Secom Data Set Classification Report

Group member ：ZHOU Bin　YU Shilong　ZHANG Liangshan　HU Fuqin

Pleader ：ZYZH　　Date ：2018.08.29

CONTENTS

Part One

Problem and data description

Problem and data description

Illustrative context

According to the test data of the wafer, it is accurate to determine whether the wafer is qualified

Model requires

Inspection accuracy of nonconforming product>72%

Secom data set

Two kinds of label

590 features

1567 samples

104 unqualified samples

1,463 qualified samples

Sample proportion 14:1 (pass: fail)

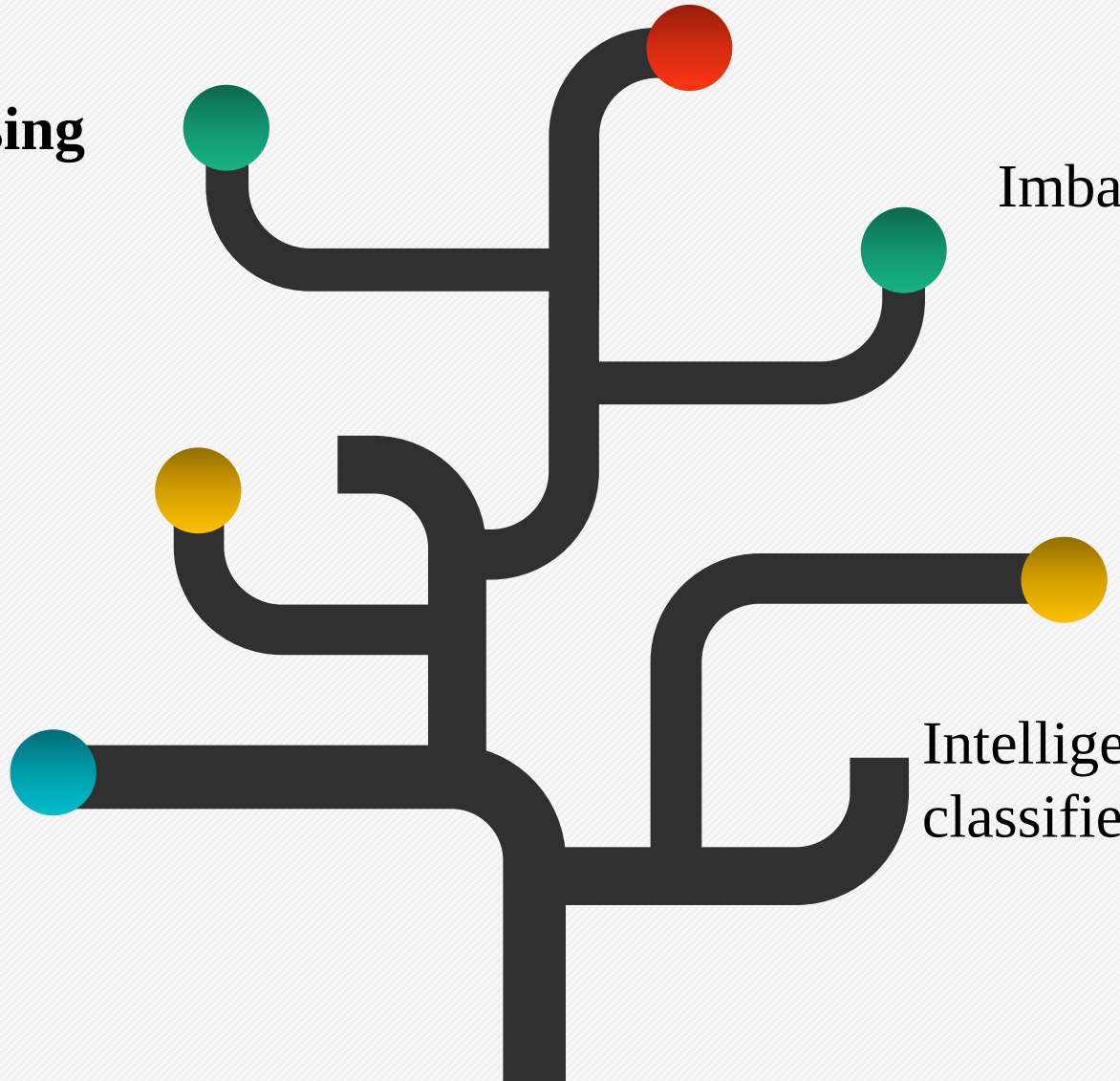
There are multiple nulls NAN

Essence: unbalanced data set biclassification problem with multiple characteristics and multiple books (abnormal point problem)

Steps and methods

Data preprocessing

Imbalance problem solved



Intelligent algorithm and classifier selection

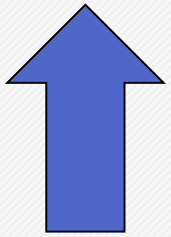
Feature selection

Part Two

Data preprocessing

normalization

Simple scaling	Standard deviation standardization	Nonlinear normalization
min-max standardization()	**x = (x - u)/σ**	**log、Exponential, Tangent, etc**
This normalization method is applicable to the case where the values are concentrated.If Max and min are unstable, it is easy to make the normalization result unstable	In classification and clustering algorithms, the second method (z-score standardization) performs better when distance is used to measure similarity, or when PCA is used to reduce dimension.	It is often used in scenarios where the data is highly differentiated, some of the values are large and some of them are small.The calculation is too large.



Deletion processing

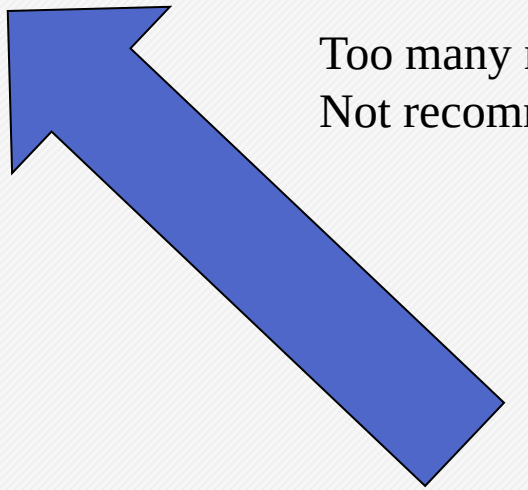
0fill

Mean replaceme

delete

异常值(NAN)替换

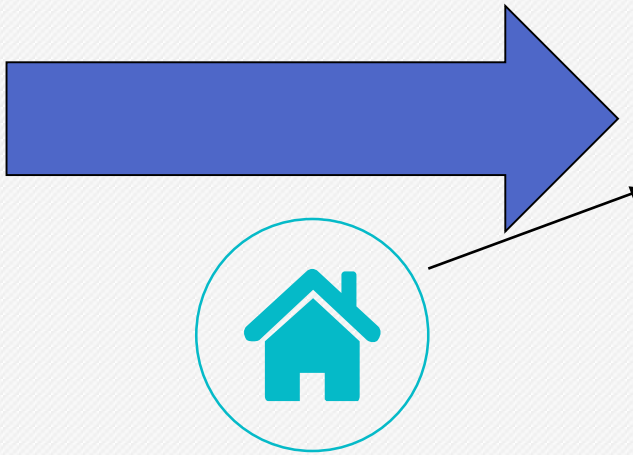
替换为0		替换为均值	
迭代次数	每代适配值	迭代次数	每代适配值
1	0.434782609	1	0.733333333
2	0.434782609	2	0.733333333
3	0.434782609	3	0.733333333
4	0.47826087	4	0.8
5	0.47826087	5	0.8
6	0.52173913	6	0.8
7	0.52173913	7	0.8
8	0.52173913	8	0.8
9	0.52173913	9	0.8
10	0.52173913	10	0.8



Too many missing samples
Not recommended

Exception handling? (out)

Partition data set



SMOTE

The training set needs to be balanced
The test set is not required

Set aside method

Divide a training set and a test set

Leave one and leave P

Cross validation is required

folding method

Divide K data sets for cross validation

Cross validation

The two need cross validation after cross validation

One class SVM

https://github.com/Meena-Mani/SECOM_class_imbalance/blob/master/secomdata_ocsvm.ipynb

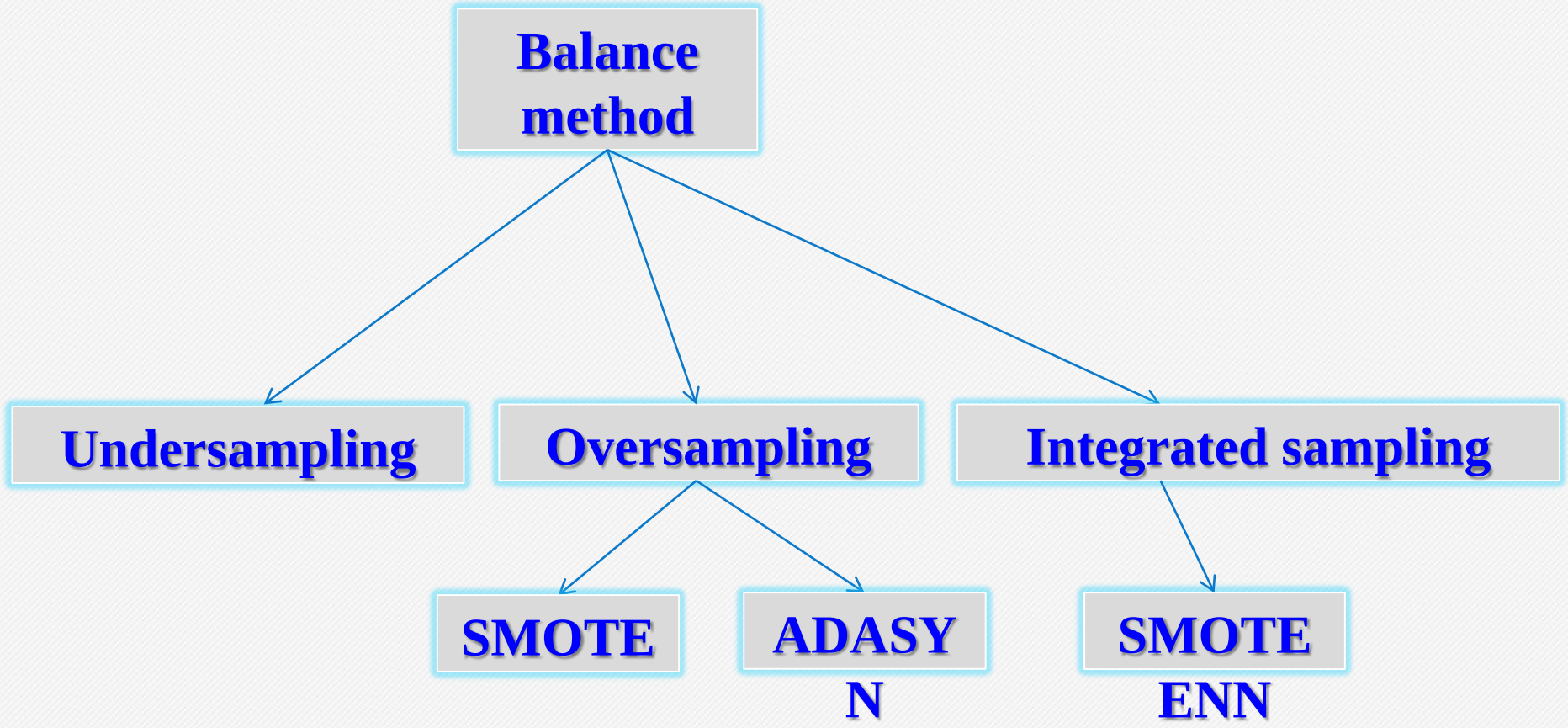
nfeatures	nu	gamma	train error	test error	outlier error
40	0.03	0.07	147(14.36%)	112(25.51%)	44(42.31%)
	0.04	0.07	124(12.11%)	112(25.51%)	44(42.31%)
	0.05	0.07	135(13.18%)	112(25.51%)	44(42.31%)
	0.03	0.08	157(15.33%)	127(28.93%)	40(38.46%)
	0.04	0.08	203(19.82%)	127(28.93%)	40(38.46%)
	0.05	0.08	193(18.85%)	127(28.93%)	40(38.46%)
	0.03	0.09	169(16.50%)	148(33.71%)	32(30.77%)
	0.04	0.09	208(20.31%)	148(33.71%)	32(30.77%)
	0.05	0.09	217(21.19%)	148(33.71%)	32(30.77%)
	0.03	0.10	203(19.82%)	175(39.86%)	29(27.88%)
	0.04	0.10	186(18.16%)	175(39.86%)	29(27.88%)
	0.05	0.10	236(23.05%)	175(39.86%)	29(27.88%)
	0.03	0.15	374(36.52%)	283(64.46%)	11(10.58%)
	0.04	0.15	373(36.43%)	283(64.46%)	11(10.58%)
	0.05	0.15	262(25.59%)	283(64.46%)	11(10.58%)
	0.03	0.20	396(38.67%)	376(85.65%)	1(0.96%)
	0.04	0.20	484(47.27%)	375(85.42%)	1(0.96%)
	0.05	0.20	508(49.61%)	375(85.42%)	1(0.96%)

Shall not be used

Part Three

Equilibration

Balance method



Class imbalance

SMOTE

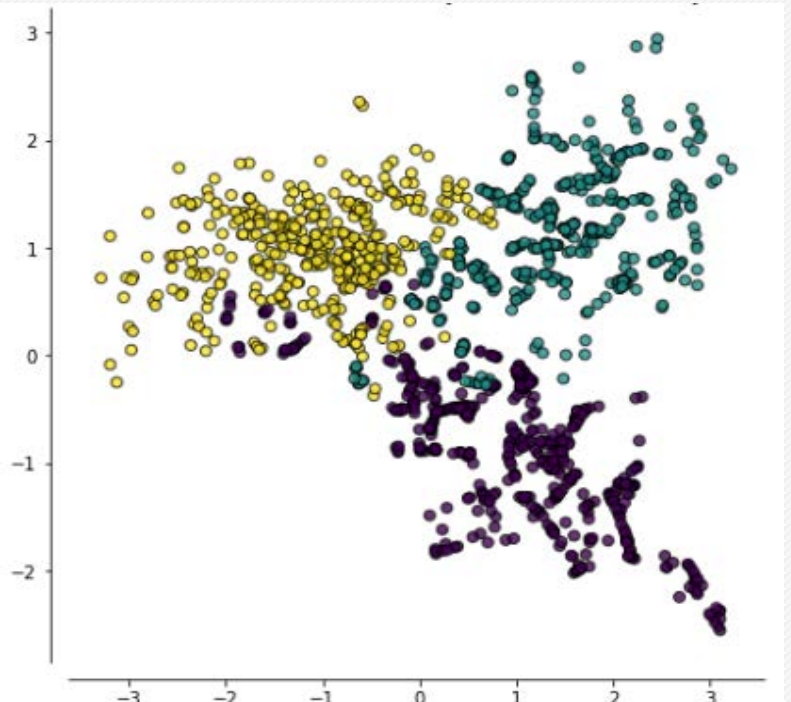
The new minority samples are synthesized in a specific way so that the two categories in the training set are roughly equal in number.

ADASYN

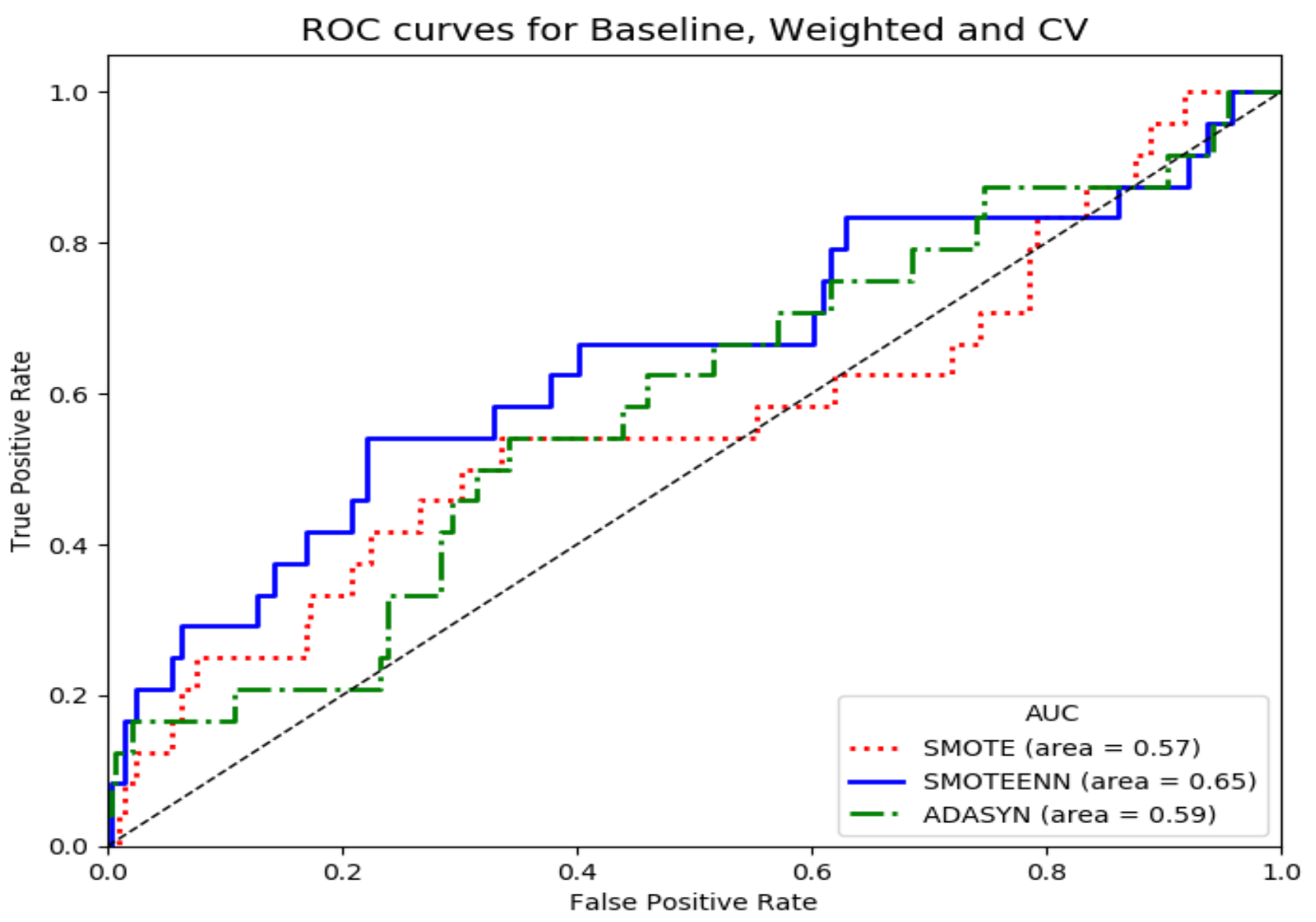
Generate different numbers of new samples for different niche samples based on data distribution.

SMOTEENN

A new minority sample is synthesized by SMOTE, and then the noise generated during the SMOTE process is cleaned by ENN.



Comparison of three methods of class imbalance



Overall SMOTEEN is better

Part Four

Feature selection and dimension reduction

Feature selection

Purpose: To speed up training, lower model complexity and better interpretability, higher accuracy (selected features), and reduced overfitting.

Recursive feature elimination

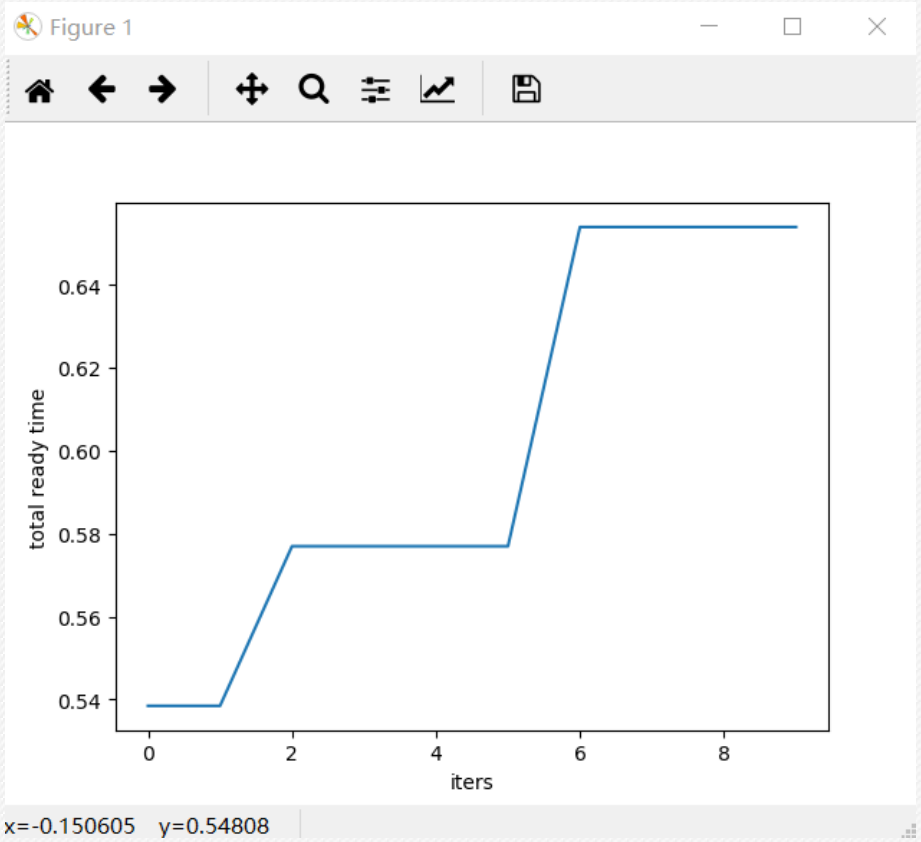
A base model (SVC / LR) is used to perform multiple rounds of training. After each round of training, the characteristics of several weight coefficients are eliminated, and the next round of training is performed based on the new feature set.

Mutual information

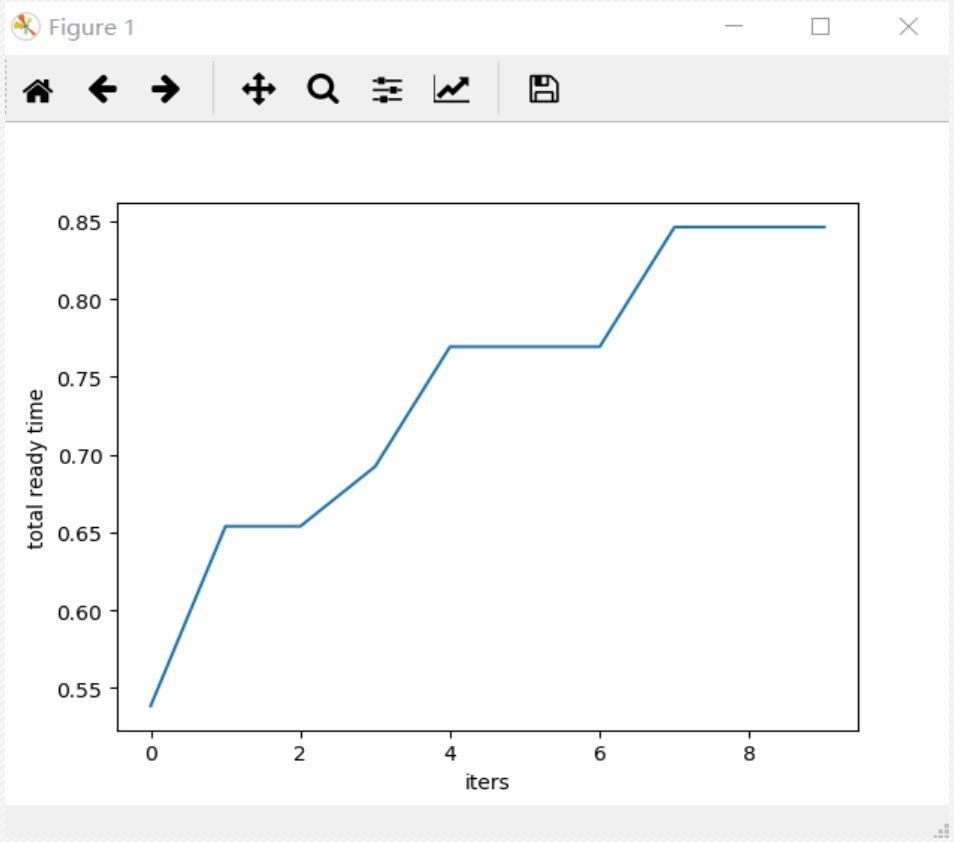
Mutual information measures the degree of interdependence between two variables, indicating that the content of information shared between two variables is not limited to a linear relationship.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}$$

Recursive feature elimination method vs mutual information

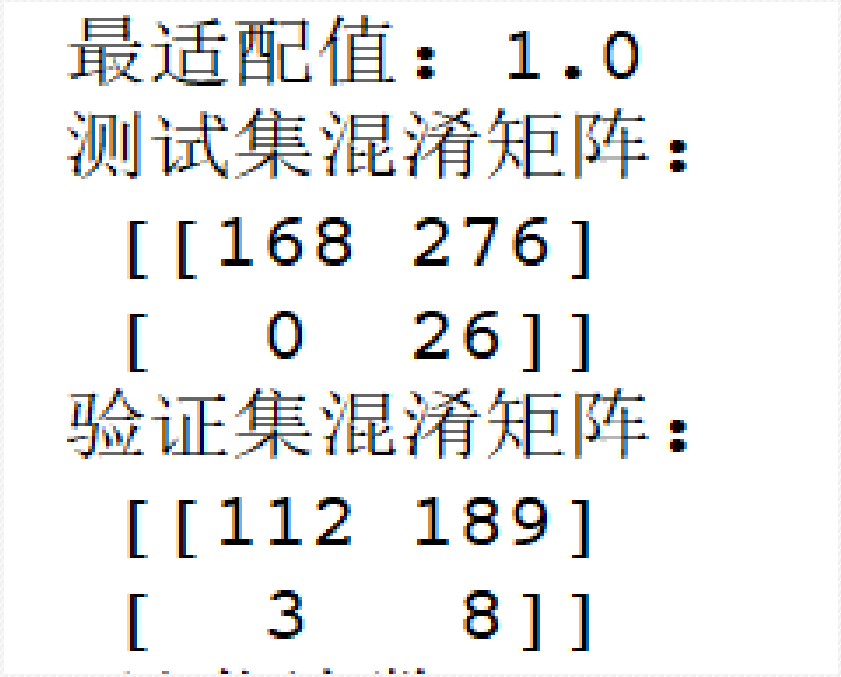


mutual information



recursive feature elimination

Dimensionality reduction



Using 200 feature PCA to reduce dimensionality

Not used!

Mapping from high dimensional feature space to low latitude feature space

Advantages of PCA:
1. Minimum error.
2. extracted the main information

Disadvantages of PCA:
1. The principal component with small contribution rate may often contain important information about sample differences.
2. over-fitting is serious

Part Five

Intelligent algorithm and classifier

Intelligent algorithm selection（GA）

01

Define hyperparameters

Number of iterations, population size, crossover probability, mutation probability, gene length

02

Initial population

The coding method adopts 01 coding, and the genetic sequence (feature selection) of each individual is randomly arranged.

03

Survival of the fittest

Choosing the appropriate fitness function, using the tournament operator to select the parent, cross mutation to produce the child

04

Genetic Variation

Fitness function

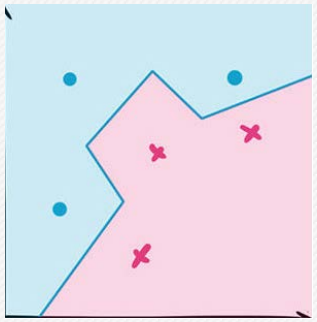
In this problem, the training set training model, the test set through the model to obtain the recall rate, the greater the recall rate represents the better classification effect, so the recall rate as a fitness function.

The recall rate (also called the recall rate) is the ratio of the number of related documents retrieved to the number of related documents in the document library. In this question, the ratio of the predicted positive samples to the total positive samples

Classifier

KNN

For the point to be judged, find the data points closest to it, and determine the type of the point to be judged according to their type.



Rf

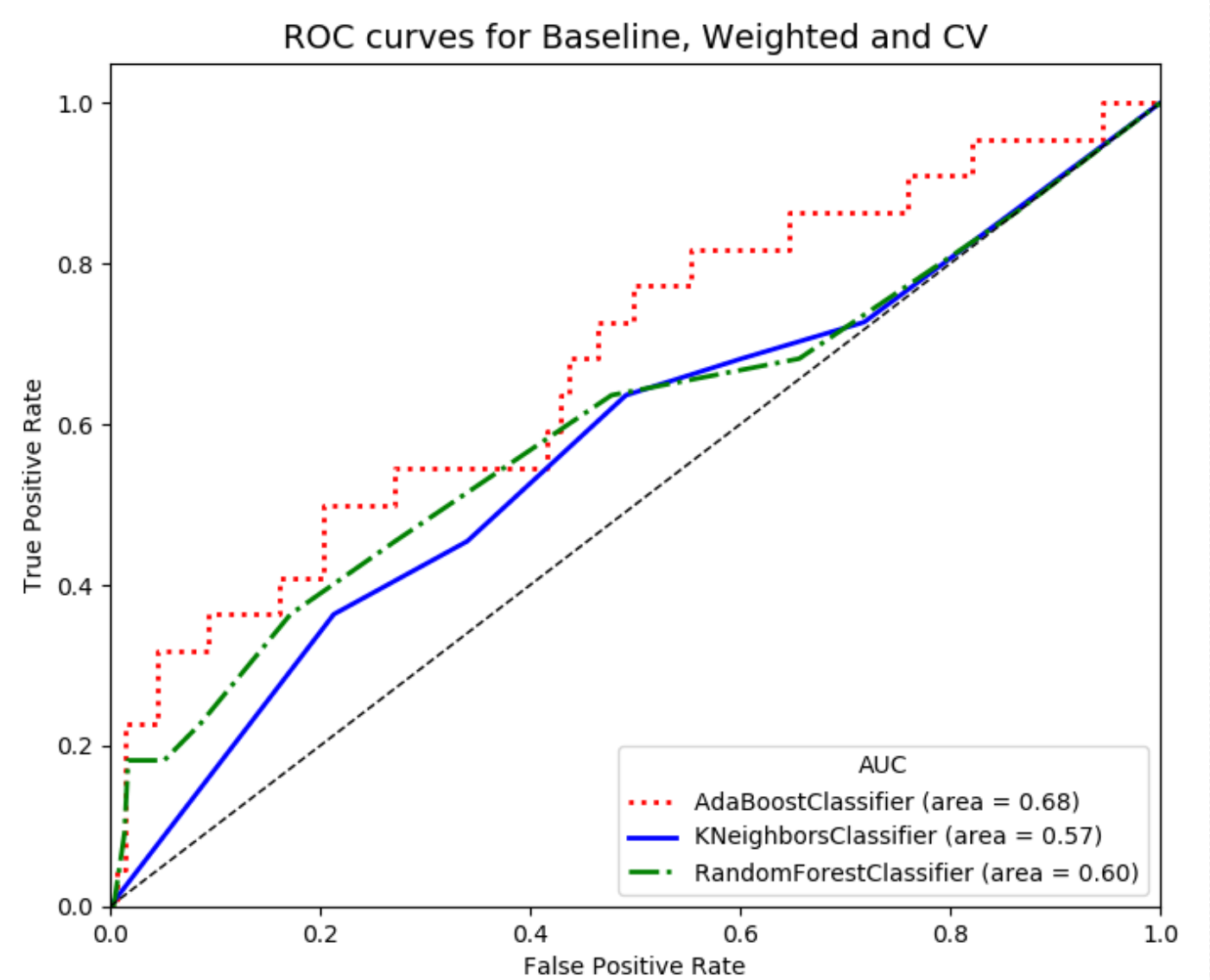
Randomly select different features and training samples, generate a large number of decision trees, and then combine the results of these decision trees to perform the final classification.

AdaBoost

Core idea: three stinkers, top one Zhu ge Liang

The weak classifiers are combined according to a certain calculation method to form a strong classifier. There is an association between the classifiers. The final classification is the result of multiple classifier combinations.

Comparison of three classifiers



With FPR as the horizontal axis and TPR as the vertical axis, the ROC space is obtained. The AUC value is the area covered by the ROC curve. The larger the AUC value, the better the classifier classification effect.

Part Six

Conclusion and summary

Flow chart

Data preprocessing

- Read data
- Standard deviation standardization
- Mean fill data
- Leave one method to divide the data set

Data balance

- Training set SOMTENN

Feature selection and dimensionality reduction

- Variance feature elimination
- Recursive feature elimination method to 100

Intelligent algorithm and classifier

- Integrated classifier classification (adaboost)
- GA algorithm selection subset

Conclusion

```
最适配个体：[1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0,
 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1,
 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1,
 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1,
 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
最适配值：0.7307692307692307
测试集混淆矩阵：
 [[308 136]
 [  7  19]]
验证集混淆矩阵：
 [[220  81]
 [  7   4]]
```

Meet the requirements

Summarize

HU Fuqin：Data preprocessing

ZHOU Bin： Balanced

YU Shilong：Pre-screening

ZHANG Liangshan：Genetic algorithm

division of work

1) Team level: the division of labor is not clear, resulting in duplication of labor;
2) Personal level: poor ability to write code;
3) Project level: the model hyperparameter tuning is not enough, the degree of visualization is poor, and the data analysis is not enough.

inadequate

Thanks for listening

will request earnestly fellow teachers to

give the criticism to point out mistakes！

Mentor ：Lord Bao　Pleader ：ZYZH