MACHINE LEARNING LAB2

Group:28
Author: Yi Li, Wenxuan Zhu, Xiaofei Li, Qitao Ye, JinShuai Chang

University of Nottingham

Abstract

This article will compare 4 methods for dealing with data imbalance, 8 different feature selection methods, and 4 different machine learning models to obtain the highest accuracy regression and classification models for the breast cancer dataset. In this experiment, accuracy, precision, and recall were used to evaluate PCR prediction accuracy, and MAE was used to evaluate RFS accuracy. The Random Forest classifier model was used to classify the breast cancer data set, and the overall accuracy and recall rate reached 90.7% and 97%, respectively. The ANN regressor model was used to perform regression prediction on the breast cancer data set, and the Mean absolute error was 21.8.

Index Terms — *classification, regression, accuracy, precision, recall.*

1. INTRODUCTION

Breast cancer is the most common cancer in women in the UK. Complete tumor elimination during surgery, known as pathological complete response (PCR), has a high potential for cure and longer recurrence-free survival (RFS) times. However, only 25% of patients receiving chemotherapy will achieve PCR, with the remaining 75% having residual disease and a range of outcomes. Better patient stratification and treatment could be achieved if information could be used to predict PCR and RFS prior to chemotherapy. The scope of our work is to use multiple methods to deal with imbalanced datasets, comparing numerous feature selection methods, classification models and regression models. Finally, the most accurate combination of PCR classification model and feature selection method, and the most accurate combination of RFS regression model and feature selection method are obtained.

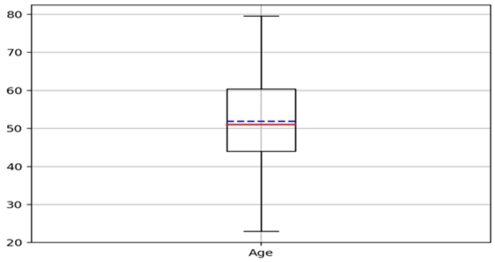
2. DATA PREPROCESSING

2.1. Missing data

To begin with, missing data, filled by value 999, are observed by searching for. It found that all missing data have 14 records for missing data, if there is data missing in the “pCR (result)” column, delete the corresponding row, and fill the missing data in other columns with the mode.

2.2. Outliers

For outliers, the experiment draws a boxplot about age to distinguish them. However, the picture shows all records are in the normal range.



2.3. Normalization

For normalization, the unit in features are various, to avoid nondimensionalization, these feature data are normalized by sklearn’s MinMaxScaler after selecting features.

3. DATA IMBALANCE PROCESSING

By observing the data set, we found that there is a relatively large difference in the number of positive and negative samples, with a ratio of about 3:1. As far as the model is concerned, the model constructed with unbalanced data will be more willing to be biased towards the labels of multi-category samples, and the actual application value is low. Therefore, to avoid the risks caused by data imbalance, we have found some common methods to solve data imbalance, including under-sampling and over-sampling methods. Due to the small number of total samples in this dataset, we finally decided to try to use over-sampled SMTOE, Borderline SMOTE， ADASYN and artificial completion to process the dataset.

3.1. SMTOE

SMOTE (Synthetic Minority Oversampling Technique) is an oversampling algorithm improved based on random sampling. SMOTE is simple to implement, but its disadvantages are also obvious. Since SMOTE treats all minority samples equally and does not consider the category information of neighboring samples, sample aliasing often occurs.

3.2. Borderline SMOTE

Borderline SMOTE is an improved oversampling algorithm based on SMOTE. This algorithm only uses the minority class samples on the border to synthesize new samples, thereby improving the class distribution of samples. The sampling process is to divide the minority class samples into three categories, namely Safe, Danger and Noise, as described below. Finally, only the minority class samples whose table is Danger are oversampled.

3.3. ADASYN (adaptive synthetic sampling)

ADASYN (adaptive synthetic sampling) is like Borderline SMOTE, assigning different weights to different minority samples, thereby generating different numbers of samples.

3.4. Artificial Completion

The category with less data in the unbalanced data set finally achieves classification balance through the method of replication, because the data obtained in this way are all real, and the corresponding labels are also trustworthy.

4. FEATURE SELECTION METHOD

DT and Random Forest performed better in classification, SelectFromModel performed better in regression.

4.1. DT

DecisionTreeClassifier model with different values of criterion using Gini and entropy with the balanced data, then get the accuracy with 80% for both train data and test data. So, I list the feature_importances_, and get the features great than 0.05.

4.2. RandomForest

Since DT achieves good expectations, RandomForest with different n_estimators from 100 to 110, I get the highest accuracy with 85%, more importantly, precision and recall are both more than 80. So, we decide to choose this method as our final option.

4.3. SelectFromModel

LinearSVC with L1 or L2 as the penalty is used as an estimator and ranking the score by calculating from the get_supoort method. Similarly, add a new possible feature adding the non-selected feature of cancer-related features, besides, adaBoostRegressor as estimator was a better choice for regression predict.

4.4. Other methods

4.4.1. variance selection

Through variance selection, 40 features with a threshold value of 0 were selected.48 features were combined by variance selection and cancer-related features.

4.4.2. Pearson correlation coefficient

Firstly, calculating the correlation coefficients, and deciding the threshold with the trying methods: max (median, min+(max-min)/2), and median, median+1. Finally, the other non-selected features of cancer-related features are added to all selected features

4.4.3. RFE

Different estimators have been used in RFE, such as logistic Regression and SVC with different kernels which are linear and RBF. Then set the n_features_to_select equal to 10,15,20. I also complement the above features with non-selected feathers of cancer-related features.

4.4.4. Manual select features

Just choose the first ten cancer-related features as selected features.

5. CLASSIFICATION

5.1. Description of Random Forest

Random Forest is a type of machine learning algorithm that can be used for classification and regression tasks. The Random Forest algorithm works by creating a large number of Decision Tree, each of which is trained on a random subset of the training data. The predictions made by the individual Decision Tree are combined to make the final prediction, which is typically more accurate than the predictions made by any individual tree.

One of the key advantages of Random Forest is that they are able to handle large and complex datasets, and they can also provide estimates of the importance of each feature in the dataset. This makes them a popular choice for many applications, including image and text classification, as well as predictive modeling. In addition, the Random Forest algorithm has good anti-noise ability and can run efficiently on large data sets. Especially in the classification of breast cancer data sets, it performs well, with high accuracy and fast training speed. Besides, there are not many parameters in the Random Forest, and it is easy to find the best parameters.

Although the Random Forest algorithm is fast enough, when the number of Decision Tree in the Random Forest is large, the space and time required for training will be large, which will cause the model to slow down. Therefore, in practical applications, if you encounter a situation with high real-time requirements, it is best to choose other algorithms.

As for other model algorithms, such as logistic regression, ANN and SVM, there are generally problems such as low accuracy and too many parameters to be adjusted in the kernel function, especially in the classification of breast cancer data sets, the performance is not ideal, so we use Random Forest for training.

5.2. Parameter of Random Forest

It is found that the optimal value of the n_estimators of the Random Forest parameter fluctuates around 100 from multiple experimental data. After multiple experiments and Matlab data analysis, the optimal value of n_estimators will appear between 100 and 110. Take the best value this time and use it as a parameter to train the Random Forest model.

5.3. Results

accuracy: 0.9079096045197741
precision: 0.8735099792453085
recall: 0.9705173858399664

Explanation and presentation of the results obtained: We used a combination of various data preprocessing, unbalanced processing, and feature selection methods (Comparison of Multiple

Classification Methods and Data Preprocessing Methods) After comparison, we finally used the Random Forest feature selection method, used the artificial completion in the oversampling technique, and finally used the Random Forest classification model, and end up with this result.

6. REGRESSION

6.1. Description of ANN

Artificial Neural Network is a type of machine learning algorithm that is designed to simulate the workings of the human brain. It contains several neurons that process and transmit information in each layer, allowing the network to learn from data and make predictions or decisions based on that information. The network learns and trains data repeatedly, gradually adjusts and changes the weights of neuron connections to achieve processing data and simulating the relationship between input and output.

ANN can process large amounts of data and identify complex patterns, leading to high accuracy in predictions and decision making. Besides, ANNs can be easily scaled up to handle larger data sets and more complex tasks. What is more important is that ANN can handle noisy and missing data, making them suitable for handling real-world data. ANN can learn and adapt over time, similar to how human brains learn and adapt to new information. This allows ANNs to improve their performance over time.

Compared with ANN, many other models show limited flexibility. Such as linear regression or logistic regression, they are limited in their ability to model complex relationships between variables. Other model also cannot learn and adapt to complex patterns and relationships based on data.

6.2. Parameter of ANN

One model used in regression prediction by ANN is MLPRegressor. It has several parameters. In this experiment, it is found that when hidden_layer_sizes value (20,), activation is 'relu', solver is 'sgd', alpha is 0.01, max_iter is 10000, this model gets an optimal predicted result after multiple experiments are observed. Besides, scheme of getting average of prediction by training models for several times also contributes to improve the accuracy and stability of prediction.

6.3. Results

MAE of train data:18.645254702885744
MAE of test data:21.87404403258685

Explanation and presentation of the results obtained: We combine various data preprocessing and feature selection methods (Comparison of Multiple Regression Methods and Data Preprocessing Methods). After comparison, we finally deal with data by filling missing values with modes and use the selectFromModel with AdaBoostRegressor as estimator to select features sorted by weight, because this model performs excellent and the results of training data and test data in MAE are close. Finally, we use the MLPRegressor model, and end up with this result.

7. CONCLUSION

Finally, we picked out the best combination, which is Random Forest feature selection method, artificial completion in the oversampling technique and Random Forest model for classification ; SelectFromModel feature selection method with AdboostRegressor as estimator and MLPRegressor model for regression.

8. REFERENCE

[1] Chawla N V , Bowyer K W , Hall L O , et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2002, 16(1):321-357.

[2] Hui H , Wang W Y , Mao B H . Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[J]. Lecture Notes in Computer Science, 2005.

[3] He H , Yang B , Garcia E A , et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]// Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008.

Task and Weighting	**Data pre-processing(10%)**	**Feature-selection(20%)**	**ML method development(30%)**	**Method Evaluation(10%)**	**Report writing(30%)**
Yi Li	0	50%	10%	10%	20%
Wenxuan Zhu	0	30%	20%	20%	20%
Xiaofei Li	0	10%	20%	0%	40%
Qitao Ye	60%	0	30%	20%	10%
JinShuai Chang	40%	10%	20%	50%	10%