMOTE (Synthetic Minority Oversampling Technique) method, the class type with unbalanced data distribution is artificially replicated and balancing is achieved in this way. In the BorderlineSMOTE method, it belongs to the minority class type. The samples located on the boundary lines of the clusters formed by the samples were multiplexed with SMOTE. In their experimental application, it was stated that the Borderline-SMOTE method provided higher success [25].

In the dataset used in this study, there are 249 stroke patients and 4861 non-stroke patients, and the dataset is unbalanced. For this reason, in the study, the results of the classification made with the initial state of the dataset using SMOTE, which is the data balancing method, and the results obtained from the balanced data were compared.

Specificity, sensitivity, accuracy, positive predictive value and negative predictive values were obtained as 0.0887, 0.9772, 0.9339, 0.9544 and 0.1679, respectively, according to the modeling result using the gardient boosting tree method using the original version of the dataset. Specificity, sensitivity, accuracy, positive predictive value and negative predictive values were obtained as 0.0887, 0.9772, 0.9339, 0.9544 and 0.1679, respectively, according to the modeling result using the gardient boosting tree method using the SMOTE applied version of the dataset.

In unbalanced datasets, classifiers tend to focus on the class set where the number of samples is in the majority during training. In other words, a bias occurs in the system and in the results, as the classification will be trained with the class label with a large number of samples. The dataset used in the study is unbalanced, and when the results obtained in the classification made with its original form are examined, it is known that the negative predictive value and specificity values are low and these values are unacceptable in a study. However, when the modeling results made with the SMOTE applied dataset were examined, it was obtained that the results were more consistent and realistic. For this reason, it is recommended that researchers apply dataset balancing methods in order to obtain more accurate results in case they encounter an unbalanced dataset problem.

Conflict of interests
The authors declare that they have no competing interests.

Financial Disclosure
All authors declare no financial support.

References

1. Baroni AFFB, Fábio SRC, Dantas RO. Risk factors for swallowing dysfunction in stroke patients. Arq Gastroenterol. 2012;49:118-24.
2. Hatano S. Experience from a multicentre stroke register: a preliminary report, Bull World Health Organ. 1976;54:541.
3. Sarikaya H, Ferro J, Arnold M. Stroke prevention-medical and lifestyle measures. Eur Neurol. 2015;73:150-7.
4. Feigin VL, Nichols E, Alam T, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019;18:459-80.
5. Stinear CM, Lang CE, Zeiler S, et al. Advances and challenges in stroke rehabilitation. Lancet Neurol. 2020;19:348-60.
6. Adams RD, Victor M, Ropper AH, et al. Principles of neurology, ed: LWW, 1997.
7. Tasci ME, Samli R. Diagnosis of Heart Disease with Data Mining. Eur J Eng Sci Tech. 2020;88-95.
8. Dogan A, Birant D. Machine learning and data mining in manufacturing. Expert Syst Appl. 2020;114060.
9. Ozmen O, Ahmad K, Engin A. Performance Comparison of Classifiers on Heart Disease Data. Fırat University J Engineering Sciences. 2018;30:153-9.
10. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57.
11. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.
12. Maldonado S, López J, Vairetti C. An alternative SMOTE oversampling strategy for high-dimensional datasets. Appl. Soft Comput. 2019;76:380-9.
13. Agrawal A, Viktor HL, Paquet E. SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling, in 2015 7Th international joint conference on knowledge discovery, knowledge engineering and knowledge management (IC3k). 2015;226-34.
14. Manju B, Nair AR. Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms, in 2019 ISED. 2019;1-7.
15. Wang Z, Wu C, Zheng K, et al. SMOTETomek-based resampling for personality recognition. IEEE Access. 2019;7:129678-89.

16. Li T, Levina E, Zhu J. Network cross-validation by edge sampling. Biometrika. 2020;107:257-6.

17. Guelman L. Gradient boosting trees for auto insurance loss cost modeling and prediction. Expert Syst Appl. 2012;39:3659-67.

18. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;1189-232.

19. Ahmed T, Bosu A, Iqbal A, et al. SentiCR: a customized sentiment analysis tool for code review interactions, in 2017 32nd IEEE/ACM International Conference on ASE. 2017;106-11.

20. Medin J, Nordlund A, Ekberg K. Increasing stroke incidence in Sweden between 1989 and 2000 among persons aged 30 to 65 years: evidence from the Swedish Hospital Discharge Register. Stroke. 2004;35:1047-51.

21. Altun Y, Aydin I, and Algin A. Demographic characteristics of stroke types in Adiyaman. Turkish J Norol. 2018;24:6.

22. Katan M, Luft A. Global burden of stroke, in Seminars in neurology. 2018;208-11.

23. Wu G, Chang EY. Class-boundary alignment for imbalanced dataset learning, in ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC. 2003;49-56.

24. He H, Garcia EA. Learning from imbalanced data, IEEE Trans Knowl Data Eng. 2009;1263-84.

25. Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in International conference on intelligent computing. 2005;878-87.