Unbalanced Data Classification Algorithm Based on Hybrid Sampling and Ensemble Learning

Mengfei Wu
Computer Department
Xi'an University of Posts and Telecommunications
Xi'an, China
15249293184@163.com

Ximing Li
Computer Department
Xi'an University of Posts and Telecommunications
Xi'an, China
lxm623867171@163.com

2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) | 978-1-6654-0553-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISKE54062.2021.9755369

***Abstract*—Due to the inherent characteristics of unbalanced data sets, the classification results are often affected by a large number of categories, resulting in the problem of low prediction accuracy of a few classes in the classifier. This paper proposes an unbalanced data classification algorithm based on hybrid sampling and ensemble learning—ESBagging. Firstly, the undersampling algorithm ENN is used to eliminate the samples affecting the classification decision boundary, and then the SMOTE algorithm is used to further mitigate the reduced data sets to make the data sets reach the balanced state. Finally, the Bagging ensemble learning framework is used to extract sub-data sets from the data sets, and then multiple basic classifiers are obtained based on these sub-data sets, Then, a relatively better prediction model is obtained by combining the classification results of the basic classifier. Taking 10 groups of KEEL imbalanced data set as experimental data, ESBagging, BalancedBagging, and SMOTEBagging were compared. The experimental results show that ESBagging obtains the highest evaluation index in 5 groups, 7 groups, and 7 groups of data sets under F1-measure, G-mean, and AUC criteria respectively. It can be seen that the proposed algorithm has better classification performance on unbalanced data.**

Keywords—Unbalanced data, Classification, Undersampling, Oversampling, Ensemble Learning

I. INTRODUCTION

An unbalanced data set [1] refers to a data set with a large number of samples in some categories and a small number of samples in some categories, which results in the unbalance of all types of samples in the data set. Generally, the category with a small number of samples is called the minority category and the category with a large number of samples is called the majority category. However, using unbalanced data sets to train traditional classifiers will lead to low prediction accuracy for a few classes, so unbalanced data learning has always been a hot research topic in the field of machine learning.Therefore, unbalanced data learning has always been a research hotspot in the field of machine learning. There are a large number of unbalanced data classification problems in real life, such as medical detection[2], healthcare insurance Fraud Detection [3], intrusion detection [4], industrial fault detection [5] and so on. For the current imbalanced data classification problem, the usual solutions are mainly divided into the data preprocessing level and the classification algorithm level [6]. The main idea of the method at the data preprocessing level is to achieve a relative balance between the number of samples in each category in the data set through resampling technology, mainly under-sampling for the majority category [7-8] and oversampling for the minority category [9-10] , And the hybrid sampling combining the two sampling methods [11]. At the algorithm level, related researchers have improved the algorithm to increase the importance of the classifier to the minority classes. The more representative algorithms are cost-sensitive [12], single-class learning [13-14] and ensemble learning [15], etc. . This paper focuses on the research progress of imbalanced data classification algorithms that combine sampling and ensemble learning. The proposed model framework, as shown in Fig.1, first divides the KEEL unbalanced data set into a training set and a test set, and most classes in the test set. Using the ENN(edited nearest neighbor) [16] algorithm, we remove the edge junctions of the categories or the minority clusters, and we will delete them. Subsequently, through SMOTE (Synthetic Minority Oversampling Technique) [17] synthesis of minority samples to achieve a balanced data set. Perform N random sampling from the balanced data set, collect 10 times in total, get 10 sample sets, construct multiple base learners, train multiple base learners at the same time, and finally combine multiple base learners to obtain a strong learner , And then predict and classify the test set.

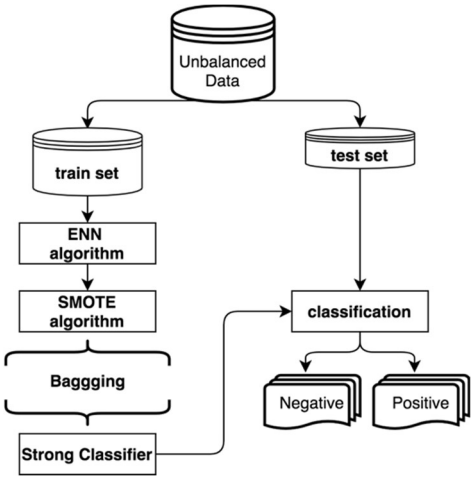

Fig.1. frame diagram of ESBagging algorithm

II. OVERVIEW OF IMBALANCED DATA CLASSIFICATION PROBLEMS

The problem of unbalanced data is mainly due to the unbalanced distribution of various samples in the data set. Due to the particularity of the classification problem of unbalanced data set, the classification operation of the unbalanced data set is often susceptible to the "majority rule" principle when using the general classification algorithm. Obviously, in order to improve the overall classification accuracy, the classifier will naturally ignore the influence of a few classes on classification and classify them into a majority class. According to this classification result, a higher classification performance can be obtained, but it cannot bring relatively high practical value.

A. Unbalanced data set classification process

The classification process of machine learning mainly includes four parts: obtaining the original data set, data preprocessing, classification model construction, and model evaluation. Unbalanced data sets often follow the following classification process.

a) Obtaining the original data set: The data set is an indispensable element of the machine learning algorithm. Obtaining the data set is also the first step in machine learning research. The more classic unbalanced test data set is the KEEL data set.

b) Data preprocessing: The original data may be messy and complicated. Directly using such a data set for modeling training will often bring extremely high training costs to the classifier, but it will not get a good classification effect. Usually after analyzing the unbalanced data set, some preprocessing operations are performed to reduce the worries of the training model.

c) Classification model construction: This is the process of learning from data to build a classification model (classifier), and then predicting the output of new inputs. Building a classification model is never a once-and-for-all thing. It is necessary to construct a suitable classification model based on the inherent characteristics of the imbalanced data set.

d) Model evaluation: Judge the classification effect of a classifier model through a series of evaluation indicators.

B. Difficulties in the classification of unbalanced data sets

Classification is the basic problem of machine learning research. Even if faced with a set of balanced data sets, the classification problem itself does not have a relatively complete set of processing algorithms. Unbalanced data sets have their inherent complexity and particularity, which makes this There are still many difficulties to be solved in the field of research. The main reasons for the difficulty in classification are as follows.

a) Difficulty in data sampling: classification problems often bring large amounts of data, but the proportion of minority samples in unbalanced data sets is often far less than one percent of the overall sample. Although a series of sampling strategies will be adopted to balance the data, the existing sampling methods generally have defects such as over-fitting, loss of most types of sample information, and increased redundant information.

b) Algorithm selection difficulty: Although common and mature classification algorithms such as decision trees, random forests, support vector machines, etc. have made considerable progress, they have a low recognition rate for minority classes in imbalanced data and cannot adapt to imbalances well. The characteristics of the data set.

c) Difficulty in data recognition: Noise is usually an inevitable factor in the data set. Noise in minority samples will undoubtedly reduce the ability of the classifier to recognize minority classes. Especially when the number of noisy data is equal to or greater than the number of samples of the minority class, there may be a risk that the classifier learns both the noise and the minority class at the same time. Therefore, it is particularly important to remove as much noise in the unbalanced data set as possible, and it is also closely related to the smooth classification of subsequent classifiers.

d) Difficulty in performance evaluation: Performance evaluation has important judging value for measuring the pros and cons of a classifier, and it also provides an indispensable reference value for choosing a suitable classifier. Evaluation indicators based on accuracy and error rate will in order to pursue a higher overall accuracy rate and a lower error rate, at the expense of the minority class, the unknown sample will be tilted to the majority class, ignoring the classification accuracy of the minority class sample. Can not reflect the quality of the model well.

III. UNBALANCED DATA CLASSIFICATION ALGORITHM BASED ON HYBRID SAMPLING AND ENSEMBLE LEARNING

A. Hybrid sampling technique

The data reconstruction strategy is a process of preprocessing the original data distribution at the data level independent of the classification algorithm. It aims to convert the unbalanced data set into a more balanced data set, and then use the classification method of the balanced data to learn classification and Performance evaluation.

1) Undersampling

The simplest undersampling strategy is Random Under-Sampling (RUS), that is, some samples are randomly selected from the majority of samples to be eliminated. Because the random under-sampling method does not consider the distribution of the sample, the sampling has great randomness and may delete important majority-type sample information. In response to the above shortcomings, Wilson et al.[16]. proposed an edited nearest neighbor (ENN) under-sampling algorithm. The rule is that if two or more sample categories of the three nearest neighbor samples of the sample are different from it, then Delete this sample. The ENN method is an under-sampling algorithm. First, it searches the 3 nearest neighbor samples of the majority class of samples. If two or more of the 3 nearest neighbor samples of the sample are not the same as the sample type, the sample will be deleted. This algorithm means most samples are deleted.

2) Oversampling

The oversampling technique corresponding to under_ sampling adopts a strategy of increasing the number of minority classes in an unbalanced data set, and synthesizes new minority class samples through a series of methods and adds them to the original data set to balance the data set. Random Over-Sampling (ROS) is also the simplest oversampling strategy. Randomly copy samples from categories with few samples, and then add the sampled samples to the data set. The realization of the oversampling strategy is simple. However, using such a simple random copy method to increase the minority samples will easily cause overfitting and make the model not good generalization ability. The addition of new synthetic samples will also increase the sample training time. In order to reduce the possibility of over-fitting of the classification algorithm, the SMOTE over-sampling technique proposed by Chawla N V et al[18]. is a powerful method. SMOTE oversampling technology is different from the traditional oversampling method of simple sample copying. It uses minority samples to control the generation and distribution of artificial samples to achieve the purpose of data set balance, and this method

can effectively solve the problem of small decision-making intervals. The classification of the over-fitting problem.

The SMOTE algorithm flow is as follows:

a) For each individual x in the minority sample, calculate its distance to all samples in the minority sample set C_1 using Euclidean distance as the standard, and sort all the distances. The k nearest samples to it are its k nearest neighbors.

b) A sampling ratio is set according to the sample imbalance ratio to determine the sampling magnification N. For each minority sample x, a sample y_i is randomly selected from its k nearest neighbors.

c) For each randomly selected neighbor y_i, construct a new sample with the original sample according to the following formula (1).

$$X_{new,i} = x + rand(0,1) \times |x - y_i| \quad (1)$$

B. Ensemble learning classification model

Integrated learning is a learning mode that integrates multiple weak learning machines to obtain better results than a single learning machine. In classification problems, regression problems, outlier detection and other issues, the integrated learning algorithm performs well, and has better accuracy and stability than traditional machine learning algorithms.

1) Bagging

Bagging algorithm is an integrated learning framework, mainly to use self-service and aggregation. Self-service is based on the data of the original sample, sampling the original sample with replacement data, and the number of sampling ensures that the number of new samples is the same as the original sample number and the scale is the same. Some samples may never be selected, and some samples may be selected many times. The base learning algorithm learns from the obtained data set to obtain the base learner. The advantage of this is that we only need to spend a few computing resources to get a more accurate estimate. Aggregation refers to the aggregation of base classifications. Bagging can handle both binary classification and polyphenols in classification problems. The results of aggregation are generally carried out by voting. After the sub-base classifier learns the samples, Bagging collects the output of all the classifiers and votes to reject the tags, and the one with the most votes is used as the final prediction result of the template. When the number of votes is the same, a tag is randomly selected.

2) CART decision tree

Decision trees have the advantages of strong interpretability and fast running speed. Decision trees with smaller depth are often used as basic classifiers in ensemble learning [20]. The ensemble learning model constructed in this article uses the CART algorithm to generate basic classifiers.

CART is a common decision tree generation algorithm, which uses Gini coefficient as an index to evaluate the optimal partition characteristics. If the Gini value of the sample set is smaller, the probability that the samples in the data set belong to the same category is higher.

For sample set D, the Gini coefficient calculation formula (2) is as follows:

$$Gini(D) = 1 - \sum_{k=1}^{k} (|C_k|/|D|)^2 \quad (2)$$

Among them: K is the number of categories in the sample set D, and C_k is the number of samples in the k category.

The calculation formula for the Gini coefficient of sample set D on feature A is as follows (3):

$$Gini(D,A) = \frac{|D_1|}{D} Gini(D_1) + \frac{|D_2|}{D} Gini(D_2) \quad (3)$$

Among them: D1 and D2 are two subsets after dividing the data set with a certain characteristic value on the characteristic value A.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In classification tasks with unbalanced data, accuracy, recall, and F1-measure values are usually used as model performance metrics. The confusion matrix for the two-class classification problem is shown in Table 1. Among them: TP (True Positive) is the case when the positive sample is classified correctly; FP (False Positive) is the case when the negative sample is classified incorrectly; FN (False Negative) is the case when the positive sample is classified incorrectly; TN (True Negative) Is the case where the classification of the counter-example sample is correct.

TABLEI. Confusion matrix

The true situation	**forecast result**	
	positive	**negative**
positive	TP	FP
negative	TN	FN

A. Evaluation index

Obviously, recall rate(refers to the proportion of the number of correctly classified positive classes to the number of all positive classes) and precision rate are(The proportion of the number of correctly classified positive classes to the number of predicted positive classes) sometimes a pair of contradictory indicators, that is, there is no guarantee that a higher recall rate will also have a higher precision rate. Due to the complexity of classification of imbalanced data sets, it is difficult to use only recall rate or precision rate such that a single indicator can more accurately evaluate the performance of the classifier. In order to comprehensively reflect the classification performance of imbalanced data sets, F1-measure, G-mean, AUC, etc. are used as evaluation indicators.

F1-measure[21] is also called F-Score, and its calculation formula is shown in equatio(4), and α is a proportional coefficient that is often taken as 1. F-measure can balance precision and recall rate and find the best combination of them.

$$\text{F1} - \text{measure} = (\alpha^2 + 1)Recall \times Prision/ (\alpha^2 Recall + Prision) \quad (4)$$

G-mean is also a comprehensive evaluation index, involving two single evaluation indexes of sensitivity and specificity, Sensitive=TP/(TP+FN), which measures the ability of the classifier to recognize positive classes; Specificity = TN/(TN + FP), which measures the classifier's

ability to recognize negative classes. Its expression is as formula (5):

$$G - mean = \sqrt{Sensitive \times Specificity} \quad (5)$$

Although F-measure and G-mean have improved and perfected the accuracy and error rate, they still cannot achieve a good evaluation effect when comparing the performance between the classifier and various distributions. The appearance of ROC [21] curve appropriately solves the problem that it is difficult to compare the performance of different classifiers on different sample distribution ranges.

The ROC curve is called the receiver operating characteristic curve (receiver operating characteristic curve), which takes the false positive rate (FP_rate) and the true rate (TP_rate) as the axis, weighs the correlation between the benefits of correct classification and the cost of misclassification, and visualizes The way is displayed intuitively. The area under the ROC curve is called AUC (Area Under Curve) [21]. AUC is used to quantitatively evaluate the accuracy of the classifier's prediction. The closer the curve is to the upper left corner, the higher the value, that is, the larger the area under the curve, the more accurate the prediction high.

AUC is not affected by the type of classifier and the prior probability, so it is widely recognized in the classification performance evaluation index of imbalanced data set. It is undeniable that the ROC curve provides a powerful visualization method for the classification and evaluation of class imbalanced data.

B. Data set description

In order to measure the performance of the ESBagging algorithm proposed in this paper, 10 data sets of the standard database in KEEL are used to train the classifier and analyze the experimental results. The Imbalance Ratio (IR) of the experimental data set ranges from 1.8 to 14.3. The data are all two-category data sets. The experimental data set information is shown in Table 2.

TABLEII. Experimental datasets description

Data set	Total number of samples	Attributes	Majority sample number	Number of minority samples	IR
Pima	768	8	500	268	1.87
glass0	214	9	144	70	2.06
yeast1	1484	8	1055	429	2.46
haberman	306	3	225	81	2.78
vehicle1	846	18	629	217	2.9
ecolo1	336	7	259	77	3.36
new-thyroid1	215	5	180	35	5.14
segment0	2308	19	1979	329	6.02
yeast3	1489	8	1321	163	8.1
yeast-l_vs_7	459	7	429	30	14.3

C. Experimental design and analysis

TABLEIII. Experimental result

Dataset	Bagging			SMOTEBagging			ESBagging		
	F1	G-mean	AUC	F1	G-mean	AUC	F1	G-mean	AUC
pima	0.6947	0.7770	0.8429	0.6800	0.7713	0.8433	**0.7143**	**0.8760**	**0.8584**
glass0	0.7059	0.7921	0.9592	**0.9000**	**0.9701**	**0.9902**	0.7200	0.8911	0.9150
yeast1	0.5541	0.6530	0.7846	**0.6171**	0.7212	0.7891	0.6122	**0.7271**	**0.8152**
haberman	0.3571	0.5259	0.5738	**0.5455**	0.6588	0.7291	0.4598	**0.7146**	**0.7484**
vehicle1	0.4598	0.5794	0.7797	0.4000	0.5395	0.7539	**0.6129**	**0.7213**	**0.9377**
ecolo1	**0.7368**	**0.8592**	0.9190	0.6364	0.8359	0.9040	0.5714	0.7739	**0.9377**
new-thyroid1	0.9524	0.9535	0.9920	0.9524	0.9535	0.9936	**0.9524**	**0.9535**	**0.9941**
segment0	**0.9710**	0.9771	0.9751	0.9640	0.9758	0.9757	0.7937	**0.9812**	**0.9771**
yeast3	0.7407	0.7987	0.7197	0.7458	0.8329	0.9385	**0.7937**	**0.8867**	**0.9449**
yeast-l_vs_7	**0.5714**	0.6325	0.7333	0.3636	0.6177	**0.7966**	0.1429	0.4262	0.7230

In this experiment, all algorithms adopt the five-fold cross verification method, and the comparison algorithms in the experiment are Bagging algorithm and SMOTEBagging algorithm. The ensemble algorithm adopts CART as the base classifier, the depth of all decision trees is 8, and the number of base classifiers of the algorithm is 10.

The experiment analyzed the F1-measure value, G-mean value and AUC value of the three algorithms. Table 3 lists the evaluation index values of this algorithm and the other two comparison algorithms in each data set, among which the bolded value is the highest evaluation index value.

The F1-measure value, G-mean value and AUC value can better measure the performance of the unbalanced data classification algorithm. It can be seen from Table 3 that ESBagging algorithm has obvious advantages in most data sets compared with other algorithms. However, in the yeast-1_vs_7 data set, the F1-measure value of the algorithm in this paper differs greatly from that of Bagging algorithm, which is due to the fact that in highly unbalanced data, Minority class only a few of the total number of samples, sample after the sample for most class undersampling, may delete many potential most valuable kind of sample, can be used to appropriately increase the majority class at this time of sampling conditions for clustering analysis, discrete values can be out, as far as possible keep sample of important value.

However, compared with SMOTEBagging algorithm using oversampling, the AUC value of the algorithm in this paper is significantly improved on the Haberman data set. This is because the algorithm in this paper obtains samples with high weight after undersampling, and then the oversampling again increases the weight of a few classes and improves the ratio column discriminating as most classes.

The proportion of the discriminant to a small number of classes is reduced, so that the influence of these samples on the base classifier is greater.

In order to make a more intuitive comparison of the three algorithms, Fig.2, Fig.3, and Fig.4 show the experimental results of the comparison algorithm and ESBagging algorithm on 10 data sets. It can be seen that the algorithm in this paper has certain advantages in processing unbalanced data.

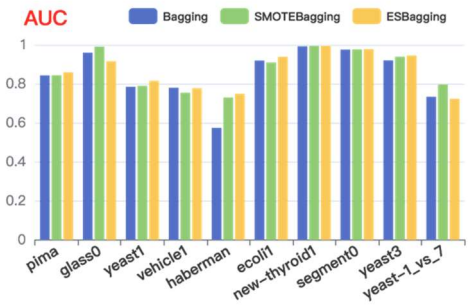

Fig.2. AUC comparison of three algorithms

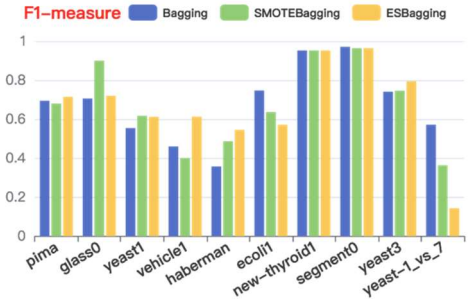

Fig.3. F1-measure comparison of three algorithms

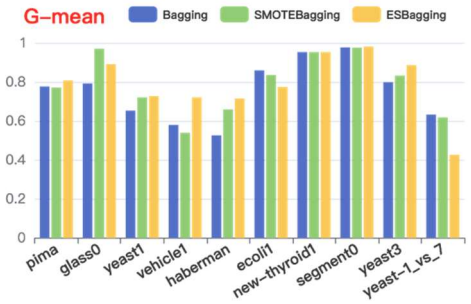

Fig.4. G-mean comparison of three algorithms

V. CONCLUSION

This article through to the unbalanced data classification study on the problems existing in the traditional Bagging algorithm, is proposed based on hybrid sampling and ensemble learning study combined with the unbalanced data classification method, the mixed sample stage as far as possible, reduce the amount of most classes have too much influence classifier performance problems, then carried out sampling to reduce the number of samples. To prevent the result misclassification caused by sampling problems in the classifier, the idea of ensemble learning is adopted to construct multiple base classifiers. According to the results of the base classifiers, voting method is adopted to ensure the robustness of the results and reduce the impact of the special case data set obtained by sampling.

However, the algorithm in this paper inherits the shortcoming that important value samples may be lost under sampling. How to ensure the prediction accuracy during the training process while reducing the impact of the lost sample data on the model, and how to find the important decision samples will be the future focus of research.

REFERENCES

[1] Liang, X. W. et al. “LR-SMOTE - An improved unbalanced data set oversampling based on K-means and SVM.” Knowl. Based Syst. 196 (2020): 105845.

[2] Xu, Zhaozhao et al. “A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data.” Inf. Sci. 572 (2021): 574-589.

[3] Mary, A. and S. P. A. Claret. “Imbalanced Classification Problems: Systematic Study and Challenges in Healthcare Insurance Fraud Detection.” 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (2021): 1049-1055.

[4] Zhang, Hongpo et al. “An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset.” Comput. Networks 177 (2020): 107315.

[5] Lee, Jeongsu et al. “Fault detection based on one-class deep learning for manufacturing applications limited to an imbalanced database.” Journal of Manufacturing Systems 57 (2020): 357-366.

[6] Yan-Xia L I , Chai Y , You-Qiang H U , et al. Review of imbalanced data classification methods[J]. Control and Decision, 2019.

[7] Lee, Yoon Sang, and Chulhwan Chris Bang. "Framework for the Classification of Imbalanced Structured Data Using Under-sampling and Convolutional Neural Network." Information Systems Frontiers (2021): 1-15.

[8] Yuanyuan W U , Shen L . Imbalanced fuzzy multiclass support vector machine algorithm based on class-overlap degree undersampling[J]. Journal of University of Chinese Academy of Sciences, 2018.

[9] Pradipta, G. et al. “Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data.” IEEE Access 9 (2021): 74763-74777.

[10] Xu, Tingting et al. “A predictive model of recreational water quality based on adaptive synthetic sampling algorithms and machine learning.” Water research 177 (2020): 115788 .

[11] Gao, Xin et al. “An ensemble imbalanced classification method based on model dynamic selection driven by data partition hybrid sampling.” Expert Syst. Appl. 160 (2020): 113660.

[12] Bejaoui, Amine et al. “Cost-sensitive design of quadratic discriminant analysis for imbalanced data.” Pattern Recognition Letters 149 (2021): 24-29.

[13] Asimit, Vali et al. “Robust Classification via Support Vector Machines.” ArXiv abs/2104.13458 (2021): n. pag.

[14] Hayashi, T., Fujita, H. “One-class ensemble classifier for data imbalance problems”. Appl Intell (2021).

[15] Wang, Xinyue et al. “Local distribution-based adaptive minority oversampling for imbalanced data classification.” Neurocomputing 422 (2021): 200-213.

[16] Tyagi, S. and S. Mittal. “Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning.” (2019).

[17] Diallo, Moussa et al. “Synthetic minority oversampling technique in stages for unbalanced climate and rice dataset: the Office Du Niger case study.” Proceedings of the 3rd International Conference on Telecommunications and Communication Engineering (2019): n. pag.

[18] Costa R , Baker J W . SMOTE-LASSO Model of Business Recovery Over Time - Case Study of the 2011 Tohoku Earthquake. 2021.

[19] Breiman L I , Friedman J H , Olshen R A , et al. Classification and Regression Trees. Wadsworth.[J]. Biometrics, 1984, 40(3):358.

[20] Zhang Z L , Dang-Ping W U . An imbalanced data classification method based on probability threshold Bagging[J]. Computer Engineering & Science, 2019.

[21] Bhamare, V. et al. “Area Under Curve Method Development for Etodolac in Bulk and Tablet dosage form.” (2021).