BioData Mining

RESEARCH

Open Access

Processing imbalanced medical data at the data level with assisted-reproduction data as an example

Junliang Zhu[1], Shaowei Pu[1], Jiaji He[1], Dongchao Su[1], Weijie Cai[1], Xueying Xu[1] and Hongbo Liu[1,2*]

*Correspondence:
Hongbo Liu
hbliu@cmu.edu.cn
[1]Department of Health Statistics, School of Public Health, China Medical University, Shenyang 110122, PR China
[2]Key Lab of Environmental Stress and Chronic Disease Control & Prevention, China Medical University, No.77 Puhe Road, Shenyang North New Area, Shenyang 110122, Liaoning Province, PR China

Abstract

Objective Data imbalance is a pervasive issue in medical data mining, often leading to biased and unreliable predictive models. This study aims to address the urgent need for effective strategies to mitigate the impact of data imbalance on classification models. We focus on quantifying the effects of different imbalance degrees and sample sizes on model performance, identifying optimal cut-off values, and evaluating the efficacy of various methods to enhance model accuracy in highly imbalanced and small sample size scenarios.

Methods We collected medical records of patients receiving assisted reproductive treatment in a reproductive medicine center. Random forest was used to screen the key variables for the prediction target. Various datasets with different imbalance degrees and sample sizes were constructed to compare the classification performance of logistic regression models. Metrics such as AUC, G-mean, F1-Score, Accuracy, Recall, and Precision were used for evaluation. Four imbalance treatment methods (SMOTE, ADASYN, OSS, and CNN) were applied to datasets with low positive rates and small sample sizes to assess their effectiveness.

Results The logistic model's performance was low when the positive rate was below 10% but stabilized beyond this threshold. Similarly, sample sizes below 1200 yielded poor results, with improvement seen above this threshold. For robustness, the optimal cut-offs for positive rate and sample size were identified as 15% and 1500, respectively. SMOTE and ADASYN oversampling significantly improved classification performance in datasets with low positive rates and small sample sizes.

Conclusions The study identifies a positive rate of 15% and a sample size of 1500 as optimal cut-offs for stable logistic model performance. For datasets with low positive rates and small sample sizes, SMOTE and ADASYN are recommended to improve balance and model accuracy.

Keywords Imbalanced data, Logistic model, Imbalanced degree, Sample size, Imbalanced data processing method

Background

With the development of information technology and medical technology, mining big data for valuable medical information has become a research hotspot in the fields of statistics, machine learning, and artificial intelligence. In current research on medical big data, predicting and classifying diseases have always been a focus [1, 2]. When establishing prediction models, a prerequisite for using existing classification models is that the training sample set is balanced or close to balanced [3, 4]. However, real medical datasets often fail to meet this prerequisite [5], especially in studies of rare cases. In the majority of cases, due to sampling difficulties and sample size constraints in clinical practice, the number of positive samples in medical datasets is far smaller than the number of negative samples, leading to imbalanced data. This imbalance is a particularly common problem when diagnosing and predicting malignant tumors [6], heart disease [7], cardiovascular disease [8], and pregnancy diseases [9].

Imbalanced classification data occur when one class (the majority class) has more instances than another class (the minority class) [10, 11]. When constructing medical prediction models, the prediction performance of the minority-class samples is crucial for determining the model's quality [12]. However, with real medical data, these positive samples are often masked by a large number of negative samples that interfere with the model's prediction effect on positive samples. For example, in a diagnostic dataset of 1000 cancer patients, only 10 samples might be diagnosed as having cancer, whereas the others are diagnosed as healthy. In this case, if the classifier predicted that all the samples were healthy, it would obtain a prediction accuracy of 99%. However such a model would not help with the goal of identifying cancer patients.

Currently, traditional classification models such as logistic regression and discriminant analysis are widely used in medical research, due to their convenience, strong variable interpretability, high prediction accuracy, and generalization [13, 14]. However the logistic model can only function effectively if the distribution of response variables in the dataset is balanced. If the probability of occurrence of an event is less than 5%, it is difficult to establish a good prediction model because there is less information about rare events [15]. Therefore, the logistic model is not suitable for the problem of imbalanced data classification, especially in datasets where the minority class is extremely small or when the data distribution is heavily skewed. In such cases, alternative predictive models, such as decision trees, support vector machines, or random forests, may provide better performance because they are more robust to data imbalance. However, in some existing studies, these problems are not considered, and logistic models are directly used for analysis due to their simplicity and interpretability. As such, their conclusions may be questioned. It is crucial to explore and discuss its performance on imbalanced medical datasets, even when other predictive models might offer better accuracy in certain situations.

At present, methods to address the above-mentioned imbalanced data problem can be applied at the algorithmic level [16, 17] or the data level [18, 19]. At the algorithmic level, imbalanced data are mainly dealt with by modifying existing algorithms or proposing new classification algorithms, often using cost-sensitive learning [20] or ensemble learning [21]. Cost-sensitive learning makes the classifier learn imbalanced data better by increasing the cost of misclassifying a few class samples. Ensemble learning integrates the results of training multiple learners according to certain standards to improve the

generalization ability of the learner. However, compared with the traditional method, models established by an algorithm-level method have higher complexity, and the model results lack intuitive interpretation. Therefore, processing at the data level may be more conducive to the analysis of imbalanced medical data.

Data level methods involve modifying the original dataset through preprocessing techniques, which include optimizing the feature space using feature selection methods [22, 23] and optimizing the sample space using resampling techniques [24, 25]. Feature selection techniques generally fall into three categories: filters, wrappers, and embedded methods [26]. These methods focus on resolving the implicit complexity of the data by finding a feature space that better represents the minority class, addressing the issue where the original space may inadequately characterize it [27, 28]. However, feature space optimization can be challenging. In non-high-dimensional imbalanced datasets, feature selection often needs to be combined with resampling and algorithmic methods to achieve better results. Resampling adjusts the dataset's imbalance to balance the two classes, making it more suitable for traditional classification methods. Studies have shown that, for several common standard classification models, the training effect of using a balanced dataset is better than the original imbalanced dataset [29]. However, some studies have also shown that some classification models trained by the original imbalanced dataset are comparable to those trained by the same resampled balanced dataset [30, 31]. The biggest controversy about resampling technology to solve the problem of imbalanced classification is that resampling changes the distribution of the sample data. According to statistical knowledge, only randomly selected samples can be used to estimate the distribution of a population [32]. Although resampling techniques cannot simulate the true distribution of original data, classification models can obtain more useful information from the balanced data than the original data [33]. According to the balanced distribution method, resampling techniques can be categorized into undersampling, oversampling, and hybrid sampling. Numerous scholars have conducted comparative studies on the performance of undersampling and oversampling techniques, but a widely accepted conclusion has yet to be reached. Some researchers have pointed out that the advantages of proposed undersampling methods become more pronounced when the dataset reaches the PB scale [34]. Conversely, other studies have suggested that oversampling outperforms undersampling, particularly for datasets with a very small number of minority-class samples [35]. Even for complex datasets, oversampling has been shown to significantly enhance classifier performance [36].

Traditional classification models and imbalanced data processing methods both have advantages for the problem of imbalanced data. This study explores the trends in classification performance indices of the models and the differences in methods of treating imbalanced data under varying degrees of imbalance and different sample sizes. Compared with simulated data, real data better reflects data distributions in real environments. Therefore, we chose assisted-reproduction data as an example to construct datasets in different situations. We constructed datasets with different imbalance degrees and sample sizes. Then, we compared the classification accuracy of the models under different circumstances and determined the optimal cutoff values for the imbalance degree and sample size. We constructed various datasets with a high degree of imbalance and small sample sizes. These datasets were processed by four processing methods for imbalanced data: Synthetic Minority Over-Sampling Technique (SMOTE)

oversampling, Adaptive Synthetic Sampling (ADASYN) oversampling, One-Sided Selection (OSS) undersampling, and Condensed Nearest Neighbor (CNN) undersampling. The effects of these different processing methods were compared. This study explores processing strategies for imbalanced data and provides insights into the selection of appropriate imbalanced data processing methods.

Materials and methods

Data source

We retrospectively collected medical records of patients who received assisted reproductive treatments from January 2015 to December 2020 at the Reproductive Medical Center of Jiangxi Maternal and Child Health Hospital in Nanchang City, Jiangxi Province, China. The dataset comprised 17,860 samples and 45 variables, covering the following seven aspects: basic information, infertility factors and comorbidities, previous treatment and maternal history, pre-pregnancy basic biochemical indicators, basic semen-quality indicators, biochemical indicators during pregnancy, and transfer information.

The outcome variable was whether cumulative live births occurred. A cumulative live birth is defined as the first live birth in a complete treatment cycle. A complete treatment cycle is defined as the transfer of all eligible embryos at one time after one ovulation induction until termination after a live delivery. If no live birth occurred, the treatment cycle was considered to have failed.

The ethical considerations for biomedical research involving humans in this project meet the requirements of the Declaration of Helsinki and the Measures for Ethical Review of Life Science and Medical Research Involving Humans. The Medical Ethics Review Committee of Jiangxi Provincial Maternal and Child Health Hospital approved the implementation of this project according to the research plan (SZYX-202305).

Data preprocessing and variable filtering

Data preprocessing

First, non-characteristic variables in the dataset, such as case numbers and admission dates were removed. Then, duplicate rows in the dataset were removed and merged, missing samples were removed, and outliers with statistical errors in the data were replaced by the mode. Finally, discrete variables in the dataset were numerically encoded.

Variable screening

To avoid problems such as over-dimensionality and model overfitting with the logistic model, the random forests algorithm was used to evaluate the importance of variables in the preprocessed dataset. There were two main evaluation indicators: mean decrease accuracy (MDA) and mean decrease Gini (MDG). MDA represents the degree of decline in the accuracy of random forest prediction, where higher values indicate greater importance of the variable. MDG calculates the effect of each variable on the heterogeneity of observations at each node of the classification tree to compare the importance of the variable, where higher values again indicate greater importance of the variable. In this study, MDA indicators were primarily used to evaluate the importance of variables.

Construction of different datasets

Construction of datasets with different imbalanced degrees

By randomly sampling the original dataset, datasets with different imbalance degrees were constructed, where the ratios of balanced to imbalanced samples were 99:1, 97:3, 95:5, 90:10, 85:15, 80:20, 75:25, 70:30, 65:35 and 60:40. In medical data, the description of the degree of imbalance is more commonly used in indicators such as the positivity rate, incidence, and response rate. Therefore, we used the positivity rate to express the different imbalance degrees. That is, datasets with positive rates of 1%, 3%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40% were constructed.

To explore the impact of different positive rates on the logistic regression model, we needed to determine a sample size that would be meaningful across various scenarios. Given that logistic models often perform poorly when the positive rate is below 5% [15], we selected a commonly considered extreme imbalance scenario, setting the overall positive rate (π) at 0.05. The following random sampling formula was used for calculation:

$$n = \frac{Z^2_{\alpha/2}\pi\ (1-\pi\ )}{\delta^{\ 2}}$$

Where the test level (α) was set at 0.05 and the allowable error (δ) at 0.01, resulting in a calculated sample size of approximately 1825. To facilitate the construction of datasets, the sample size was appropriately increased, and the sample size of each positive rate dataset was fixed at 2000. Additionally, we conducted studies on datasets with sample sizes of 1000 and 5000 to verify the stability and consistency of the positive rate cutoff value across different sample sizes.

To construct the datasets, the positive samples and negative samples in the original dataset were randomly sampled according to a certain proportion, and then the positive samples and negative samples were combined to form datasets with different positive rates. For example, 100 positive and 1900 negative samples were selected from the original dataset to construct a dataset with a positive rate of 5%, and 200 positive and 1800 negative samples were selected from the original dataset to construct a dataset with a positive rate of 10%, and so on.

In order to avoid the chance of random sampling, 100 datasets were constructed for each positive rate, and the 100 datasets were tested separately. Finally, the average value of 100 experiments was taken as the evaluation index of the model to minimize random sampling errors.

Construction of datasets with different sample sizes

Based on the optimal cut-off value of 15% of the positivity rate obtained from the first part of this study, we studied the effect of different sample sizes on the classification model under the same imbalance degree. The random sampling method was used to construct datasets of different sample sizes, in which the sample sizes of the datasets were 500, 800, 1000, 1200, 1500, 2000, 3000, 4000, and 5000, and the sample positivity rate was fixed at 15%.

The process of dataset construction was consistent for all of the datasets of differently imbalanced degrees. Similarly, to avoid the chance of random sampling, 100 datasets were constructed for each sample size, and each was tested separately. Finally, the

average value of 100 experiments was taken as the evaluation index of the model to minimize random sampling error.

Construction of datasets with low positive rate and low sample size

According to the optimal cut-off value of positive rate of 15% and the optimal cut-off value of sample size of 1500 obtained in the previous two steps, we compared the various imbalanced data processing methods under the condition of a low positive rate and low sample size. The distribution of the constructed dataset is shown in Supplemental Table 1.

To avoid the chance of random sampling, 500 datasets were constructed for each condition, and each was tested separately. Finally, the average value of these 500 experiments was taken as the evaluation index of the model, to minimize random sampling errors.

Introduction to classification models and resampling methods

Logistic model

The logistic model is a generalized linear regression model primarily utilized for addressing regression problems involving categorical dependent variables, particularly in the context of binary classification. In the realm of classification problems, the logistic model finds extensive application in medical research due to its convenience, interpretability, and efficient algorithmic approach. Thus, we adopted the logistic classification model as a representative to explore strategies for analyzing imbalanced data.

Resampling methods

The main idea of resampling is to balance the data distribution by reasonably adding some minority class samples or reducing some majority class samples, to reduce the impact of a skewed class distribution in the classification process. Resampling methods are divided into undersampling, oversampling, and mixed sampling techniques. In this study, four commonly used oversampling methods and undersampling methods were used to deal with imbalanced data.

(1) SMOTE oversampling

The basic principle of SMOTE oversampling is to synthesize new minority-class samples using linear interpolation between two minority-class samples [37]. Compared with the random oversampling method, this method greatly avoids the problem of overfitting in model training.

(2) ADASYN oversampling

The main feature of the ADASYN oversampling algorithm is that the number of new samples that needs to be synthesized by each minority class sample during sample synthesis is determined by the difficulty of its learning [38]. Specifically, in the *k*-nearest neighbors of the minority class sample, the more samples that belong to the majority class, the more difficult it is to learn, and the more minority samples it synthesizes. This is the most important difference between ADASYN oversampling and SMOTE oversampling, which requires the same number of synthetic samples for each minority class sample.

(3) OSS undersampling

OSS undersampling is an undersampling algorithm proposed by Kubat et al. in 1997, which is a combination of the Tomek link and CNN methods [39]. According to the distance between samples, the majority class samples are divided into four categories: noise samples, boundary samples, redundant samples, and safe samples. Then, the noise, boundary, and redundant samples are removed, and only safe samples are left, so as to balance the number of samples of the majority and minority classes.

(4) CNN undersampling

CNN undersampling deletes redundant samples in the majority class that are far away from the classification decision surface by looking for a consistency subset, so as to achieve undersampling [40].

Evaluation indicators

All datasets were divided into a training set and test set at a ratio of 7:3, and then the training set was used to train the logistic model and the test set to evaluate the model. For the dataset with low positive rate and low sample size, the logistic model was established after the training set was processed by an imbalance processing method after the dataset was divided, and the model was evaluated in the test set.

To compare the performance of the model constructed by different datasets and objectively and fully measure the classification effect of imbalanced data, we used the metrics of accuracy, recall, precision, F1-Score, geometric mean (the G-mean), and the area under the receiver operating characteristic curve (AUC) for the evaluation.

Results

Description of characteristics of the dataset

This dataset comprised 17,860 samples and 45 variables. After data cleaning, the dataset retained 15,764 samples and 43 variables, including 10,874 cases (68.9%) with cumulative live birth outcome events and 4890 cases (31.1%) without such events.

To avoid the problems of high dimensionality and model overfitting in logistic regression, the random forests algorithm was used to evaluate the importance of variables and select the variables that are important to the prediction objectives. In the parameter settings of the random forests algorithm, ntree=200 and mtry=6. According to the MDA index, the importance of variables was ranked and the 15 variables with the highest importance were selected in turn, and the results are shown in Table 1. These variables include 10 quantitative variables and 5 qualitative variables. The quantitative variables were maternal age, paternal age, basal antral follicle counts, basal endometrial thickness, basal luteinizing hormone, normal sperm rate, HCG day estradiol, number of oocytes obtained, number of embryos transferable, and number of high-quality embryos. The qualitative variables were the treatment scheme, delivery history, number of transplanted embryos, embryo development days, and scarred uterus. These 15 variables were used for the subsequent analysis.

Impact of imbalance degree on the classification model

Model performance indicators under different positive rates

Table 2 shows the values of the Akaike information criterion (AIC) and Bayesian information criterion (BIC) of the model performance evaluation indicator output by the

Table 1 Ranking of variable importance of random forests

Variable	MDA	MDG
Number of transferable embryos	53.2	903.0
Maternal age	38.1	433.3
Paternal age	20.3	285.5
Number of oocytes obtained	19.0	300.5
Basal antral follicle counts	17.7	229.3
Number of high-quality embryos	17.3	299.0
HCG day estradiol	16.8	330.3
Treatment scheme	11.7	75.6
Basal endometrial thickness	10.0	256.5
Basic luteinizing hormone	8.4	249.2
Delivery history	8.0	56.1
Number of transplanted embryos	7.9	33.0
Embryo development days	6.9	29.3
Scarred uterus	6.8	25.5
Normal sperm rate	6.6	228.2

Table 2 Comparison of performance indexes of logistic model under different positive rates ($n = 2000$)

Positive rates	AIC			BIC		
	Mean	SD	CV (%)	Mean	SD	CV (%)
1%	145.1	13.4	9.2	229.0	13.4	5.9
3%	310.7	19.8	6.4	394.6	19.8	5.0
5%	444.8	24.4	5.5	528.7	24.4	4.6
10%	699.3	29.4	4.2	783.2	29.4	3.8
15%	890.1	28.8	3.2	974.0	28.8	3.0
20%	1050.5	33.8	3.2	1134.4	33.8	3.0
25%	1166.0	32.7	2.8	1249.9	32.7	2.6
30%	1261.5	31.6	2.5	1345.4	31.6	2.3
35%	1332.2	43.7	3.3	1416.1	43.7	3.1
40%	1388.3	39.9	2.9	1472.2	39.9	2.7

Note: In the case of the same positive rate, the standard deviations of AIC and BIC are consistent, which is related to its calculation formula (AIC = -2ln(L) + 2k, BIC = -2ln(L) + k*ln(n). Where ln(L) represents the maximum log-likelihood function value of the model, k represents the number of parameters of the model, and n represents the sample size)

logistic model in the training set. As seen in the table, the coefficient of variation (CV) for the model's AIC and BIC values gradually decreases with an increase in the positive rate. This indicates that the more balanced the dataset, the more stable the logistic model results. Similar trends were observed for the sample sizes of 1000 and 5000, as shown in Supplemental Tables 2 and 3.

Classification effect indicators of the model under different positive rates

To evaluate the classification effect of the model in the test set, the AUC, G-mean, F1-Score, accuracy, precision, and recall were used. Supplemental Table 4 shows the evaluation results of the classification effect of the logistic model in the test set under different positive rates when the sample size is 2000. The results for sample sizes of 1000 and 5000 are presented in Supplemental Tables 5 and 6, respectively.

According to the experimental results in Supplemental Table 4, a raincloud plot was drawn to visualize the distribution of the evaluation indicators of the logistic model in each positive rate situation. As can be seen from the change trend of the mean value of the six evaluation indicators in Fig. 1, the AUC, G-mean, F1-Score, precision, and recall

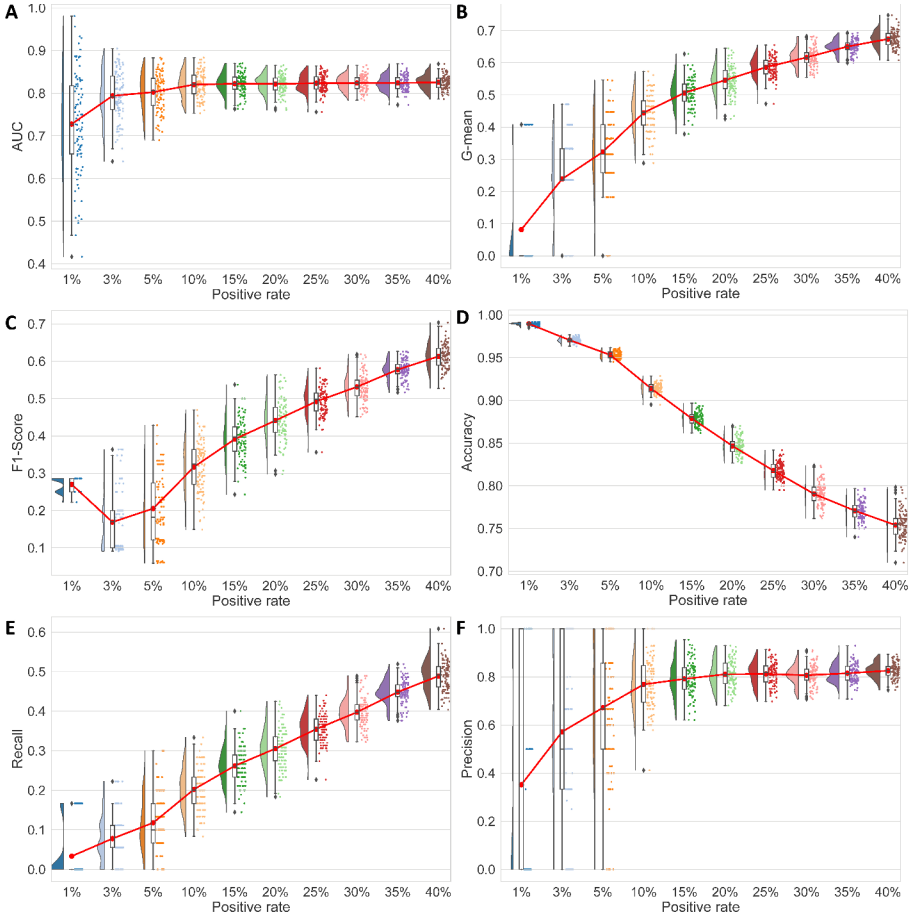

Fig. 1 Raincloud plot of the classification effect of the logistic model under different positive rates ($n = 2000$)

all show a gradually increasing trend as the data become more balanced, whereas the accuracy shows a decreasing trend. This occurs because, under significant data imbalance, the logistic model tends to sacrifice minority class accuracy to achieve higher overall accuracy.

Since the imbalanced data problem requires a comprehensive consideration of two types of errors, special attention should be paid to three key indicators: the AUC, G-mean, and F1-Score. According to the AUC index in Fig. 1(A), when the positive rate of the dataset is less than 10%, the AUC increases rapidly. When the positive rate reaches 10%, the AUC stabilizes with a small range of change. According to the G-mean index in Fig. 1(B), when the positive rate of the dataset is less than 10%, the G-mean rapidly increases, and when the positive rate reaches 10%, it increases slowly with a small range of change. As for the F1-Score in Fig. 1(C), it initially decreases and then increases when the positive rate is below 10%. This is because, when the data are extremely imbalanced, the number of minority samples is small and the number of minority samples divided into the test set is even smaller. When the logistic model is tested in the test set, the true positive (TP) value is often small or even zero. In this case, the model will sacrifice a small number of class samples to obtain higher accuracy, resulting in small false negative (FN) values and large fluctuations in recall and precision. When the positive rate of dataset reaches 10%, the F1-Score steadily and slowly increases. In addition, when the sample size is 1000 and 5000 respectively, AUC, G-mean, and F1-Score also exhibit the same

changing trend (Supplemental Figs. 1 and 2). It can be seen that a positive rate of 10% in the dataset may be an important cut-off value affecting the classification model. When the positive rate of dataset is less than 10%, the classification accuracy of the model is low. When the positive rate of dataset is higher than 10%, the classification performance of the model gradually stabilizes. Considering the robustness of selecting a truncation value, the optimal cut-off value for the positive rate that affects the classification model is determined to be 15%.

Impact of sample size on classification model

Model performance under different sample sizes

Table 3 shows the model performance evaluation index AICs and BICs output by the logistic model in the training set. As seen in the table, the CV for the model's AIC and BIC values gradually decreases with an increase in the sample size. This indicates that the larger the sample size of the dataset, the more stable the results of the logistic model.

Classification effect of the model under different sample sizes

To evaluate the classification effect of the model in the test set, the AUC, G-mean, F1-Score, accuracy, precision, and recall were used to evaluate the classification effect. Supplemental Table 7 shows the evaluation results of the classification effect of the logistic model in the test set under different sample sizes.

According to the experimental results in Supplemental Table 7, a raincloud plot was drawn to visualize the distribution of the evaluation indicators of the logistic model in each sample size situation. As observed in the change trend of the mean values of the six evaluation indicators in Fig. 2, the AUC, G-mean, F1-Score, recall and precision all gradually increase with the increase in the dataset's sample size, while the accuracy remains relatively stable.

Since the problem of imbalanced data requires a comprehensive consideration of two types of errors, it is necessary to pay more attention to three indicators: the AUC, G-mean, and F1-Score. As shown in Fig. 2(A–C), the AUC, G-mean, and F1-Score of the three indexes show a small and slow upward trend with increasing sample size. After the sample size reaches 1200, the three indexes gradually stabilize. Hence, a sample size of 1200 may be an important cut-off value affecting the classification model. When the sample size is less than 1200, the classification accuracy of the model is low, but when

Table 3 Comparison of performance indexes of logistic model under different sample sizes (positive rate = 15%)

Sample sizes	AIC			BIC		
	Mean	SD	CV (%)	Mean	SD	CV (%)
500	236.1	18.2	7.7	297.9	18.2	6.1
800	366.8	19.8	5.4	436.1	19.8	4.5
1000	454.3	21.8	4.8	527.1	21.8	4.1
1200	541.3	24.5	4.5	617.0	24.5	4.0
1500	673.1	28.0	4.2	752.4	28.0	3.7
2000	889.3	30.5	3.4	973.2	30.5	3.1
3000	1332.6	41.3	3.1	1423.0	41.3	2.9
4000	1763.2	45.7	2.6	1858.2	45.7	2.5
5000	2197.9	40.3	1.8	2296.5	40.3	1.8

Note: In the case of the same positive rate, the standard deviations of AIC and BIC are consistent, which is related to its calculation formula (AIC = -2ln(L) + 2k, BIC = -2ln(L) + k*ln(n). Where ln(L) represents the maximum log-likelihood function value of the model, k represents the number of parameters of the model, and n represents the sample size)

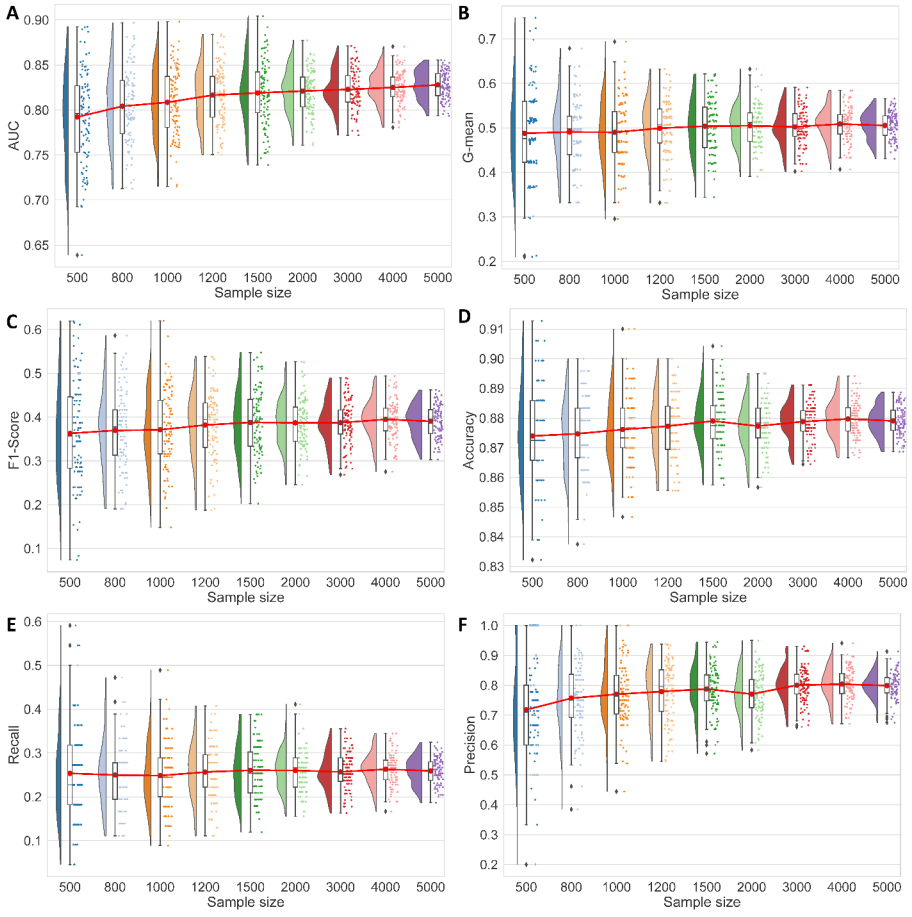

Fig. 2 Raincloud plot of the classification effect of the logistic model under different sample sizes (positive rate = 15%)

it is higher than 1200, the classification performance of the model gradually stabilizes. Considering the robustness of selecting a truncation value, the optimal cut-off value for the sample size that affects the classification model is determined to be 1500.

Impact of imbalanced data processing method on classification model

Supplemental Table 8 described the basic situation of the datasets after using an imbalanced data processing method. After SMOTE oversampling, the dataset had an equal number of minority and majority samples, resulting in a positive rate of 50%. After ADASYN oversampling, the number of minority samples was close to that of the majority, and the positive rate of the dataset remained about 50%. After the OSS undersampling process, the imbalance of the dataset improved to a certain extent, but remained in an imbalanced state. After CNN undersampling, the imbalance of the dataset significantly improved, but the sample size of the dataset decreased significantly, due to the large number of deleted class samples.

Figure 3 showed the AUC of the logistic model after the data were processed by imbalanced data processing methods. It can be seen that, compared to the untreated group, the AUC values slightly decreased after SMOTE oversampling, ADASYN oversampling, and CNN undersampling, whereas the difference was very small for OSS undersampling compared to the untreated group.

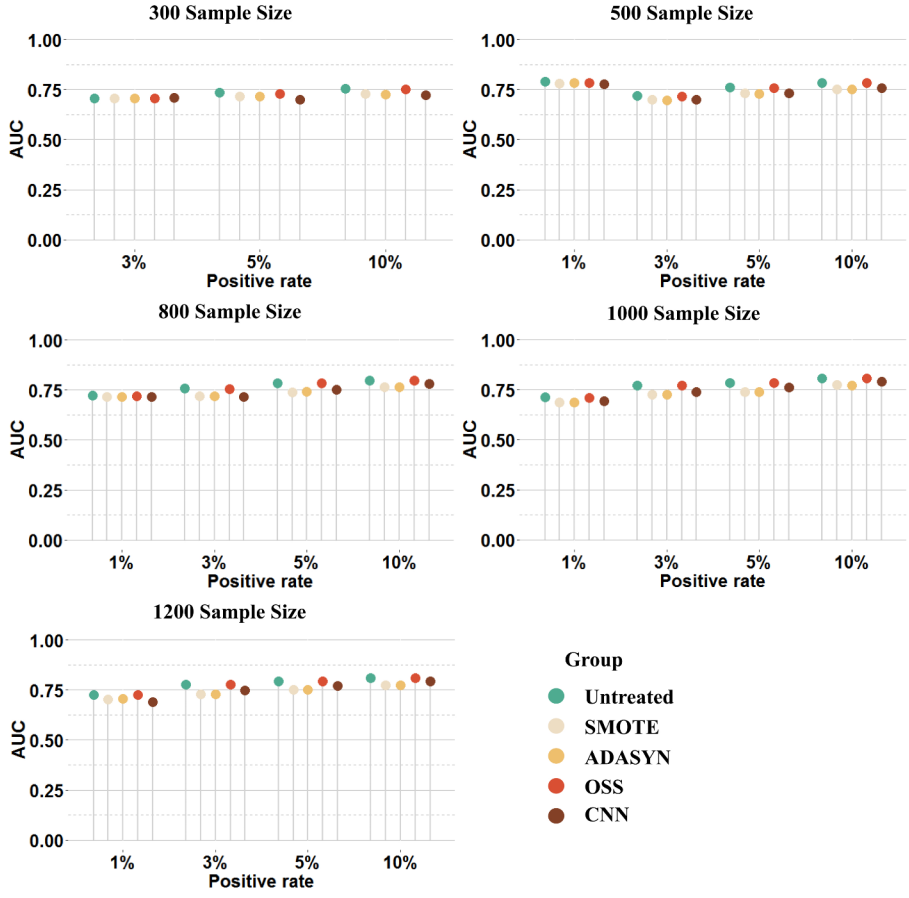

Fig. 3 AUC values of Logistic model after imbalanced processing method in various cases

For the G-mean metric (Fig. 4), SMOTE oversampling, ADASYN oversampling, and CNN undersampling showed a significant increase compared to the untreated group, while the G-mean slightly increased compared to the untreated group with OSS undersampling. When the positive rate was 1%, the CNN undersampling method was superior to the two oversampling methods, but when the positive rate was greater than 1%, the two oversampling methods outperformed the two undersampling methods. Furthermore, the larger the sample size, the more significant the improvement brought by the imbalanced data processing methods to the logistic model.

For the F1-Score metric (Fig. 5), when the positive rate was 10%, SMOTE oversampling, ADASYN oversampling, and CNN undersampling showed improvement compared to the untreated group. The accuracy, recall, and precision are shown in Supplemental Figs. 3, 4, and 5, respectively.

Discussion

In this study, datasets with different imbalanced proportions were constructed by random sampling of the original datasets, and then logistic models were established for each dataset. Logistic models are widely recognized and applied in the medical field, but they require a balanced or nearly balanced training sample set [3, 4]. Our results indicate that the classification performance of the logistic model gradually declines as the imbalance degree of the dataset increases, which aligns with the findings of other

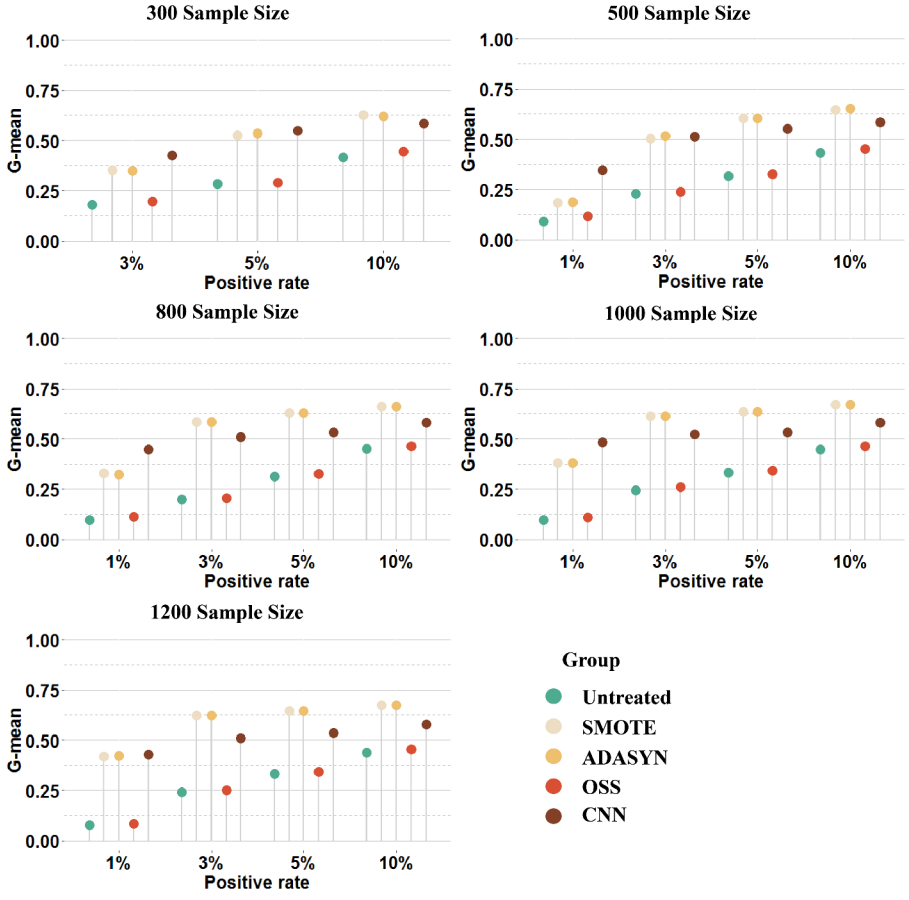

Fig. 4 G-mean values of Logistic model after imbalanced processing method in various cases

studies [41, 42]. This indicates that the imbalance of the dataset will hinder the classification performance of the model. Therefore, the logistic model is unsuitable for addressing data classification problems when the dataset is extremely imbalanced. Because of the many advantages of logistic models, our results showed that when the positive rate of the dataset is greater than 10%, logistic analysis can also get better results. Considering the robustness of the selection of the truncation value, the value should be appropriately expanded. Therefore, this study determined the optimal truncation value of the imbalance degree of the dataset affecting the logistic model at a positive rate of 15%. Data resampling methods such as oversampling and undersampling change the distribution of original data to some extent [32]. Compared to logistic models, machine learning methods such as ensemble learning and cost-sensitive learning have higher complexity and lack intuitive interpretations of model results. Therefore, when weighing the choice between traditional classification models and an imbalanced data treatment method, the positive rate of 15% obtained in this study is a good reference.

Our study found that sample size affects the classification performance of the logistic model. When the sample size is small, the classification accuracy of the model is low, making it susceptible to noise and randomness. As the sample size increases, the model better learns the characteristics and distribution of the minority class, thereby improving its classification accuracy and generalization ability. We found that when the sample size reached 1200, the performance of the logistic model tended to stabilize.

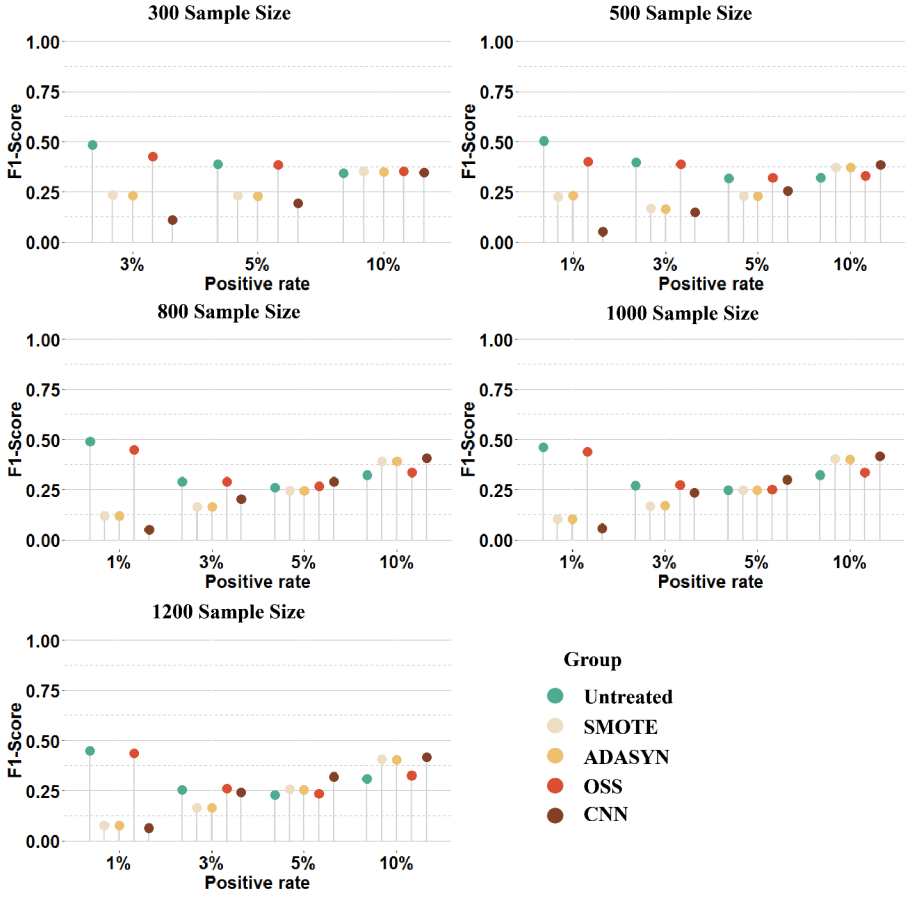

Fig. 5 F1-Score values of Logistic model after imbalanced processing method in various cases

This suggests that further increasing the sample size does not significantly enhance the model's performance. This may be because, when the sample size is sufficiently large, the model has already effectively learned the distinctions between the minority and majority class samples, rendering further increases in sample size of limited benefit to model performance. Therefore, considering the robustness of the selection of the truncation value, the value is appropriately expanded, so the optimal truncation value of the sample size of the imbalanced dataset affecting the logistic model is determined to be 1500. However, this sample-size truncation value is obtained when the data imbalance is a 15% positive rate. When it is generalized to data that are seriously imbalanced, this truncation value may be restricted by certain conditions.

To explore the impact of imbalanced data processing methods on the model, we used four methods to process the dataset: SMOTE oversampling, ADASYN oversampling, OSS undersampling, and CNN undersampling. SMOTE and ADASYN are two classic oversampling methods, with numerous improved algorithms for these methods having been applied across various scenarios [43, 44]. As can be seen from the study results, under various conditions of a low positive rate and low sample size, SMOTE oversampling and ADASYN oversampling significantly improved the classification effect of the logistic model compared with unprocessed data, and there was little difference between them. For undersampling, CNN undersampling realizes data undersampling by deleting redundant samples far from the classification decision surface in most class samples

[40]. The results showed that under various conditions of a low positive rate and low sample size, CNN undersampling significantly improved the classification effect of the logistic model, but the improvement effect was slightly lower than that with SMOTE oversampling and ADASYN oversampling. Undersampling involves deleting a substantial number of majority class samples, which can result in the loss of valuable information to some extent. OSS undersampling balances data by removing some noisy data, boundary data, and redundant data [39]. Compared with SMOTE oversampling, ADASYN oversampling, and CNN undersampling, OSS undersampling did not significantly improve the classification effect of the logistic model under various conditions of a low positive rate and low sample size. This also shows that there are some differences between different resampling methods and imbalanced data treatment methods should be selected carefully to obtain better results. In addition, according to the results of this study, under various conditions of a low positive rate and low sample size, oversampling outperformed undersampling. Therefore, when it is necessary to adopt an imbalanced data processing method, we recommend using an oversampling method for processing.

This study also has some limitations. Although we thoroughly explored various imbalance degrees and sample sizes, the imbalance degrees were analyzed based on a fixed sample size of 1000, 2000, and 5000, and the sample size analysis was conducted with a fixed positive rate of 15%. There may be some limitations in generalizing these findings to other data scenarios. Further research is needed to evaluate these relationships across a broader range of conditions. Additionally, our study focused exclusively on the classical logistic model, without considering other classification models such as discriminant analysis, decision trees, and support vector machines. These alternative models might produce different cut-off values for imbalanced proportions and sample sizes. Future studies should include multiple models to achieve more comprehensive and generalizable results.

Conclusions

When the positive rate of a dataset reaches 15%, the logistic model's classification performance stabilizes, establishing this rate as the optimal cut-off for imbalance. Similarly, a sample size of 1500 ensures stable model performance, making it the recommended minimum sample size. For datasets with a positive rate below 10% and a sample size under 1200, oversampling techniques such as SMOTE and ADASYN are advised to achieve better balance and improve classification accuracy.

Abbreviations

ADASYN	Adaptive Synthetic Sampling
CNN	Condensed Nearest Neighbor
CV	Coefficient of Variation
MDA	Mean Decrease Accuracy
MDG	Mean Decrease Gini
OSS	One-Sided Selection
ROC	Receiver Operating Characteristic curve
SMOTE	Synthetic Minority Over-Sampling Technique

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13040-024-00384-y.

Supplementary Material 1

Acknowledgements
We gratefully acknowledge the financial support from the Science Research Project of Education Department of Liaoning Province (Project ID LJKZ0765), the Science Research Project of Shenyang City (Project ID 23-506-3-01-21), and the Science and Technology Planning Project of Liaoning Province (Project ID 2021JH4/10200008). Special thanks to the medical staff at the medical staff of the Jiangxi Maternal and Child Health Hospital for their collaboration in conducting this study, as well as the patients who participated in this research.

Author contributions
JLZ: Writing - review & editing, Writing - original draft, Methodology, Visualization, Software, Project administration. SWP: Resources, Investigation, Formal analysis. JJH: Writing - review & editing, Visualization, Validation. DCS: Software, Resources, Data curation. WJC: Software, Validation. XYX: Methodology, Conceptualization. HBL: Writing - review & editing, Resources, Project administration, Funding acquisition, Conceptualization. All authors read and approved the final manuscript.

Funding
This research was partially supported by the Science Research Project of Education Department of Liaoning Province (Project ID LJKZ0765), the Science Research Project of Shenyang City (Project ID 23-506-3-01-21), and the Science and Technology Planning Project of Liaoning Province (Project ID 2021JH4/10200008).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
The ethical matters of biomedical research involving humans in this project meet the requirements of the Declaration of Helsinki and the Measures for Ethical Review of Life Science and Medical Research Involving Humans. The Medical Ethics Review Committee of Jiangxi Provincial Maternal and Child Health Hospital approved the implementation of this project according to the research plan (SZYX-202305). Signed informed consent was obtained from all patients before enrollment.

Consent for publication
All authors have read and agreed to the published version of the manuscript.

Competing interests
The authors declare no competing interests.

Received: 21 June 2024 / Accepted: 27 August 2024
Published online: 04 September 2024

References

1. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–18.
2. Lu S, Yang J, Gu Y, He D, Wu H, Sun W, et al. Advances in machine learning processing of big data from disease diagnosis sensors. ACS Sens. 2024;9(3):1134–48.
3. Shi SN, Li J, Zhu D, Yang F, Xu Y. A hybrid imbalanced classification model based on data density. Inf Sci. 2023;624:50–67.
4. Zhao JK, Jin J, Chen S, Zhang RF, Yu BL, Liu QF. A weighted hybrid ensemble method for classifying imbalanced data. Knowl-Based Syst. 2020;203:106087.
5. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput. 2013;3(2):224.
6. Fotouhi S, Asadi S, Kattan MW. A comprehensive data level analysis for cancer diagnosis on imbalanced data. J Biomed Inf. 2019;90:103089.
7. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: a systematic literature review. Artif Intell Med. 2022;128:102289.
8. Drouard G, Mykkänen J, Heiskanen J, Pohjonen J, Ruohonen S, Pahkala K, et al. Exploring machine learning strategies for predicting cardiovascular disease risk factors from multi-omic data. BMC Med Inf Decis Mak. 2024;24(1):116.
9. Ren Y, Wu D, Tong Y, López-DeFede A, Gareau S. Issue of data imbalance on low birthweight baby outcomes prediction and associated risk factors identification: establishment of benchmarking key machine learning models with data rebalancing strategies. J Med Internet Res. 2023;25:e44081.
10. Dablain D, Krawczyk B, Chawla NV. DeepSMOTE: fusing deep learning and SMOTE for imbalanced data. IEEE Trans Neural Netw Learn Syst. 2023;34(9):6390–404.
11. Rezvani S, Wang X. A broad review on class imbalance learning techniques. Appl Soft Comput. 2023;143:110415.
12. Gong J, Kim H, RHSBoost. Improving classification performance in imbalance data. Comput Stat Data An. 2017;111:1–13.
13. Zhang L, Geisler T, Ray H, Xie Y. Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. J Appl Stat. 2022;49(13):3257–77.
14. Charizanos G, Demirhan H, İçen D. A Monte Carlo fuzzy logistic regression framework against imbalance and separation. Inf Sci. 2024;655:119893.
15. Li J, Fong S, Mohammed S, Fiaidhi J. Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms. J Supercomput. 2016;72(10):3708–28.
16. Kim KH, Sohn SY. Hybrid neural network with cost-sensitive support vector machine for class-imbalanced multimodal data. Neural Netw. 2020;130:176–84.

17. Wang Z, Zheng M, Liu PX. A novel classification method based on stacking ensemble for imbalanced problems. IEEE Trans Instrum Meas. 2023;72:1–13.
18. Maldonado S, Vairetti C, Fernandez A, Herrera F. FW-SMOTE: a feature-weighted oversampling approach for imbalanced classification. Pattern Recognit. 2022;124:108511.
19. Ng WWY, Xu S, Zhang J, Tian X, Rong TW, Kwong S. Hashing-based undersampling ensemble for imbalanced pattern classification problems. IEEE Trans Cybern. 2022;52(2):1269–79.
20. Peng P, Zhang W, Zhang Y, Xu YY, Wang HW, Zhang HM. Cost sensitive active learning using bidirectional gated recurrent neural networks for imbalanced fault diagnosis. Neurocomputing. 2020;407:232–45.
21. Alves Ribeiro VH, Reynoso-Meza G. Ensemble learning by means of a multi-objective optimization design approach for dealing with imbalanced data sets. Expert Syst Appl. 2020;147:113232.
22. Parlak B. Class-index corpus-index measure: a novel feature selection method for imbalanced text data. CONCURR COMP-PRACT E. 2022;34(21):e7140.
23. Fu GH, Wu YJ, Zong MJ, Pan J. Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data. BMC Bioinformatics. 2020;21(1):121.
24. Kosolwattana T, Liu C, Hu R, Han S, Chen H, Lin Y. A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare. BioData Min. 2023;16(1):15.
25. Beinecke J, Heider D. Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. BioData Min. 2021;14(1):49.
26. Parlak B, Uysal AK. On feature weighting and selection for medical document classification. In developments and advances in intelligent systems and applications. Stud Comput Intell. 2018;718:269–82.
27. Labory J, Njomgue-Fotso E, Bottini S. Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data. Comput Struct Biotechnol J. 2024;23:1274–87.
28. Parlak B, Uysal AK. A novel filter feature selection method for text classification: extensive feature selector. J Inf Sci. 2023;49(1):59–78.
29. Moniz N, Monteiro H. No free lunch in imbalanced learning. Knowl-Based Syst. 2021;227:107222.
30. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal. 2002;6(5):429–49.
31. Batista GE, A P A, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6(1):20–9.
32. Vimalraj S, Porkodi Dr R. A review on handling imbalanced data. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT). IEEE, 2018.
33. Wei JN, Huang HS, Yao LG, Hu Y, Fan QS, Huang D. NI-MWMOTE: an improving noise-immunity majority weighted minority oversampling technique for imbalanced classification problems. Expert Syst Appl. 2020;158:113504.
34. Beckmann M, Ebecken NFF, De Lima BSLP. A KNN undersampling approach for data balancing. JILSA. 2015;7(4):104.
35. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl. 2004;6(1):20–9.
36. Nakamura M, Kajiwara Y, Otsuka A, Kimura H. LVQ-SMOTE - learning vector quantization based synthetic minority oversampling technique for biomedical data. BioData Min. 2013;6(1):16.
37. Li J, Fong S, Sung Y, Cho K, Wong R, Wong KKL. Adaptive swarm cluster-based dynamic multi-objective synthetic minority oversampling technique algorithm for tackling binary imbalanced datasets in biomedical data classification. BioData Min. 2016;9:37.
38. Munshi RM. Novel ensemble learning approach with SVM-imputed ADASYN features for enhanced cervical cancer prediction. PLoS ONE. 2024;19(1):e0296107.
39. Jia C, Zuo Y, S-SulfPred:. A sensitive predictor to capture S-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique. J Theor Biol. 2017;422:84–9.
40. Devi D, Biswas SK, Purkayastha B. Redundancy-driven modified Tomek-link based undersampling: a solution to class imbalance. Pattern Recognit Lett. 2017;93:3–12.
41. Zhou L, Lai KK. Benchmarking binary classification models on data sets with different degrees of imbalance. Front Comput Sci Chi. 2009;3(002):205–16.
42. Yang H, Li XX, Cao HY, Cui YH, Luo YH, Liu JC. Using machine learning methods to predict hepatic encephalopathy in cirrhotic patients with unbalanced data. Comput Meth Prog Bio. 2021;211:106420.
43. Zhang AM, Yu HL, Huan ZJ, Yang XB, Zheng S, Gao S. SMOTE-RkNN: a hybrid re-sampling method based on SMOTE and reverse k-nearest neighbors. Inf Sci. 2022;595:70–88.
44. Özdemir A, Polat K, Alhudhaif A. Classification of imbalanced hyperspectral images using SMOTE-based deep learning methods. Expert Syst Appl. 2021;178:114986.

Publisher's note