Cluster-Based Ensemble Learning Model for Rapid Detection of Aortic Dissection

Yan Gao
Central South University
Min Wang
Central South University
Guogang Zhang
Central South University
Lingjun Zhou
Central South University
Jingming Luo
Central South University
Lijue Liu (ljliu@csu.edu.cn)
Central South University

Research Article

Keywords: aortic dissection, resampling, imbalanced data, ensemble learning, bagging

Posted Date: July 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-653868/v1

Cluster-based ensemble learning model for rapid detection of Aortic Dissection

Yan Gao[1], Min Wang[1], Guogang Zhang[2], Lingjun Zhou[1], Jingming Luo[2], Lijue Liu[1§]
[1]School of Automation, Central South University, Hunan Province, China.
[2]Xiangya Hospital, Central South University, Hunan Province, China.

§ Corresponding author
Lijue Liu: ljliu@csu.edu.cn

Email addresses:
Yan Gao: gaoyan@csu.edu.cn
Min Wang: wm490602605@csu.edu.cn
Lijue Liu: ljliu@csu.edu.cn
Guogang Zhang: zhangguogang@csu.edu.cn
Lingjun Zhou: csu_0918@163.com
Jingming Luo: jmluo0618@139.com

Abstract

Background

Aortic dissection (AD) is a rare and high-risk cardiovascular disease with dangerous morbidity and high mortality, so it needs rapid diagnosis and timely treatment. However, due to its complex and changeable clinical manifestations and the lack of special symptoms and signs, it is easy to cause missed diagnosis and misdiagnosis.

Methods

The data set used in this paper comes from 53213 patients, which collected from XiangYa Hospital in Hunan Province from 2008 to 2016. The data includes 802 patients with aortic dissection and 52411 patients with non-aortic dissection. In order to help clinicians predict AD, we designed an ensemble learning model based on clustering: Cluster Random Under-sampling Smote-Tomek-link Bagging (CRST-Bagging). This model combines the advantages of clustering-based compound resampling (CRST) method and Bagging ensemble classifier. It achieves good results on aortic dissection data sets.

Results

The model validates the effectiveness of the CRST sampling method on the AD data set. We compared the CRST-Bagging model with the classical ensemble models RUSBoost and SMOTE-Bagging on the AD data set. The experimental results show that the CRST-Bagging model has the best performance in the detection of AD. Model's accuracy and recall rate are 83.6% and 80.7% respectively. And the F1 value is 82.1%, which is 4.8% and 1.6% higher than that of RUSBoost and SMOTE-Bagging model.

Conclusions

The model proposed in this paper can be used as an auxiliary diagnostic model of AD to provide reference for clinical medicine. The model can also help doctors to judge whether patients need further imaging examination. Thus help to reduce the rate of clinical misdiagnosis and missed diagnosis of AD.

Keywords: aortic dissection, resampling, imbalanced data, ensemble learning, bagging

1. Background and introduction

Aortic dissection (AD) is a medial rupture caused by intramural hemorrhage, which leads to the separation of aortic wall layer, followed by the separation of true and false lumen[1]. AD is a dangerous cardiovascular disease with dangerous morbidity, many complications and high mortality. The clinical manifestations of AD are complex and changeable. They lack of special symptoms and signs. And the location, lesion degree and scale of AD are different. So the clinical manifestations and severity are different. In addition, clinicians tend to observe the common symptoms of AD, such as chest pain and back pain. But for painless patients, atypical symptoms make the diagnosis more difficult. It is easy to cause missed diagnosis and misdiagnosis[2]. According to the clinical statistics, the misdiagnosis rate of AD is more than 1/3 in the actual cases of AD [3][4][5]. Mortality can reach as high as 50% within a week of onset and between 60 and 70% within a month. With the help of scientific methods and effective techniques, the timely diagnosis of AD by clinicians is the most effective means to save patients' lives.

In recent years, the application of artificial intelligence in the field of medical and health care has attracted much attention. At present, various auxiliary diagnosis systems in the medical field have emerged one after another. Huang et al. [6] used an enhanced resampling method of electronic medical records to classify and predict Major adverse cardiac events (MACEs) of acute coronary syndrome (ACS). Zhou et al. [7] proposed an interpretable pattern discovery method from the perspective of statistical learning methods to interpret clinical chest data and make classification predictions. Song et al. [8] used deep learning technology to develop a histopathological detection system for gastric cancer detection that can be used for clinical diagnosis. Kamal et al. [9] proposed a novel and interpretable PAVE model for sepsis attack prediction and mortality prediction estimates. Song et al. [10] used Bayesian subgroup analysis to evaluate the conditioned treatment effect of adjuvant treatment for patients with synovial sarcoma and help clinicians choose treatment options. Xia et al. [11] proposed a multi-level hypoglycemia early warning system based on sequential pattern mining based on continuous blood glucose monitoring data. With the accumulation of a large amount of medical data, researchers have used deep learning methods to achieve classification prediction[12][13] in the field of cardiovascular diseases. For example, Guo et al. [14] used the LSTM model to predict cardiovascular health trajectories in time series electronic health records; Cheng et al. [15] used deep learning techniques to classify AD using contrast-enhanced CT images. Among 1,000 CT images from 20 patients, the accuracy rate reached 85.0%. However, in the actual situation, due to the lack of clinician experience or unsupported examination equipment, it's difficult to carry out relevant imaging examination in time resulting in patient AD may not be diagnosed by means of CT image, leading to missed diagnosis and misdiagnosis, which threatens the life safety of patients. Therefore, different from the above methods, we studied the diagnostic and predictive

method of AD based on the routine examination data of patients, so as to help doctors judge whether patients need further imaging examination.

The rarity of AD leads to the significant imbalance in the data set. If the traditional machine learning technology is applied to the aortic dissection data set, the model tends to fit large samples, showing high accuracy and low recall rate, so the model's generalization ability is low. Therefore, imbalanced learning [16][17] and ensemble learning [18][19][20][21] were combined to predict aortic dissection data.

According to the characteristics of AD data, this paper proposes a cluster-based ensemble learning model: Cluster Random Under-sampling Smote-Tomek-link Bagging (Hereinafter referred to as CRST-Bagging) to help clinicians detect AD in clinical practice. This model includes two parts: Cluster-Based resampling (CRST) and Bagging classifier. The resampling method CRST combines the advantages of the over-and-under sampling method. It overcomes the difficulties in the detection of AD caused by imbalanced data. Bagging classifier is used to improve the generalization ability of the learning model. To demonstrate the effectiveness of the CRST-Bagging approach, we compared it with the classic ensemble models RUSBoost and Smote-Bagging on AD datasets. Experimental results show that the proposed model is superior to other models, which prove the effectiveness of the model.

The main contributions of this paper can be summarized as follows:

- In this paper, missing value processing , feature screening and dimension-reduction visualization were performed in the AD data set. These methods enable us to have a priori knowledge of the distribution of AD data, which can be used in clinical medicine to explore the pathological mechanism of AD.
- A new compound resampling method, CRST is proposed. This method combines the advantages of clustering ideology and SMOTE + Tomek-Link sampling methods. This method not only makes the collected samples effectively represent the characteristics of different samples, but also ensures the randomness of sampling, which can effectively reduce the imbalanced ratio of AD data.
- CRST-Bagging ensemble learning model is proposed to predict AD disease. Through experimental comparison and analysis, the model shows excellent performance and good generalization ability on AD data sets. Therefore, this model can be used for clinical auxiliary diagnosis.

The rest of this article is arranged as follows. In the second section, we introduce the data set used in this paper, our resampling method CRST and the imbalanced algorithm integration model. In the third section we present our experimental results and model performance evaluation. In the fourth section, the experimental results are discussed. Finally, the summarization and the discussion of future work are described in the last section.

2. Data

2.1 Data Overview

The dataset used in this paper comes from the examination indicators of 53213 patients, which collected from XiangYa Hospital in Hunan Province from 2008 to 2016. The data includes 802 patients with AD and 52411 patients without AD. The dataset has 71-dimensional features and 1-dimensional tags (The table of Description of Features is included as additional file 1 of the supplementary material). This dataset has a high imbalanced ratio, with the number of AD samples approximately 67 times that of non-AD samples. In addition, this paper also uses a test set to better verify the classification performance and generalization ability of our model. The test set includes the examination indicators of 235 patients from the same hospital, and the data format is the same as the above data set.

2.2 Data preprocessing

Firstly, some non-numerical indicators are normalized by binary coding, and then standardized. Secondly, we made statistics on the missing rate of samples and features in the original AD dataset as shown in Figure 1 (the abscissa represents the features and the ordinate represents the missing rate). Six features with a deletion rate of more than 50% were found, namely Plasma antithrombin Ⅲ antigen determination(r_28)(missing rate is 81.5%), Plasma plasminogen antigen determination (r_20)(missing rate is 80.7%), Hypersensitivity thyrotropin (r_64) (missing rate is 75.6%), erythrocyte sedimentation rate (r_59) (missing rate is 63.8%), D-dimer (r_19) (missing rate is 62.6%), free triiodothyronine (FT3) (r_62) (missing rate is 51.9%).

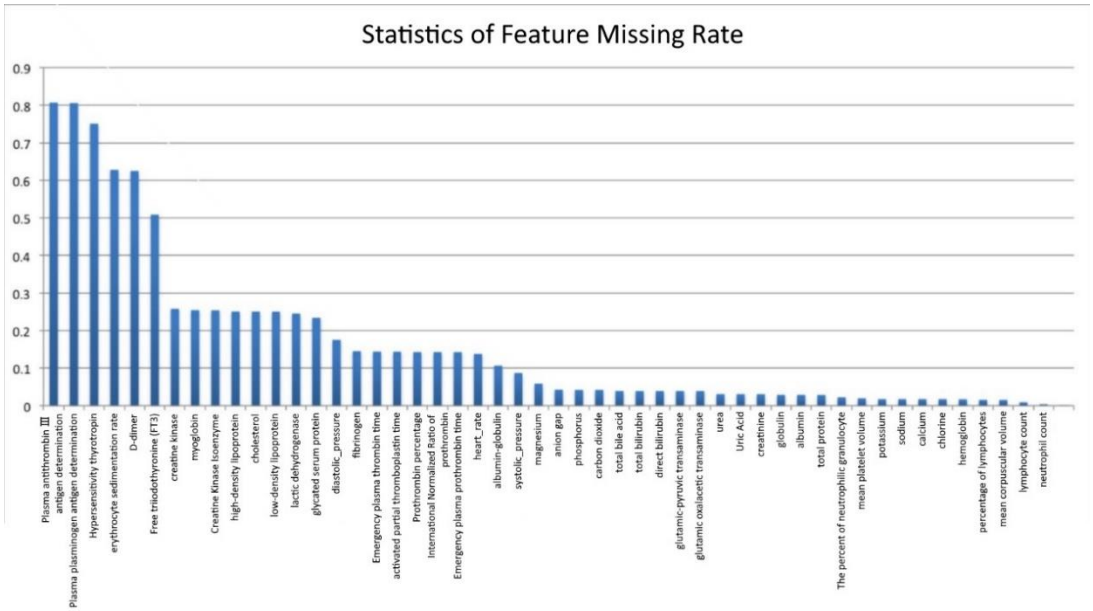


Figure 1. Feature Missing Rate Statistics

Due to the high-missing rate of the above-mentioned six-dimensional features, it is difficult to complete them. The general method is to delete them. However, the

etiology and related diagnostic indicators of AD are not yet clear, we cannot directly determine whether the missing feature indicators are key indicators, so they cannot simply be deleted. Therefore, the XGboost method is used to analyze feature importance [22][23]. The result is shown in Figure 2, with the horizontal coordinate as feature numbers and the vertical ordinate as feature importance scores.

From figure 2, we find that among the 6-dimensional features with a deletion rate greater than 50%, the feature importance scores of free triiodothyronine (FT3) and D-dimer are ranked in the top 10, which indicates that these two features are important for detecting whether a patient suffers from AD. Therefore, we only remove the four characteristics of Plasma antithrombin Ⅲ antigen determination, Plasma plasminogen antigen determination, Hypersensitivity thyrotropin, erythrocyte sedimentation rate. Free triiodothyronine (FT3) and D-dimer remain and are complemented with the remaining features. The adjusted new sample set size is (53213,67).

In this paper, the data set is filled by the method of classified random filling method. Compared with ordinary random filling, the method of random filling by class is to fill the positive and negative samples respectively. The missing values of the samples are randomly filled with the non-null values of the same kind of samples. This filling method can effectively avoid the intersection of feature values.

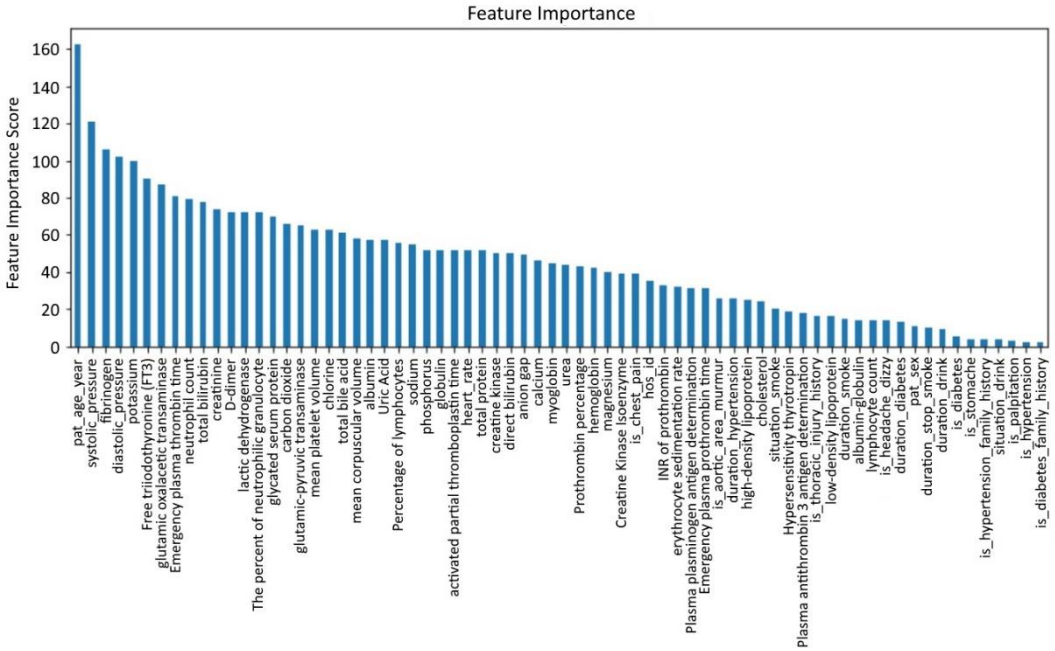


Figure 2. Feature importance analysis

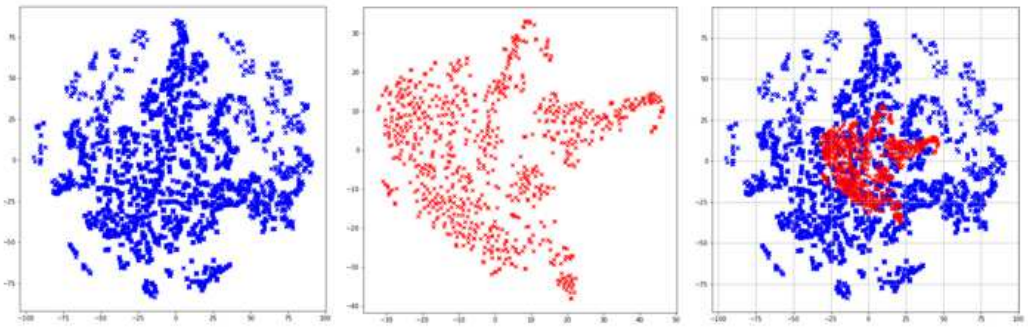
2.3 Dimensionality reduction visualization

In order to have a more systematic and in-depth understanding of the high-dimensional AD data set, this paper uses the method of dimensionality reduction visualization to analyze the data distribution. This allows us to understand the data more intuitively and provide information for the design of AD prediction algorithm.

We analyze the existing methods of dimensionality reduction. T-SNE[24] algorithm

can retain both global and local data structures. Therefore, we used the T-SNE algorithm to reduce the dimensionality of the dataset. We analyze the distribution of samples through dimensionality reduction and visualization. By observing the data distribution, it is concluded that the clustering algorithm is feasible to improve the under-sampling method.

The results of the AD dataset using the t-SNE method for dimensionality reduction are shown in Figure 3. The red sample point is the positive sample, and the blue sample point is the negative sample. As can be seen from figure 3, the data distribution of positive samples is agglomerated. This shows that there is a certain similarity between the cases of AD. Therefore, from the analysis of data distribution, the under-sampling method of clustering is effective. At the same time, it can also be seen that there is obvious overlap between the two kinds of samples in space from the visualization of data distribution. Therefore, it is necessary to construct a nonlinear classification model.



(a) (b) (c)

Figure 3. AD dataset t-SNE dimensionality reduction map. (a) The data distribution of negative samples after dimensionality reduction. (b) Data distribution of positive samples after dimensionality reduction. (c) The data distribution of positive and negative samples in the same space after dimensionality reduction.

3. Methods

According to the data characteristics of AD, this paper proposes an ensemble learning model based on clustering: Cluster Random Under-sampling Smote-Tomek-link Bagging (CRST-Bagging). The model structure is shown in figure 4. Data pre-processing is carried out first (see section 2). Then in the second step, a clustering-based resampling algorithm is proposed to resample the imbalanced data set to reduce the imbalanced ratio of the data. Finally, the Bagging ensemble model is used to construct a powerful nonlinear classifier to predict AD. The methods are described in detail below.

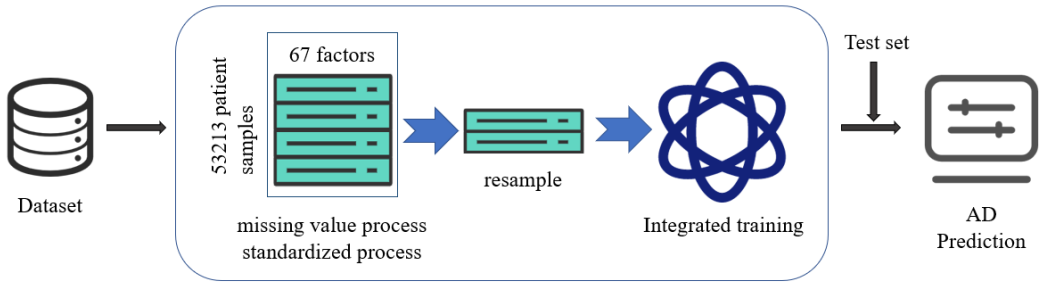


Figure 4. Model structure diagram

3.1 Cluster Random Under-sampling Smote + Tomek-link Approach（CRST）

Resampling technique is to obtain a balanced data set from the original non-balanced data set by using different sampling methods. From the perspective of data level, resampling methods are mainly divided into three kinds: over-sampling, under-sampling and combined sampling methods. The classic and commonly used methods are: SMOTE and its improved method[25][26][27], Tomek-links method[28], SMOTE + Tomek-Link, SMOTE + ENN[29] combination method, etc.

In view of the imbalanced and clustered characteristics of AD data, we proposed Cluster Random Under-Sampling Smote + Tomek-link Approach (hereafter, CRST). This method is an under-sampling method which takes the cluster center as the representative point. It combines the advantages of K-means++ and Smote + Tomek-link sampling method.

Firstly, the training samples in majority class were clustered by K-means++ algorithm, in which K is obtained by super-parameter optimization. Then random under-sampling is carried out for each cluster. The degree of sampling p% can be determined by the actual situation. After the under-sampling is completed, SMOTE + Tomek-Link combined sampling method is used to form a new balanced data set. By iterating the above operations many times, we get several new balanced sub-datasets.

The clustering of samples in majority classes is visualized in the figure 5. The green dots are the selected remaining of majority class sample points after p% random under-sampling for each cluster. It can be seen that in this way the sample points can be uniformly sampled in each cluster by under-sampling, maintaining the original data distribution.

Finally, Smote + Tomek-link(S-T) sampling method is applied to generate some minority samples, thus the sample loss caused by the under-sampling method is compensated and the imbalanced ratio is alleviated. As shown in Figure 6, S-T generates minority samples through SMOTE method, while Tomek-Link method is adopted to solve the problem of fuzzy boundary caused by excessive generation of minority samples. This method can reduce the redundancy of samples. The procedure of algorithm is shown in Table 1.

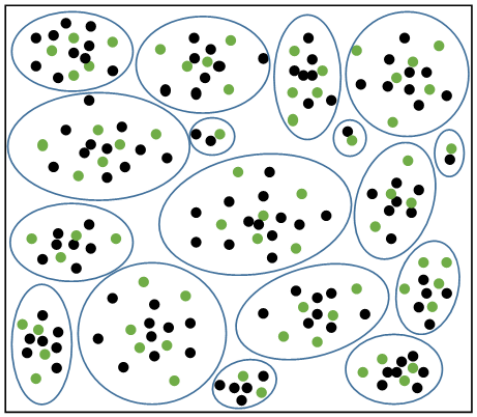


Figure 5. Schematic Diagram of the Under Sampling Process of majority samples

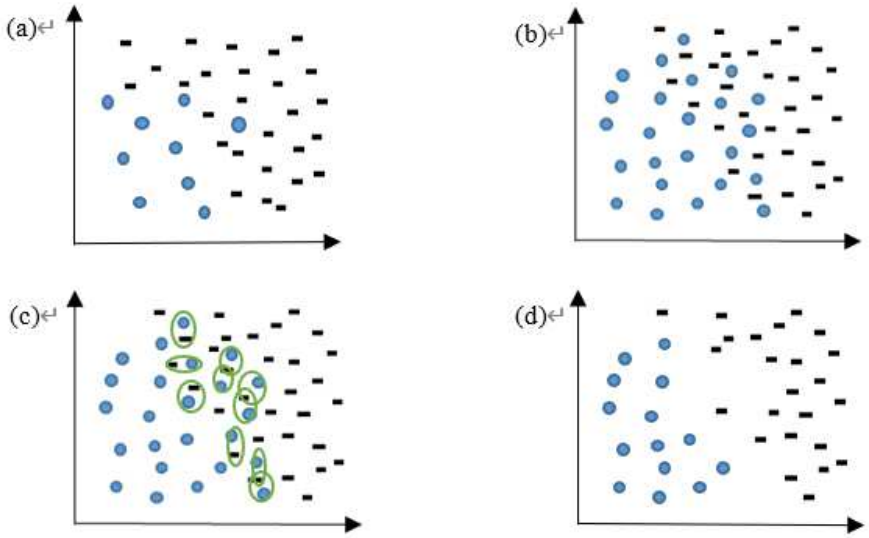


Figure 6. Raw data set (a); Data set after SMOTE over-sampling(b); Tomek-Link recognition process (C); Data set after the boundary and noise samples are removed (D).

Table 1. CRST Sampling Method

Input: All standardized training sample set P;

Output: The new balanced sub-datasets Z

Method:

1. Divide the input training sample set $\boldsymbol{P}$ into the majority sample set $\boldsymbol{P_{max}}$ and the minority sample set $\boldsymbol{P_{min}}$ according to the labels, and then remove the labels.
2. Use the K-means++ algorithm to cluster the training sample set $\boldsymbol{P_{max}}$ of majority classes to get K clusters. K is obtained by super-parameter optimization. denoted as: $\boldsymbol{\Omega} = \{\boldsymbol{C_1}, \boldsymbol{C_2}, \cdots, \boldsymbol{C_K}\}$, wherein, $\boldsymbol{C_i} = \{\boldsymbol{d_{i1}}, \boldsymbol{d_{i2}}, \cdots, \boldsymbol{d_{im}}\}$ m represents the sample's number of $\boldsymbol{C_i}$
3. Take p% samples for each cluster class to obtain a new K cluster class sample set randomly, denoted as $\boldsymbol{\Omega}' = \left\{\boldsymbol{C_1}', \boldsymbol{C_2}', \cdots, \boldsymbol{C_K}'\right\}$, wherein $\boldsymbol{C_i}' = \{\boldsymbol{d_{i1}}, \boldsymbol{d_{i2}}, \cdots, \boldsymbol{d_{in}}\}, \boldsymbol{n} = \boldsymbol{m} * \boldsymbol{p}\%$
4. Combine the majority class sample set $\boldsymbol{\Omega}'$ and the minority class sample set $\boldsymbol{P_{min}}$ to synthesize the sample set $\boldsymbol{Q}$, S-T method is used for $\boldsymbol{Q}$ to obtain a balanced data set $\boldsymbol{Z}$.

3.2 ensemble model based on CRST

As can be seen from the visualization analysis results of the data in Section 2.1, the data of aortic dissection has high overlap. So the classification boundary is blurred, and it is necessary to construct a nonlinear classification model with strong generalization ability. On the basis of the CRST sampling method proposed in Section 2.2, integrated with the idea of Bagging[30] ensemble learning, CRST-Bagging ensemble learning algorithm is proposed in this section. It overcomes the limitation of a single classifier.

The CRST-Bagging algorithm is to generate a new sample set $B = \{Z_1, Z_2, \cdots, Z_T\}$ by using CRST sampling method iteratively. Then each sub-sample set Z_i is used to construct a sub-classifier M_i separately. A complete ensemble model classifier can be obtained by integrating the results of the T sub-classifier. The integration rule used in the algorithm is Majority Vote rule[31]. For the classifier, if P_{i1} is greater than or equal to P_{i2}, then R_1 gets one vote; if P_{i1} is less than P_{i2}, then R_2 gets one vote. R_1 and R_2 represents the sample category. This rule can be expressed by formula (1)(2). After the construction of the classifier, we send the verification set to the classifier for verification to evaluate the effect of the model. The model structure is shown in Figure 7.

$$R_1 = \sum_{i=1}^{T} f(P_{i1}, P_{i2}), R_2 = \sum_{i=1}^{T} f(P_{i2}, P_{i1}), where \quad f(x,y) = \begin{cases} 0, x < y \\ 1, x > y \end{cases} \tag{1}$$

$$C = \underset{j}{argmax} R_j \tag{2}$$

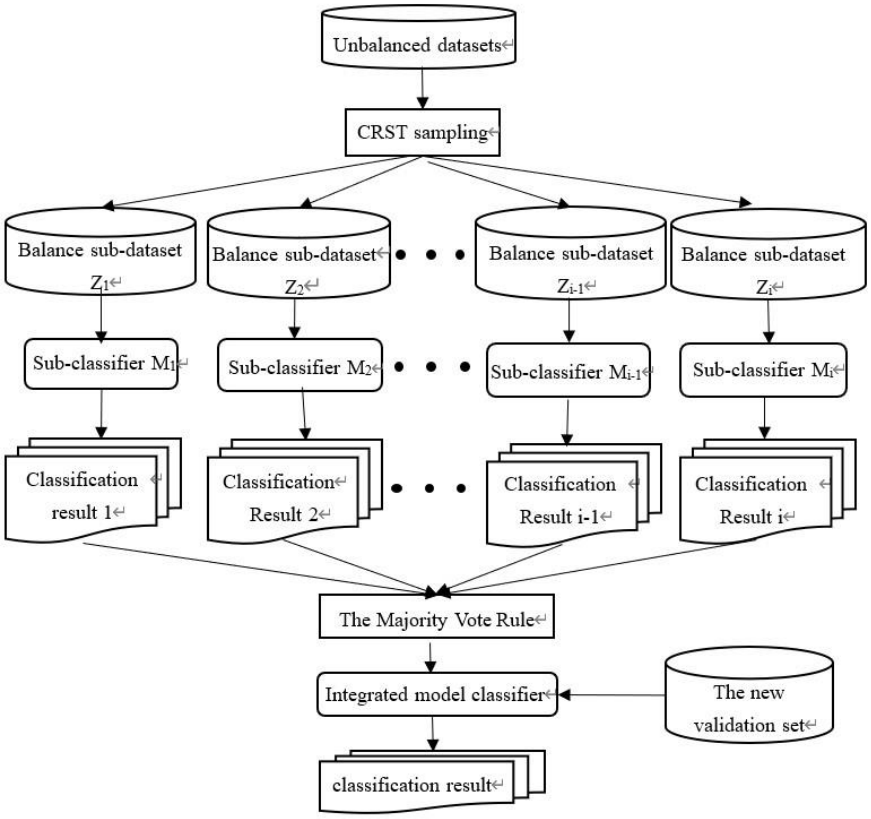


Figure 7. CRST-Bagging algorithm structure diagram

4. Experiment

In this section, the CRST sampling method we proposed is compared with the classical sampling methods Smote and Smote + Tomek-link, to verify its effectiveness on the AD data set. And then, on this basis, we compare the classification performance of CRST-Bagging model, RUSBoost[32] and SMOTEBagging[33] on AD data sets. RUSBoost and SMOTEBagging are ensemble learning algorithm for imbalanced data sets. The effects and advantages of CRST-Bagging method are analyzed.

4.1 Validity experiment of the CRST sampling method

4.1.1 Experiment settings

We apply Smote, Smote + Tomek-link, CCST and CRST sampling methods to the AD data set, and use the seven-fold cross-validation method to measure the sampling performance. In order to ensure the consistency of the experiment, XGboost is used as the classifier for this comparative experiment. The specific practices of the three methods are as follows:

- **Smote**, and **Smote + Tomek-link(S-T)**: At first, the samples in majority class are randomly sampled so that the ratio of majority to minority is 2:1. Then Smote and S-T are carried out respectively so that the ratio of positive and negative samples after resampling is 1:1. The number of samples in each category is 1604.

- **CCST:** CCST method refers to the clustering of most class training sample sets through K-means++ algorithm. Then, N sample points closest to the center of each cluster were selected. Finally, SMOTE + Tomek-Link method was applied to balance the data. In this experiment, we set K=802, N=2. After under-sampling, the number of majority samples is 1604.
- **CRST**: The number of clusters K is obtained by super-parameter selection, which is 30. Randomly select 3.1% of each cluster class for under sampling, and then apply S-T to the samples in minority class. The number of samples in each category is 1624.

4.1.2 Experimental results and analysis

Table 2. Experimental results for the Smote, S-T, CCST, CRST methods

Method	Precision	Recall	F1
Smote	0.789	0.711	0.748
S-T	**0.793**	0.723	0.749
CCST	0.778	0.765	0.771
CRST	0.782	**0.774**	**0.778**

The experimental results of the seven-fold cross-validation on the original dataset (53213,67) are shown in Table 2. Compared with the single over-sampling method Smote and S-T method, CCST and our proposed method CRST has a great improvement in recall rate and F1 value. As CCST selects the samples which is closest to the sample center of each cluster, the sample distance between different clusters after sampling is too far and the sample distribution is uneven, so the effect is inferior to that of CRST. By contrast, the CRST method performed better. It shows that CRST can reduce the occurrence of missed diagnosis in patients with AD to some extent. This is because the CRST is an over-and-under sampling. While clustering and under-sampling the majority samples, the CRST method uses the S-T method to generate the same amount of minority class samples. In this way, we can not only retain the original distribution of majority samples and select representative sample points, but also balance the number of minority samples through over-sampling.

4.2 CRST-Bagging Model Effect Comparison Experiment

4.2.1 Experimental settings

The seven-fold cross-validation method is used in the comparative experiments. The first group of experiments was performed in the original data set (53213, 67). The second group experiment was tested on the test set. The experimental details of the various algorithms of this experiment are as follows:

- **RUSBoost**: the base learner type is decision tree C4.5, the number is 100, the depth is 5;
- **SMOTEBagging**: Set the number of clusters K = 5, the base learner is decision tree C4.5, the number is 100, the depth is 6;
- **CRST-Bagging**: Set the number of clusters K = 50, p%=3.1%.

4.2.2 Experimental results and analysis

The experimental results of the seven-fold cross-validation on the original dataset (53213,67) are shown in Table 3. Figure 8 shows the ROC curve of the algorithm. It can be seen that CRST-Bagging performs best on the AD dataset, SMOTEBagging algorithm is second, and the RUSBoost algorithm performs worst. Compared with RUSBoost and SMOTEBagging methods, CRST-Bagging improved classification performance, and significantly improved accuracy and F1 value. It indicates that the rate of missed diagnosis and misdiagnosis in patients with AD are significantly reduced by CRST-bagging.

Table 3. Average experimental results of RUSBoost, SMOTE Bagging, CRST-Bagging on the original data set of seven-fold cross-validation

Method	**Precision**	**Recall**	**F1**
RUSBoost	0.774	0.751	0.762
SMOTEBagging	0.791	0.780	0.785
CRST-Bagging	**0.841**	**0.783**	**0.811**

Table 4. The average experimental results of the RUSBoost, SMOTEBagging, CRST-Bagging methods on the test sample set

Method	**Precision**	**Ecall R**	**F1**
RUSBoost	0.787	0.760	0.773
SMOTEBagging	**0.842**	0.772	0.805
CRST-Bagging	0.836	**0.807**	**0.821**

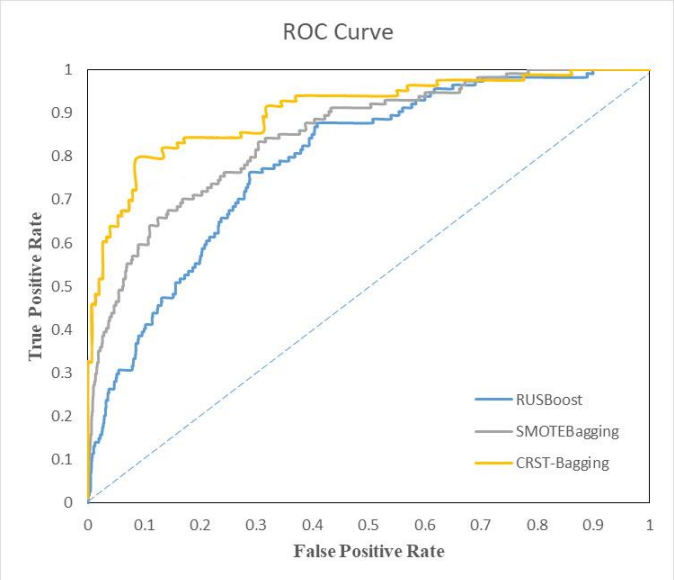


Figure 8. ROC graph of RUSBoost, SMOTE Bagging, CRST-Bagging on the original sample set

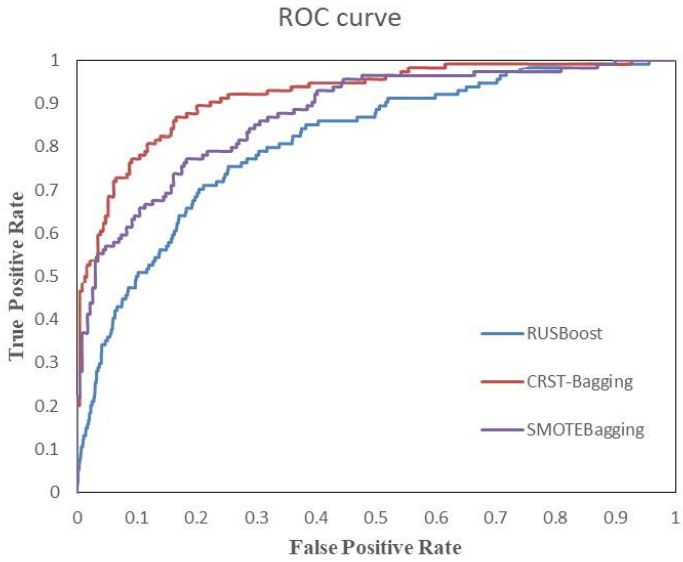


Figure 9. ROC graph of RUSBoost, SMOTEBagging, CRST-Bagging on the test sample set

To test the generalization ability of the model, we performed the prediction of AD on the test set (235, 67). The experimental results of the three algorithms are shown in Table 4. Figure 9 shows the ROC curve of the algorithm. As can be found from the table 4 and figure 9, the SMOTEBagging algorithm has the highest accuracy on the test sample set, and the CRST-Bagging algorithm has the highest recall rate and F1 value. CRST-Bagging's algorithm performance is significantly improved compared to RUSBoost and SMOTEBagging algorithms on the test sample set. CRST-Bagging algorithm has stronger generalization ability. In other words, CRST-bagging algorithm is more likely to detect potential patients with AD.

5. Discussion

AD is a rare and high-risk cardiovascular disease. Its complex clinical manifestations and various atypical symptoms lead to serious misdiagnosis and missed diagnosis. The rarity of the disease also leads to a significant imbalance in the data set. This paper studies the misdiagnosis and missed diagnosis of AD and the imbalanced characteristics of data.

For the original AD dataset, in the data preprocessing stage, we performed missing value processing, feature screening, data standardization. And at the level of data distribution the data is understood through dimensional reduction visualization. This is different from the general "black box" approach of machine learning algorithms. These methods enable us to have a priori knowledge of the distribution of actual medical data sets. This prior knowledge can inspire clinical medicine to explore the etiology and diagnostic criteria of AD.

Aiming at the high imbalance of AD data set, this paper proposes a resampling method CRST based on clustering. This method combines the advantages of traditional sampling methods Smote + Tomek-link and clustering algorithm K-means++. In CRST, a certain percentage of samples are randomly selected from

clusters, which not only makes the selected samples effectively represent the characteristics of most kinds of samples, but also ensures the randomness of sampling. Experiments show that CRST scientifically and effectively reduces the imbalanced ratio of rare disease medical data and relieves the obstacles that imbalanced data bring to the construction of classification models.

On this basis of CRST, this paper proposes the CRST-Bagging learning model combined with the idea of ensemble learning. After experimental comparative analysis, the CRST-Bagging model presented in this paper shows excellent performance on the AD data set. Not only the accuracy and recall rate of the model on the original AD data set have been improved, but also the generalization ability of the model on the test sample set is also very good. This shows that this model is a good diagnostic model of AD. Clinically, the misdiagnosis rate of AD is close to 40%[3][4][5]. The model performance data show that this algorithm can not only reduce the workload of doctors, but also reduce the misdiagnosis rate of AD to save patients' lives effectively.

6. Conclusions

In this paper, a cluster-based ensemble learning model named CRST-Bagging is proposed to assist in diagnosing aortic dissection through the patient's inspection results. Compared with the ordinary classification model, our model pays more attention to the processing of medical data sets with high imbalance ratio. While ensuring high accuracy, CRST effectively improves the recall rate. That is, the missed diagnosis and misdiagnosis rate is reduced. In addition, the algorithm demonstrates a strong ease of use. In many basic hospitals where the equipment is not advanced enough, it is difficult for patients to perform more examination items such as CT, magnetic resonance angiography (MRA), etc. The model proposed in this paper can reduce the burden of doctors and patients to a certain extent and help diagnose the AD.

The diagnosis of AD remains one of the most difficult problems in the cardiovascular field. In the future, based on the analysis results of aortic dissection data and the proposed auxiliary diagnostic method for aortic dissection in this paper, we will study the pathological mechanism and key diagnostic indicators of aortic dissection from the perspective of interpretability, and explore whether there is a more definite clinical diagnostic method for aortic dissection.

Abbreviations

AD: aortic dissection

LSTM: long short-term memory

PAVE: Pattern Attention model with Value Embedding

CT: Computed Tomography

T-SNE: t-distributed stochastic neighbor embedding

CRST-Bagging: Cluster Random Under-sampling Smote-Tomek-link Bagging

CRST: Cluster Random Under-Sampling Smote + Tomek-link Approach
S-T: Smote + Tomek-link
CCST: Cluster-Center Under-Sampling and Smote+Tomeklink Approach

Declarations

Ethics approval and consent to participate

Ethical approval for this study was obtained from the Ethics Board of Xiangya Hospital, Central South University (201502042).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

Data are provided by Xiangya Hospital and it cannot be shared with other research groups without necessary permission. The data used during the current study is available from the corresponding author on reasonable request. The data description supporting the conclusions of this article is included in the article (and its Appendix).

Funding

The study was financially supported by National Natural Science Foundation of China (No. 61502537), and Strategic Emerging Industry Technological Research and Major Technological Achievement Transformation Project, High-tech Development and Industrialization Office (No. 2019GK4013). The funding body had no role in design of the study, collection, analysis, and interpretation of data or in writing the manuscript.

Authors' contributions

All of the authors had full access to all of the data in the study and take responsibility for the content of the manuscript. YG designed the model and experiment implementation. MW and LJZ wrote the code. LJL, GGZ and JML contributed to data collection and feature selection. YG, LJL and GGZ perform the results analysis. MW and LJZ drafted the initial manuscript. YG revised the manuscript. All authors read and approved the final draft of the manuscript for publication.

Acknowledgements

Not applicable.

References

[1] Erbel R, Aboyans V, Boileau C, et al. 2014 ESC guidelines on the diagnosis and treatment of aortic diseases: document covering acute and chronic aortic diseases of the thoracic and

abdominal aorta of the adult. The task force for the diagnosis and treatment of aortic diseases of the European Society of Cardiology (ESC). Eur Heart J 2014;35:2873-926.
[2] Mussa FF, Horton JD, Moridzadeh R, Nicholson J, Trimarchi S, Eagle KA. Acute Aortic Dissection and Intramural Hematoma: A Systematic Review. JAMA. 2016;316:754-763.
[3] Marroush TS, Boshara AR, Parvataneni KC, Takla R, Mesiha NA. Painless Aortic Dissection. Am J Med Sci. 2017;354:513-520.
[4] Yin XB, Wang XK, Xu S, He CY. Type A aortic dissection developed after type B dissection with the presentation of shoulder pain: A case report. World J Clin Cases 2021; 9(1): 232-235 [PMID: 33511190 DOI: 10.12998/wjcc.v9.i1.232]
[5] Ahmed, T., et al., Two Intriguing Cases of Stanford Type A Acute Aortic Dissection. Cureus, 2020. **12**(2).
[6] Zhengxing Huang, Tak-Ming Chan, Wei Dong. MACE prediction of acute coronary syndrome via boosted resampling classification using electronic medical records. 2017,
[7] Zhou, PY., Wong, A.K.C. Explanation and prediction of clinical data with imbalanced class di stribution based on pattern discovery and disentanglement. BMC Med Inform Decis Mak 21, 16 (2021).
[8] Zhigang Song, Shuangmei Zou, Weixun Zhou, et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. 2020, 11(1):7-34.
[9] Kamal, S.A., Yin, C., Qian, B. et al. An interpretable risk prediction model for healthcare with pattern attention. BMC Med Inform Decis Mak 20, 307 (2020).
[10] Seo, S.W., Kim, J., Son, J. et al. Evaluation of conditional treatment effects of adjuvant treatments on patients with synovial sarcoma using Bayesian subgroup analysis. BMC Med Inform Decis Mak 20, 320 (2020).
[11] Yu, X., Ma, N., Yang, T. et al. A multi-level hypoglycemia early alarm system based on sequence pattern mining. BMC Med Inform Decis Mak 21, 22 (2021).
[12] Huo D, Kou B, Zhou Z, et al. A machine learning model to classify aortic dissection patients in the early diagnosis phase. Sci Rep 2019;9:2701.
[13] Dwivedi AK. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl 2018;29:685-93
[14] Guo,. Beheshti, R., Khan, Y.M. et al. Predicting cardiovascular health trajectories in time-series electronic health records with LSTM models. BMC Med Inform Decis Mak 21, 5 (2021).
[15] Junlong Cheng, Shengwei Tian, Long Yu, Xiang Ma, Yan Xing, A deep learning algorithm using contrast-enhanced computed tomography (CT) images for segmentation and rapid automatic detection of aortic dissection, Biomedical Signal Processing and Control, Volume 62, 2020, 102145, ISSN 1746-8094,
[16] JAPKOWICZ N. Supervised Versus Unsupervised Binary-Learning by Feedforward Neural Networks[M]. 2001.
[17] SHENG V S, LING C X B T-N C on A I. Thresholding for Making Classifiers Cost Sensitive[C]//National Conference on Artificial Intelligence. AAAI Press. 2006.
[18] MIRZA B, KOK S, LIN Z et al. Efficient Representation Learning for High-Dimensional Imbalance Data[M]. 2016.

[19] XIE Y N, YU L, GUAN G H et al. An Overlapping Cell Image Synthesis Method for Imbalance Data[J]. Analytical cellular pathology (Amsterdam), Hindawi, 2018, 2018(1): 1-12.
[20] HU J, YANG H, LYU M R et al. Online Nonlinear AUC Maximization for Imbalanced Data Sets[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(4): 882-895.
[21] SHENG C, HAIBO H, GARCIA E A. RAMOBoost: Ranked Minority Oversampling in Boosting[J]. IEEE Transactions on Neural Networks, 2010, 21(10): 1624-1642.
[22] CHEN T, GUESTRIN C B T-A S I C on K D & D M. XGBoost: A Scalable Tree Boosting System[C]. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2016.
[23] ZHANG D, QIAN L, MAO B et al. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGBoost[J]. IEEE Access, 2018, PP(99): 1.
[24] VAN DER MAATEN L, HINTON G. Visualizing Data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9: 2579–2605.
[25] CHAWLA N V, BOWYER K W, HALL L O et al. SMOTE : Synthetic Minority Over-sampling Technique[J], 2002, 16(January): 321–357.
[26] NGUYEN H M, COOPER E W, KAMEI K. Borderline over-sampling for imbalanced data classification[J]. International Journal of Knowledge Engineering and Soft Data Paradigms, 2011, 3(1): 4.
[27] LUSA L, OTHERS. SMOTE for high-dimensional class-imbalanced data[J]. BMC bioinformatics, 2013, 14(1): 106.
[28] TOMEK I. Two Modifications of CNN[J]. IEEE Transactions on Systems Man and Communications AMC-6, 1976: 769–772.
[29] BATISTA G E A P A, PRATI R C, MONARD M C. A study of the behavior of several methods for balancing machine learning training data[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 20.
[30] BREIMAN L. Bagging predictors[M]. 1996.
[31] KITTLER J, HATEF M, DUIN R P W et al. On Combining Classifiers[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1998, 20(3): 226-239.
[32] SEIFFERT C, KHOSHGOFTAAR T M, VAN HULSE J et al. RUSBoost: A hybrid approach to alleviating class imbalance[J]. IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans, 2010, 40(1): 185-197.
[33] WANG S, YAO X. Diversity analysis on unbalanced data sets by using ensemble models[J]. IEEE Symposium on Computational Intelligence&Data Mining, 2009, 1(5):324-331.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx