EasyChair Preprint

№ 4040

Classification of Cancer Subtypes Based on Imbalanced Data Sets

Yimin Fan, Lin Qi and Yun Tie

August 16, 2020

Classification of cancer subtypes based on imbalanced data sets

1st Yimin Fan	2st Lin Qi	3st Yun Tie
School of Information Engineering	School of Information Engineering	School of Information Engineering
Zhengzhou University	Zhengzhou University	Zhengzhou University
Zhengzhou, China	Zhengzhou, China	Zhengzhou, China
641170669@qq.com	ielqi@zzu.edu.cn	ieytie@zzu.edu.cn

Abstract. Cancer is an important factor affecting human health. Many cancers contain different subtypes and have high complexities. Different subtypes have different mechanisms of occurrence, so the correct classification of cancer subtypes is essential for early diagnosis and preventive treatment. With the development of high-throughput technologies, the Cancer Genome Atlas (TCGA) project has been continuously improved to provide comprehensive cancer genome data. However, many of these cancer data have the characteristics of unbalanced sample distribution, high data feature dimensions, and many redundancy, which will affect the classification effectiveness of a few classes, thereby affecting the overall classification performance. In this paper, for the DNA methylation data of liver cancer, breast cancer, gastric cancer and three types of cancer, a model based on balanced feedback sampling and Tomek link is used. First, the balanced feedback sampling algorithm is used to sample the different subtypes, and then the Tomek Link is used to clean up the data and eliminate noise to obtain the optimal sample data. Use the equally divided Lasso algorithm for feature selection, remove redundant features, and avoid overfitting. Finally, the support vector machine, random forest and convolutional neural network are used to classify, and four commonly used classification performance evaluation indicators are used to verify the effect of the balancing method. Three sets of cancer data were classified by subtype, and the best classification effect was obtained on the gcForest model.

Keywords. DNA methylation; imbalance; multiple classification; cancer subtype classification

1 Introduction

With the continuous development and progress of modern society, the living standards of human beings have gradually improved, and the material and technological life has also become richer. The changes in life and production methods provide many conveniences for everyone's life, but they also cause certain harms, such as environmental pollution and unreasonable diet. Some cancers caused by these factors have seriously threatened human health. With the increase in the world's population and the aging population, the global burden of cancer continues to increase, and nearly half of the patients will die as a result. Cancer is caused by abnormalities in the mechanisms that control the growth and division of normal cells. Cancers can be divided into various types according to their origin and nature: lung cancer, breast cancer, stomach cancer, liver cancer, etc. Different cancers are divided into a variety of cancer subtypes due to different molecular mechanisms [1]. Identifying cancer subtypes is an important branch of personalized medicine [2] research. The determination of subtypes can increase the chance of patients getting the best treatment. It also provides guidance for the analysis of cancer mechanisms and the

development of drug targets, while also laying the foundation for precision medicine and personalized treatment. It also provides guidance for the analysis of cancer mechanisms and the development of drug targets, while also laying the foundation for precision medicine and personalized treatment.

Cancer has a complex pathogenesis. Cancer cells are the source of cancer and have a variety of characteristics, including cell-level morphological characteristics and molecular-level expression characteristics. Cancer subtypes are specific types of cancer based on certain characteristics of cancer cells. The identification of cancer subtypes is the key to precise and personalized treatment. Therefore, it is very important to choose appropriate methods and models to accurately distinguish different molecular subtypes. At present, the research steps of the existing cancer classification model are mainly: first, preprocessing gene expression data, clinical information data, etc., then performing feature selection, and finally entering the classification model for training and testing. In the process of model learning, the structure and parameters of each part are constantly improved, so as to establish a model with high stability and generalization. The basic flow of cancer subtype classification is shown in Figure 1.

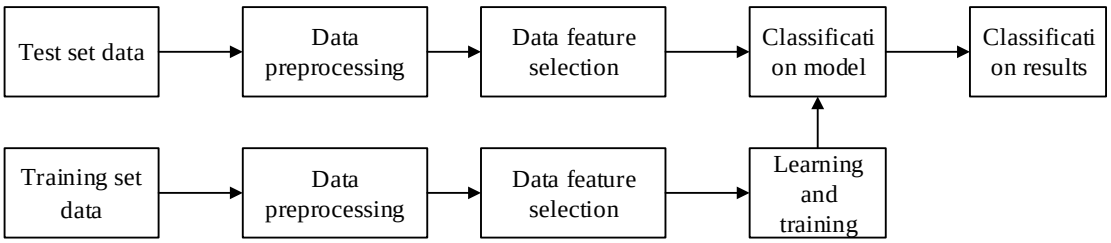


Figure 1. Basic flow chart of cancer subtype classification

The data mining process is affected by many factors, and the data obtained usually cannot reach the ideal state, and many data sets have imbalance problems. This phenomenon will lead to a large difference between the classification accuracy of the majority class and the minority class, thus affecting the overall performance of the classification model [3]. Data imbalance is a common phenomenon in biomedical databases. In the study of classification problems, when the number of samples in each category has a large gap, the recognition rate of the classification model for the minority class decreases, and the model cannot learn effective information, which makes the classification results biased and affects the performance of the classification model. The high accuracy of medical research is very important. Therefore, in order to ensure the accuracy of classification, data balancing has become an important research content. At present, there are many theoretical studies and experimental analyses on the data imbalance phenomenon of the binary classification problem. Compared with binary classification, multi-classification problems are more complex and diverse, and how to solve the problem of multi-class data imbalance still needs further research. Therefore, this paper proposes a sampling method that feeds back the information of category balance [4] to balance the samples before feature selection, so that the distribution of samples of different subtypes is relatively balanced to improve the accuracy of multi-classification.

2 Literature Survey

In the recent years, biomedicine, genetic informatics and computer technology have intermingled with each other, and more and more researches have been carried out on the classification and prediction of

medical data through machine learning and neural network methods. A lot of research has been done on the classification of cancer imbalance data in domestic and foreign studies. The following are several related studies in this area.

Literature [5] uses machine learning techniques to predict drugs that interact with adenosine receptors. Because the number of drugs interacting with adenosine molecules differs greatly from the number of non-interacting drugs, the data set is unbalanced, which makes the classification model less sensitive. For this problem, the SMOTE method is used in this paper to make the number of the two drugs almost equal, so that the data is balanced. Finally, Random Forest (RF), Decision Tree (DT) and Support Vector Machine (SVM) are used to classify the balanced data set. Among them, the accuracy of RF is the highest, reaching 75.09%.

Literature [6] proposed a pretreatment method for the imbalance of data in three common diseases such as diabetes, spondylosis, and Parkinson's disease. This method combines the SMOTE method with the data cleansing technology Tomek to further identify the boundaries between different categories after generating new samples. In this paper, eight classifiers are used for classification experiments and compared with the classification results that only use SMOTE to balance the data. The results show that the SMOTE balancing method combined with Tomek is better than the classification using SMOTE alone. It can solve the problem of medical data imbalance more effectively.

Professor Dong [4] summarized data balancing methods based on data sets and based on integrated algorithms to solve the problem of unbalanced data affecting the performance of classification models. Aiming at the imbalance in the multi-classification problem, a sampling method using balance to provide category information is proposed to complete the preprocessing of experimental data. Finally, random forest is used to classify the balanced data. Experimental results prove that the data balancing method in this paper can improve the performance of the classification model on unbalanced data sets.

At present, most of the data balancing methods are aimed at binary classification problems. For the imbalance phenomenon in the multi-classification problem of cancer subtypes, this paper proposes an effective data balancing method.

3 Methods

3.1 Data balancing method

The method of data balancing is mainly used to avoid the interference of unbalanced data sets to subsequent research. For the imbalance problem, the existing solutions mainly include two types:

(1) Balanced method based on data set [7]: By simply randomly copying or deleting samples or generating samples by some rules, the balance can be achieved by directly changing the data distribution. It mainly includes random oversampling, random under sampling, synthetic sampling method SMOTE, etc. Although these methods can satisfy the sample distribution balance in number, they will miss key features or bring a lot of noisy data, etc., and cannot fundamentally solve the problems caused by data imbalance.

(2) Algorithm-based balancing method [8]: Starting with the algorithm, the sampling method is adjusted appropriately to adapt to the unbalanced data set during sample extraction, thereby reducing the impact of unbalanced data, such as: Bagging method, Boosting Methods, cost-sensitive methods, etc. These methods need to be completed in accordance with the specific research content and data types.

After weighing the advantages and disadvantages of the two types of data balancing methods, as well as the characteristics of small sample size, many sample types, and strong complexity of cancer data, this

paper chooses the first balancing method. Referring to the sample generation idea of the SMOTE [5][6] method in the binary classification problem, the feedback information of the data balance is introduced to realize the expansion of the minority sample, and then the sample screening is completed in combination with the Tomek link, and finally the multi-category data is effectively balance treatment.

Data balance

The imbalance in the binary classification problem is usually obvious, and the degree of data balance is also relatively intuitive. You can directly use the ratio of the majority class to the minority class sample as an indicator to measure balance. For multi-classification problems, you need to calculate the standard deviation to obtain the balance of multi-classification, as shown in formula (1).

$$B(S) = -\sqrt{\frac{\sum_{n=1}^{N}\left(c_n - \overline{c}\right)^2}{N}} \tag{1}$$

$$\overline{c} = \frac{\sum_{n=1}^{N} c_n}{N} \tag{2}$$

The formula $B(S)$ is the balance function; S represents the data set; N is the number of samples in the category; c_n is the number of samples in the n category; $\overline{c}$ is the average of the sample size of different categories, as shown in formula (2). When B is larger, the standard deviation of the number of categories is smaller. The more stable sample data set, the more balanced the data, so the data set can be balanced by maximizing the balance function.

Balanced method based on balance feedback sampling

Aiming at the imbalance in the multi-classification problem, this section uses the sample balance as a reference to construct a method for balancing cancer samples based on balance feedback sampling. Based on the idea of information gain [9], the algorithm refers to the balance value during random extraction, so that the generated samples meet the goal of maximizing the balance. The class with the highest balance gain in formula (3) is listed as the class where new samples will be constructed to gradually improve the balance.

$$Gain(\boldsymbol{S}, y) = B(\boldsymbol{S} \cup \{(\hat{x}, y)\}) - B(\boldsymbol{S}) \tag{3}$$

In the formula $(\hat{x}, y)$ is the constructed new sample of category y .Considering that the initial $B(S)$ of each category is the same, the category that maximizes the balance value is obtained by formula (4).

$$argmaxB(\boldsymbol{T}, N) = argmax_{y\in\{1,\ldots,N\}} B(\boldsymbol{T} \cup \{(\hat{x}, y)\}) \tag{4}$$

In the formula, T is the balanced data set. The overall steps of the feedback sampling algorithm based on balance are as follows:

(1) Input the original sample data set, and calculate the degree of balance B by formula (1).

(2) Set a balance standard according to the research content, and judge whether the balance degree reaches the balance requirement;

(3) If the requirements are not met, the balance gain is calculated to obtain the category n with the largest gain.

(4) Find the sample set Xn of category n in the overall data set and calculate the center $\overline{x}$ of Xn.

(5) Randomly select one of the samples x in Xn, and perform linear random interpolation between x and the sample center $\overline{x}$, and the new sample of category n is constructed.

(6) Add new samples to the data set, then loop through the steps of (2), (3), (4), (5), and continuously construct new samples until the expected balance effect is achieved.

Tomek link method

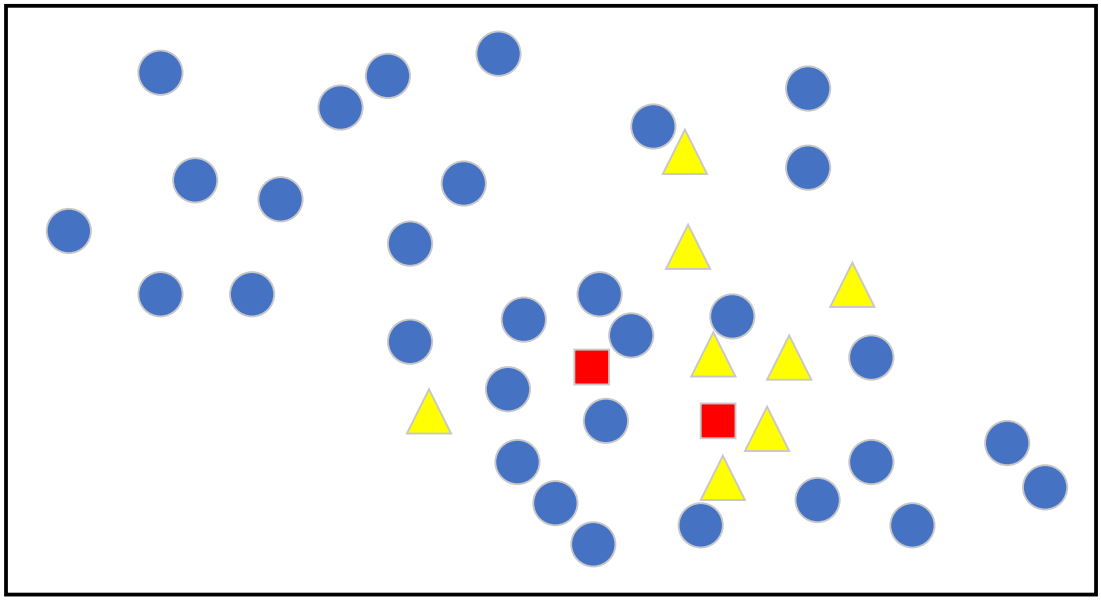
Due to the large sample complexity of the multi-classification problem, there will often be some noise and redundancy in the newly expanded sample, so the Tomek link algorithm [10] is introduced as a sample quality discrimination mechanism to complete the sample cleaning. The basic principle of the algorithm is: For a pair of Tomek link sample points $x1$ and $x2$, if there is no other point $x3$ so that formulas (5) and (6) hold, then one of $x1$, $x2$ is noise. Or all data are on the category boundary. Traverse the sample data, discard the sample points that meet the above conditions.

$$d\left(x1,x2\right) > d\left(x1,x3\right) \tag{5}$$

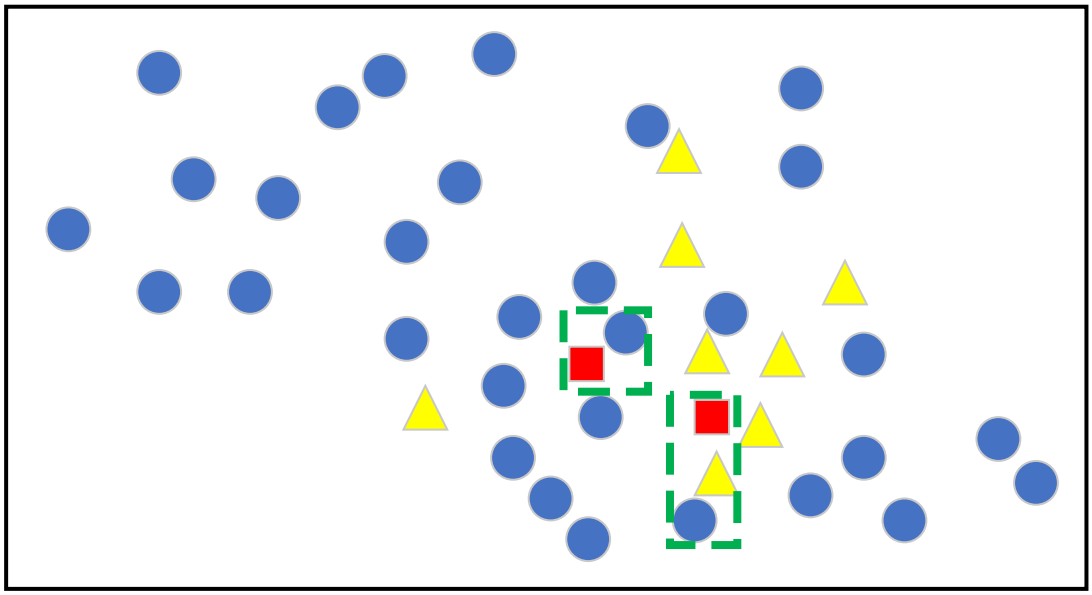
$$d\left(x1,x2\right) > d\left(x2,x3\right) \tag{6}$$

Figure 2 is a schematic diagram of the Tomek link algorithm. By introducing the Tomek link algorithm to formulate the sampling point trade-off rules, remove the noise data and boundary data, ensure the quality of the generated samples, and obtain the optimal sample set.

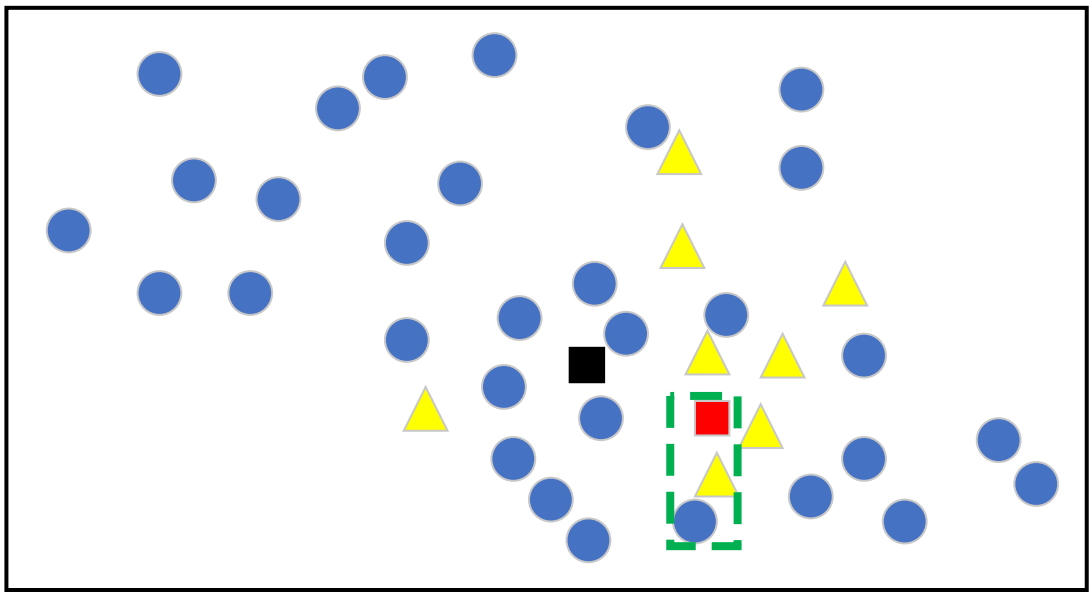
The core of the entire balancing process lies in three aspects: one is to introduce a feedback mechanism and a balance gain index, and adaptively determine the sample generation during the balancing process; The second is to avoid blindness of random sampling by continuously calculating the sample center of the new data set; The third is to judge the quality of the sample through Tomek link and decide the choice of sample points.



(a) Generate a new sample of the red box



(b) The point closest to the new sample



(c) Remove noise and boundaries

Figure 2. Schematic diagram of Tomek link principle

3.2 Feature selection method

The data features in medical research generally have thousands or even tens of thousands, and the number of features is much larger than the number of samples. This kind of high-dimensional data is prone to cause the problem of "dimensional disaster", and it is prone to overfitting during model training, which affects the performance of the model. Therefore, before the classification, the feature selection method is used to remove noise, redundancy, and irrelevant data to obtain a simplified set of features, reduce the complexity of the classification process, and improve the performance of the classification model. The DNA methylation data of cancer samples obtained in this paper has high-dimensional characteristics, and there is more noise and redundancy. It is necessary to effectively filter the data features, improve the performance and efficiency of the classification model, and enhance the model generalization ability. Considering the large difference between the sample size and the feature amount, the traditional Lasso algorithm requires a large amount of calculation and is prone to overfitting. Therefore, this paper uses the K-Lasso algorithm: an even-split Lasso algorithm, which balances the number of features and the number of samples by means of feature sharing, reducing the amount of calculation for each feature selection. The algorithm has strong selectivity, good stability and easy implementation. When selecting features, it can fully eliminate redundant features and unrelated attributes in the medical data set and identify features related to specific diseases. Therefore, it is often applied to medical small sample data.

Lasso algorithm

Lasso (Least absolute shrinkage and selection operator), also known as the minimum absolute shrinkage and selection operator [11], is a stable and efficient Embedded feature selection algorithm proposed by Tibshirani in 1996. By introducing L1 regular items, the data is sparse, leaving a subset of features with strong correlation, which achieves the purpose of data dimensionality reduction. The principle of the algorithm is to reduce the feature coefficients by constructing a penalty function, so that the sum of the

absolute values of the regression coefficients is less than a constant, and at the same time minimize the sum of the squared residuals, and then generate some regression coefficient values that are zero. The Lasso estimates of these regression coefficients are defined by formula (7):

$$\arg\min\left\{\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j)^2\right\} \tag{7}$$

In the formula β_j is the regression coefficient of the first variable, and the constraints are as shown in formula (8):

$$s.t.\sum_{j=1}^{p}\left|\beta_j\right| \le s \tag{8}$$

$s \ge 0$ is a penalty constraint for β_j, and s can take values from 0 to infinity. When the value of s is large, the constraint condition has no constraint effect, and all features will not be deleted. When the value of s is small, the Lasso estimated value of the feature coefficient with a small correlation is compressed to 0, and the corresponding variable is deleted, so as to realize the feature filtering.

Implementation of KLasso feature selection

The even-split Lasso algorithm (K-Lasso) improves on the basis of the classic Lasso algorithm, splits the sample feature set into multiple subsets, and performs parallel selection and parallel matrix operation on the feature subsets to improve the efficiency of the algorithm. Improve feature quality and classification performance. The specific process of the algorithm is shown in Figure 3.3. First, all features are equally divided into K, and then Lasso feature selection is performed on each feature to obtain K feature subsets. These subsets are combined, and finally the combined feature subsets are combined. Make a Lasso selection to avoid redundancy and get the best combination of features. In view of the problem of the value of K in the model, this paper sets K to 5 after multiple classification tests, which appropriately narrows the difference between the cancer sample size and the characteristic amount at each screening.

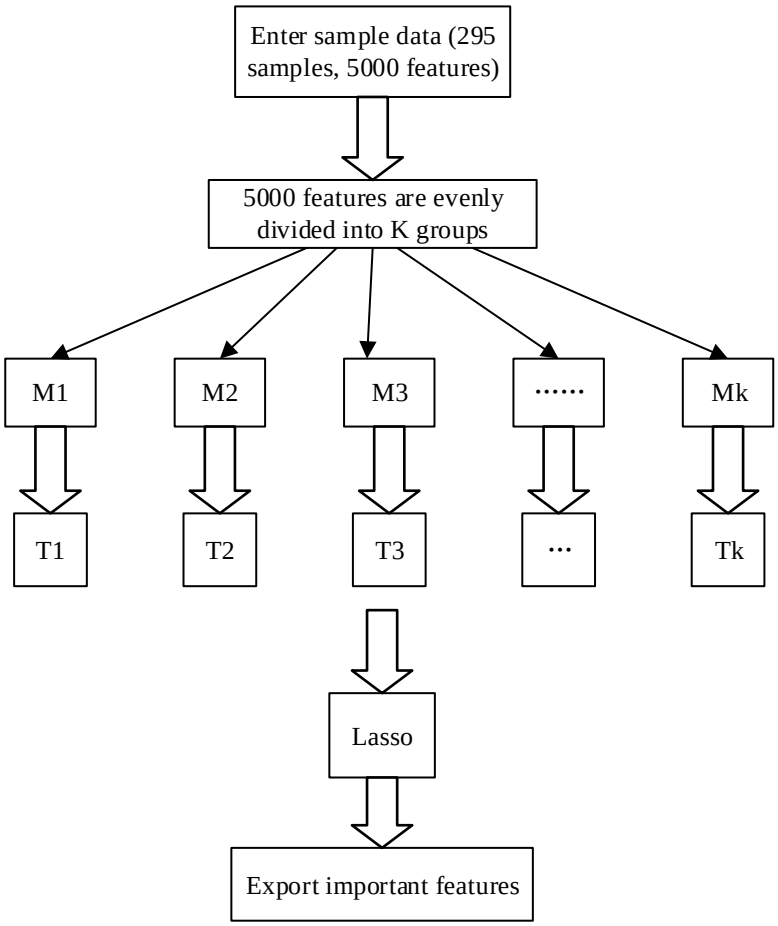


Figure 3. K-Lasso algorithm flow chart

3.3 Classification

The classification problem exists in many fields and is an important research content. The research of cancer classification model includes data processing, feature selection, classification model design, model performance evaluation and other links, and classification model design is its core link. There are many existing classification methods. Considering that the deep learning method has many parameters and the model training requires a large sample size, it is widely used in the research of two-dimensional images and videos. Cancer research is based on one-dimensional small sample data at the molecular level, and traditional machine learning methods are more suitable for such data.

Deep forest

The multi-Grained Cascade Forest (gcForest) algorithm is a deep forest structure proposed by Professor Zhou Zhihua of Nanjing University in 2017. It is a deep tree integration algorithm using cascade structure for representation learning. Compared with deep neural networks, this model is easier to adjust parameters and has strong scalability. The model training process does not require a large amount of data, and can be applied to small-scale data well [12]. The overall structure of gcForest is shown in Figure 2.3. The model includes two parts: Multi-Grained Scanning module and Cascade Forest module. The multi-granularity scanning module scans the input features through a sliding window to obtain a series of

feature vectors, and then sends them to a set of random forests and completely random forests for training, and finally combines the results of these forests as the input of the cascade forest module.

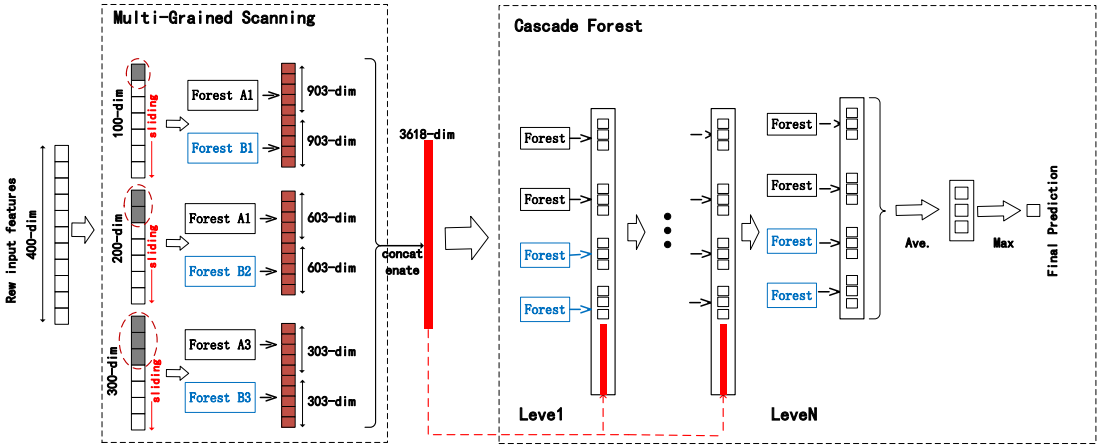


Figure 4. gcForest model structure diagram

The cascade forest module mainly expands the original features by adding new features. Multiple forests are integrated in the module, and each forest is composed of multiple subtrees. The cascade module of the model in Figure 2.3 is composed of random forests and completely random forests. These forests are the integration of decision trees. Suppose there are three prediction categories, each forest outputs a three-dimensional class vector, and the output of all forests in each layer is combined to connect it to the original input as the input of the next layer. The number of model cascade layers is adaptively determined according to data complexity [13]. During the training process, an overall performance evaluation is performed for each expanded layer. If the classification performance is not improved, the cascade layer automatically ends the expansion and the number of cascade layers is determined. If there is improvement, continue to input the class vector output from this layer with the original features to the next layer, and continue the training of the new layer.

3.4 Classification performance evaluation index

The classification performance evaluation index aims to analyze the effect of the model objectively and quantitatively. It is very important to select the appropriate index in the classification research. At present, there are many performance evaluation indexes for models, and different learning tasks correspond to different evaluation methods. For classification problems, the widely used indicators are: Accuracy, Precision, Recall [14], F1_score, etc. As shown in Table 1, the confusion matrix (Confusion Matrix) is the basis of the performance evaluation index of the binary classification problem.

Table 1. Confusion matrix

Confusion Matrix		Real	
		Positive	Negative
Predict	Positive	TP	FP

	Negative	FN	TN

TP (True Positive), predicted to be positive, but actually positive;

TN (True Negative), predicted to be negative, but actually negative;

FP (False Positive), predicted to be positive, but actually negative;

FN (False Negative), predicted to be negative, but actually positive.

According to the confusion matrix, calculate the performance evaluation indicators in formulas (9), (10), (11), and (12):

(1) Accuracy: The ratio of the number of accurately classified samples to the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

(2) Precision: Precision is also called the precision rate, and the proportion of the samples that are predicted to be positive is actually the positive.

$$\Pr ecision = \frac{TP}{TP + FP} \tag{10}$$

(3) Recall: Recall rate is also called recall rate, the proportion of all positive samples that are accurately predicted.

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \tag{11}$$

(4) F1_score: Evaluation index that comprehensively considers the precision rate and the recall rate.

$$F1 = 2 \times \frac{\Pr ecision \times \mathrm{Re}\, call}{\Pr ecision + \mathrm{Re}\, call} \tag{12}$$

4 Experiment and Analysis

4.1 The overall process of the experiment

The overall framework of cancer subtype classification research based on unbalanced data is shown in Figure 3.1. First, the DNA methylation data of the three cancers were obtained from the TCGA database and standardized. Then use the balance method based on the feedback feedback sampling combined with Tomek link proposed in this paper to balance the three sets of data. Then use K-Lasso algorithm for feature selection, and finally send it to the classification model for classification.

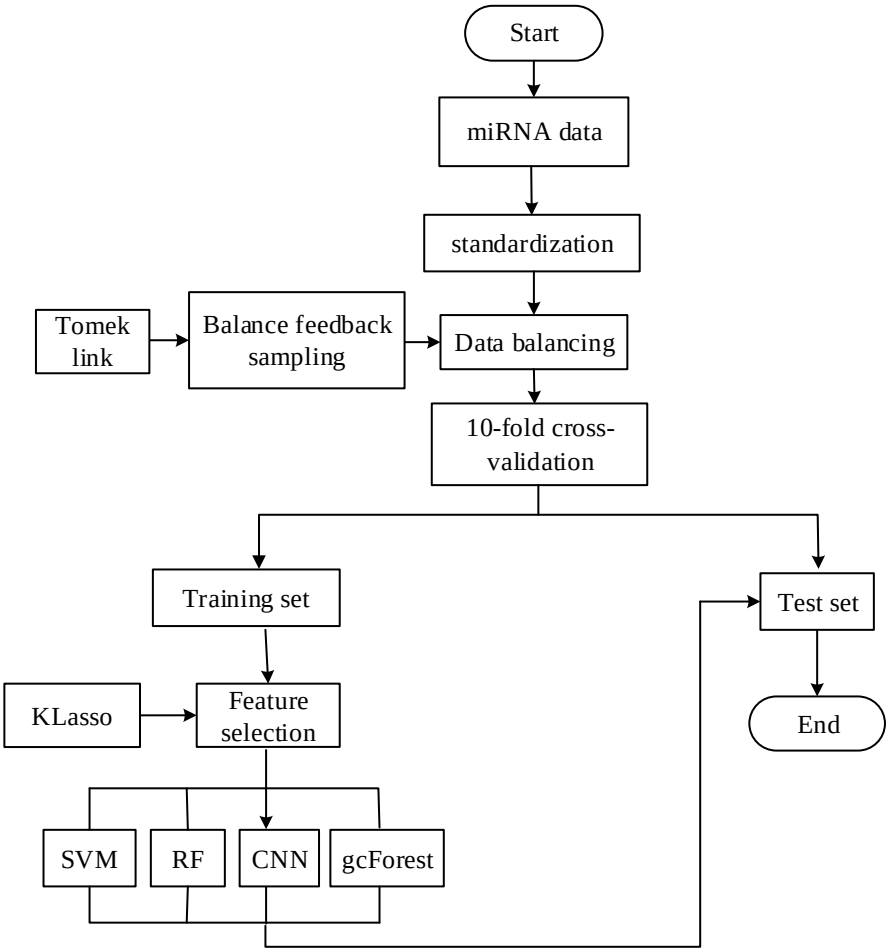


Figure 5. Overall framework

4.2 Experimental data

Three kinds of cancer data in the TCGA database were selected. Table 2 shows the details of the data. Lung cancer includes lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), with 159 and 36 cases. Breast cancer has four subtypes: Luminal A (LA) type, Luminal B (LB) type, HER2-rich (HER2-E) type and basal-like (BL) type 109, 47, 14 and 41 cases. Gastric cancer data includes four molecular subtypes: viral infection type (EBV), microsatellite unstable type (MSI), chromosomal unstable type (CIN), and genetically stable type (GS), with 26 cases, 64 cases, and 147 cases.

Table 2. Details of cancer data

Dataset	***Sample size***	***Number of classes***
Lung	195	2
Breast cancer	294	4
Gastric cancer	295	4

4.3 classification

In recent years, SVM [15] and RF [16] have been widely used and relatively effective in cancer classification research. The classic model of deep learning CNN [12] is often used in classification research. Therefore, these three models are selected as the comparative models of experiments.

(1) SVM

Implement a SVM classification model with Gaussian kernel function (RBF) in formula (13) as the core.

$$k(x, y) = \exp(-\gamma \| x - y \|^2) \quad (13)$$

(2) RF

Implement an RF classification model composed of decision trees, and vote according to the principle of "minority obeys the majority" to determine the subtype to which the sample belongs.

(3) CNN

A one-dimensional CNN model of convolutional neural network is constructed, which includes an input layer, two convolutional layers, two pooling layers, a fully connected layer and an output layer. Use the ReLU function as the activation function. The pooling layer selects the maximum pooling method.

4.4 Performance comparison

For the performance evaluation indicators of multi-classification problems, there are two calculation methods: micro and macro. The micro method is to split the K category problem into K binary classification problems to obtain K confusion matrices. Then add the corresponding TP, FP, TN, and FN values in the K confusion matrices to find the average, and finally calculate the corresponding performance index. The macro method is to split the multiple classifications into K binary classifications, and obtain K confusion matrices to find the accuracy, precision, and recall of each confusion matrix. Finally, the average value is calculated to obtain multi-class evaluation indicators. The classification of cancer subtypes studied in this paper is a multi-classification problem, and the macro method is used to calculate the model performance evaluation index.

5 discuss

The first part of the experiment is based on the DNA methylation data of cancer samples in TCGA. Table 3 is the experimental scheme for balanced comparison. The A1 scheme directly selects the data through the K-Lasso algorithm. The A2 scheme uses the mixed model based on feedback sampling and Tomek link proposed in this paper to balance multiple subtypes of data to obtain relatively balanced samples. After completing a series of data processing, the training set and test set are divided by 10-fold cross-validation, and the training set is sent to gcForest for learning. The resulting classification performance indicators are shown in Table 4.

Table 3. Balanced comparison experiment plan

Experimental program	Scheme details
A1	Unbalanced

A2	Balance

Table 4. Classification experiment results of the two schemes (%)

	Rec			*Pre*			*Acc*			*F1-score*		
	Lung	Breast	Gastric	Lung	Breast	Gastric	Lung	Breast	Gastric	Lung	Breast	Gastric
A1	88.7	90.4	90.9	90.7	92.6	93.7	90.0	91.3	92.8	89.7	91.5	92.2
A2	91.1	93.0	94.3	92.4	93.6	96.2	91.8	93.5	95.1	91.7	93.3	95.2

The second part of the experiment mainly verifies the performance of the classification model. Compare the classification effects of several common methods in medical research to select the optimal model. After the data is balanced, feature selection is performed, and then four classification models are sent to each for training and testing. Table 4 shows the experimental results of the four models, measured by four performance evaluation indicators. Figure 6 is a comparison chart of corresponding performance indicators.

Table 5. Classification experiment results of the four methods (%)

	Rec			*Pre*			*Acc*			*F1-score*		
	Lung	Breast	Gastric	Lung	Breast	Gastric	Lung	Breast	Gastric	Lung	Breast	Gastric
SVM	87.3	.88.2	89.0	87.9	88.5	88.7	87.5	88.3	88.2	87.6	88.3	88.8
RF	86.1	90.5	89.8	86.6	91.6	90.6	86.4	91.0	90.3	86.3	91.0	90.19
CNN	83.4	89.7	90.1	85.6	90.4	91.5	85.2	90.2	90.8	84.5	90.0	90.8
gcForest	91.1	93.0	94.3	92.4	93.6	96.2	91.8	93.5	95.1	91.7	93.3	95.2

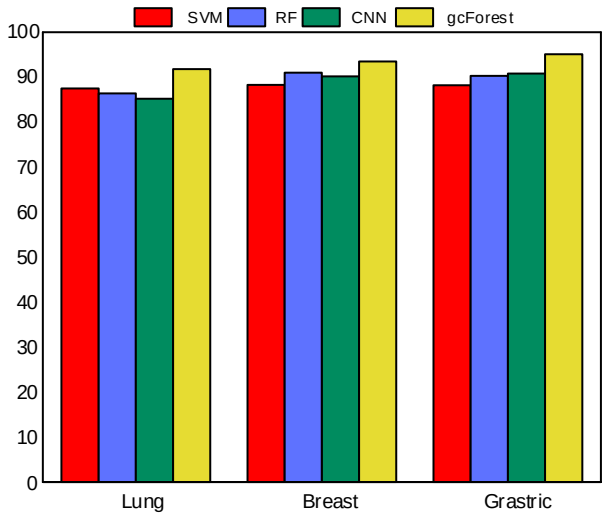


Fig. 6. The comparison of Accuracy shows the detail performance of different methods on three datasets. (%)

Observe and analyze Table 4 to see the effectiveness of the balancing method. The experimental results of A1 and A2 show that after balancing the distribution of sample subtypes in the data set, each classification performance index has been improved, and the model's ability to classify a small number of samples has been improved. It can be seen from Table 5 that the classification effect of the gcForest

model is significantly better than its SVM, RF, CNN, and the classification accuracy on gastric cancer data is the highest, reaching 95.09%

6 summary

This article mainly deals with the problems caused by the imbalance of multi-category data. Using a mixed model based on balanced feedback sampling and Tomek link, after balancing the three sets of cancer data, a classification experiment was conducted to verify the effect of the balancing method. The results show that after balancing the distribution of sample subtypes in the data set, the accuracy, precision, recall rate, and F1 classification performance indicators have improved. And achieved the best results on the gcForest model. Further research will optimize the classification algorithm to achieve higher accuracy.

References

1. Yang B, F., Liu S, S., Pang S, T., et al.: Deep Subspace Similarity Fusion for the Prediction of Cancer Subtypes. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, pp. 566-571. Spain (2018).
2. Waykole R N, F., Thakare A D, S.: Intelligent Classification of Clinically Actionable Genetic Mutations Based on Clinical Evidences. In: 2018 Fourth International Conference on Computing Communication Control and Automation, pp. 1-4. Pune (2018).
3. Zhu Rui, F., Guo Yiwen, S., Xue Jing-Hao, T.: Adjusting the imbalance ratio by the dimensionality of imbalanced data. Pattern Recognition Letters, 133(2020).
4. 董立岩,F.,王越群,S.,李永丽,T.,等.:基于最大平衡度的自适应随机抽样算法.东北大学学报(自然科学版)39(06),792-796(2018).
5. Saad A I, F., Omar Y M K, S., Maghraby F A, T.: Predicting Drug Interaction With Adenosine Receptors Using Machine Learning and SMOTE Techniques. IEEE Access, pp.7,146953-146963(2019).
6. Zeng M, F., Zou B, S., Wei F, T., et al.: Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data. In:2016 IEEE International Conference of Online Analysis and Computing Science. IEEE, pp. 225-228. Chongqing (2016).
7. Rustogi R, F., Prasad A, S., Swift Imbalance Data Classification using SMOTE and Extreme Learning Machine. In: 2019 International Conference on Computational Intelligence in Data Science. IEEE, pp.1-6. Chennai (2019).
8. Chen D, F., Wang X, S., Zhou C, T., et al.: The Distance-Based Balancing Ensemble Method for Data With a High Imbalance Ratio. IEEE Access 7,68940-68956(2019).
9. Wang H, F., Ge J, S., Zhang D, T., et al.: Sensor selection for target tracking based on single dimension information gain. The Journal of Engineering 2019(20),6562-6565(2019).
10. Rodolfo M. Pereira, F., Yandre M.G. Costa, S., Carlos N. Silla Jr. MLTL, T.: A multi-label approach for the Tomek Link undersampling algorithm. Neurocomputing 383,95-105(2020).
11. Febianti F, F., Pharmasetiawan B, S., Mutijarsa K. T.: Predictive System Based Multi-layered Clustering Model and Least Absolute Shrinkage and Selection Operator (LASSO). In: 2019 International Conference on Information and Communications Technology. IEEE, pp.371-376. Yogyakarta (2019).
12. Zhou Z H, F., Feng J, S.: Deep Forest: Towards An Alternative to Deep Neural Networks. In: IJCAI (2017).
13. Zhu Q, F., Pan M, S., Liu L, T., et al.: An Ensemble Feature Selection Method Based on Deep Forest for Microbiome-Wide Association Studies. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine. IEEE, Madrid, 248-253 (2018).

14. Xu J, F., Wu P, S., Chen Y, T., et al.: A Novel Deep Flexible Neural Forest Model for Classification of Cancer Subtypes Based on Gene Expression Data. IEEE Access 7,2086-22095(2019).
15. Karim M R, F., Wicaksono G, S., Costa I G, T., et al.: Prognostically Relevant Subtypes and Survival Prediction for Breast Cancer Based on Multimodal Genomics Data. IEEE Access 7,133850-133864(2019).
16. Aprilliani U, F., Rustam Z. S.: Osteoarthritis Disease Prediction Based on Random Forest. In: 2018 International Conference on Advanced Computer Science and Information Systems. Yogyakarta, 237-240(2018).