Genetic algorithm based Engineering problems in wafer fabrication testing

Members: Junjie He, Xin Liu, Liling Zuo

April 29 2018

東華大學
DONGHUA UNIVERSITY

Contents

Problem analysis

➢ Problem object

➢ Dataset source：

- Data measured by various sensors in the wafer fabrication test process

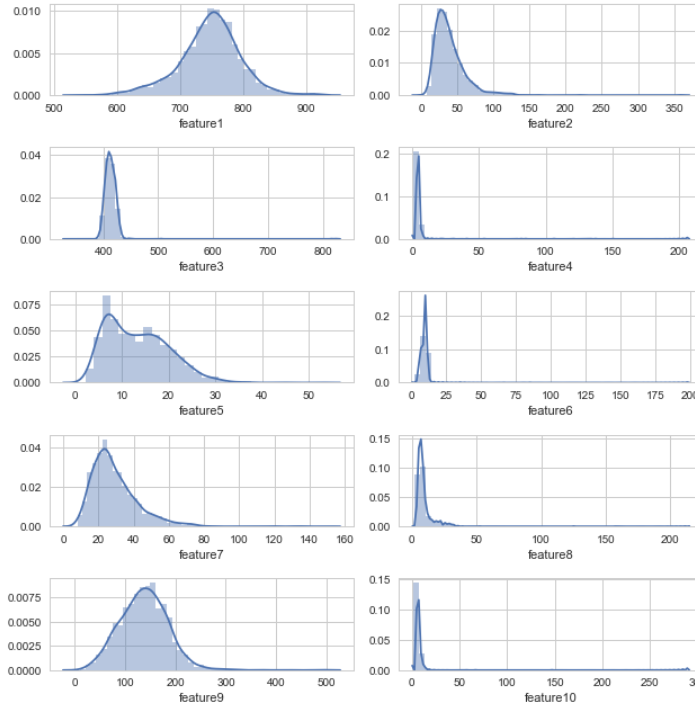
Data Set Characteristics:	Multivariate	**Number of Instances:**	1567	**Area:**	Computer
Attribute Characteristics:	Real	**Number of Attributes:**	591	**Date Donated**	2008-11-19
Associated Tasks:	Classification, Causal-Discovery	**Missing Values?**	Yes	**Number of Web Hits:**	69656

- Engineering problems
- According to the wafer test data, accurately determine whether the wafer is qualified or not, and reduce the false positive rate.

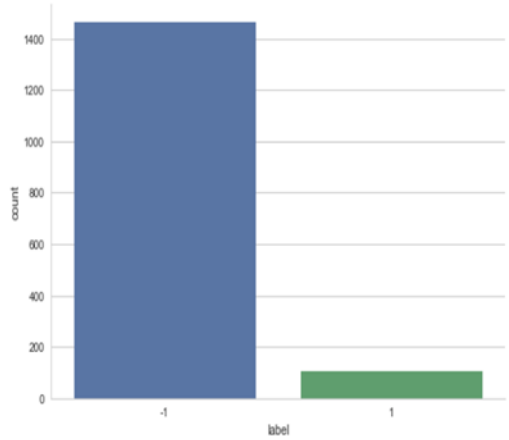
Problem analysis

data analysis

- Problem characteristics Too high feature dimension : 590



- Inconsistent distribution of each feature



- Sample data imbalance

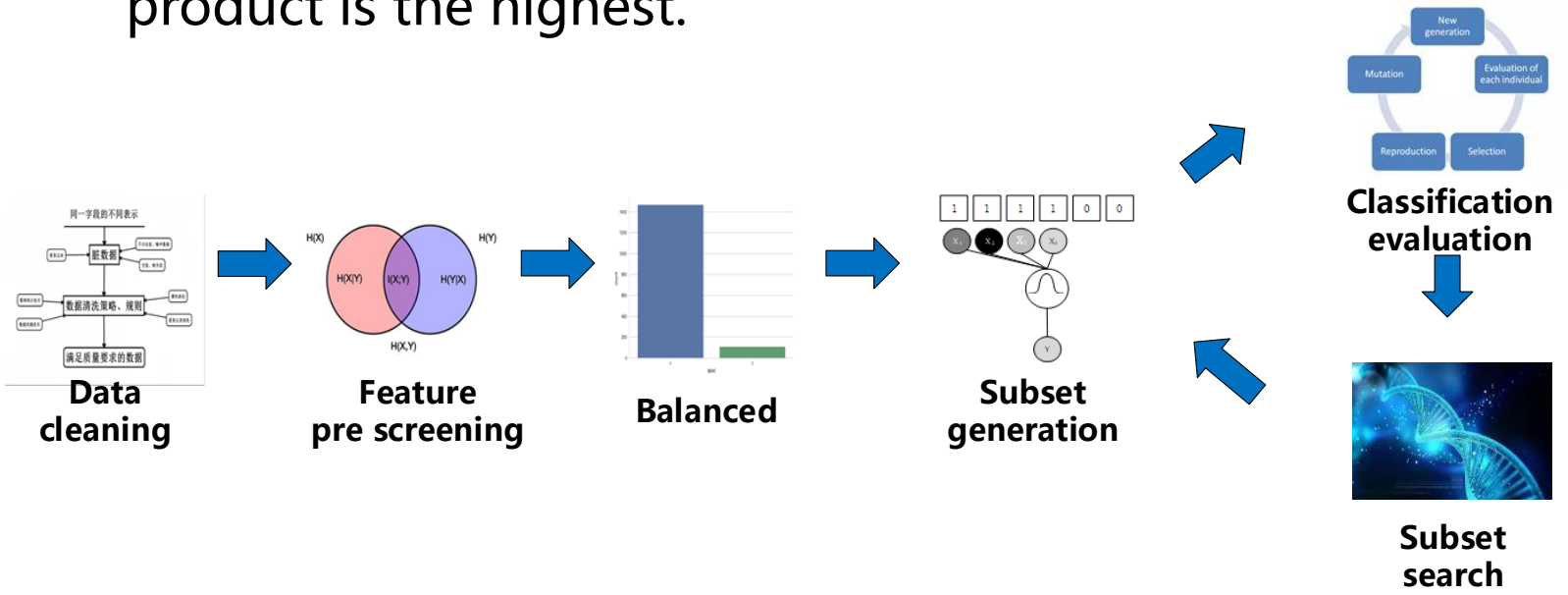
2270.256	1258.456	1.395	100
2207.389	962.5317	1.2043	100
2208.856	1157.722	1.5509	100
NaN	NaN	NaN	NaN
2207.389	962.5317	1.2043	100
2207.389	962.5317	1.2043	100
2160.367	899.9488	1.4022	100
2203.9	1116.413	1.2639	100
2257.167	1437.957	1.4918	100

- Null value exists in the data set

Contents

Problem solving

- Data cleaning for data loss issues
- To reduce irrelevant features and redundant features, mutual information calculation and feature pre-screening
- Data collection into data imbalance issues
- The genetic algorithm is used to further optimize the features of the pre-screening, and a certain feature subset combination is selected, so that the detection accuracy of the non-conforming product is the highest.



Contents

Detailed explanation of key parts

Data cleaning

- Eliminate duplicate data
- Fill in missing values

2270.256	1258.456	1.395	100
2207.389	962.5317	1.2043	100
2208.856	1157.722	1.5509	100
NaN	NaN	NaN	NaN
2207.389	962.5317	1.2043	100
2207.389	962.5317	1.2043	100
2160.367	899.9488	1.4022	100
2203.9	1116.413	1.2639	100
2257.167	1437.957	1.4918	100

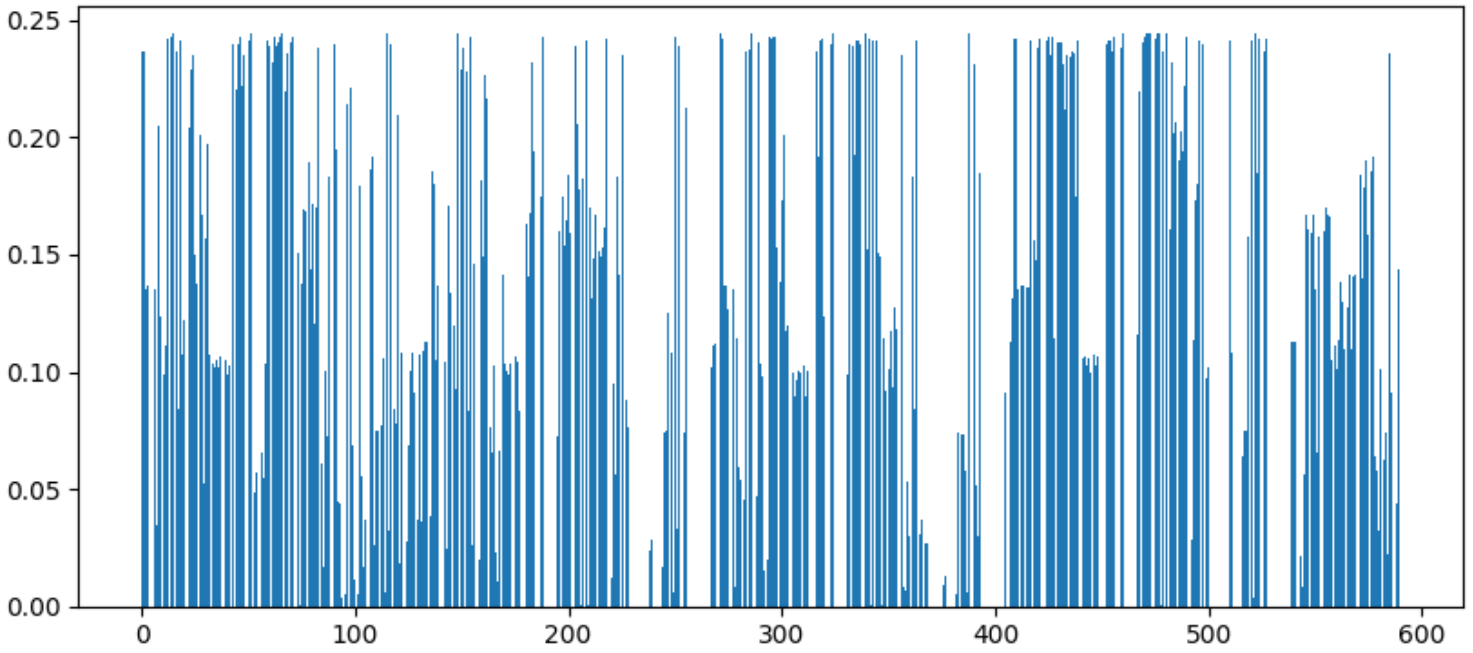
Mean fill

2270.256	1258.456	1.395	100
2207.389	962.5317	1.2043	100
2208.856	1157.722	1.5509	100
2200.547	1396.377	4.197013	100
2207.389	962.5317	1.2043	100
2207.389	962.5317	1.2043	100
2160.367	899.9488	1.4022	100
2203.9	1116.413	1.2639	100
2257.167	1437.957	1.4918	100

- Normalized

Detailed explanation of key parts

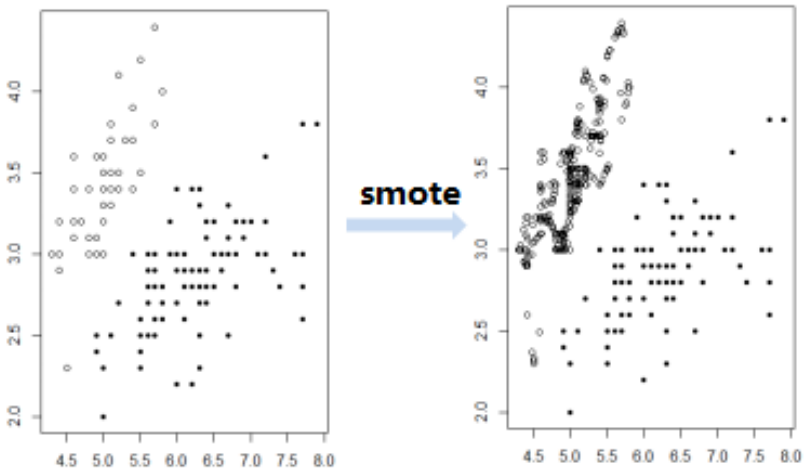
- **Pre-screening**
- **Screening basis: mutual information**
- Calculate the correlation between 590 features and labels, and use greedy algorithm to select the top 100 features with large mutual information.



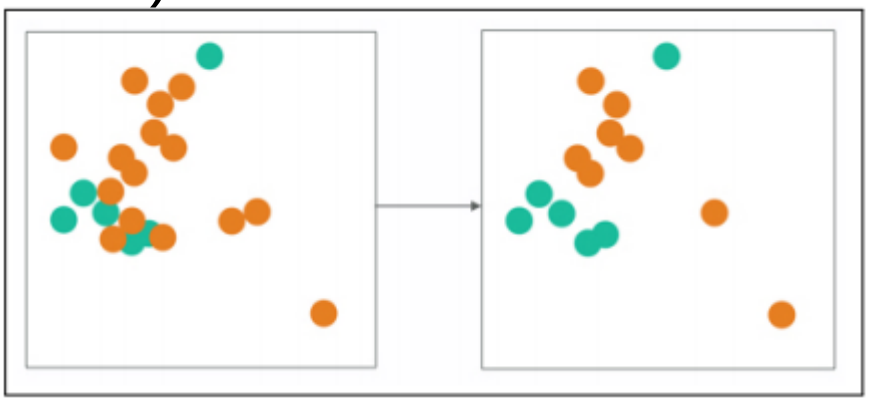
Detailed explanation of key parts

Balanced

- Synthetic Minority Oversampling Technique（SMOTE)
  - Oversampling balance data



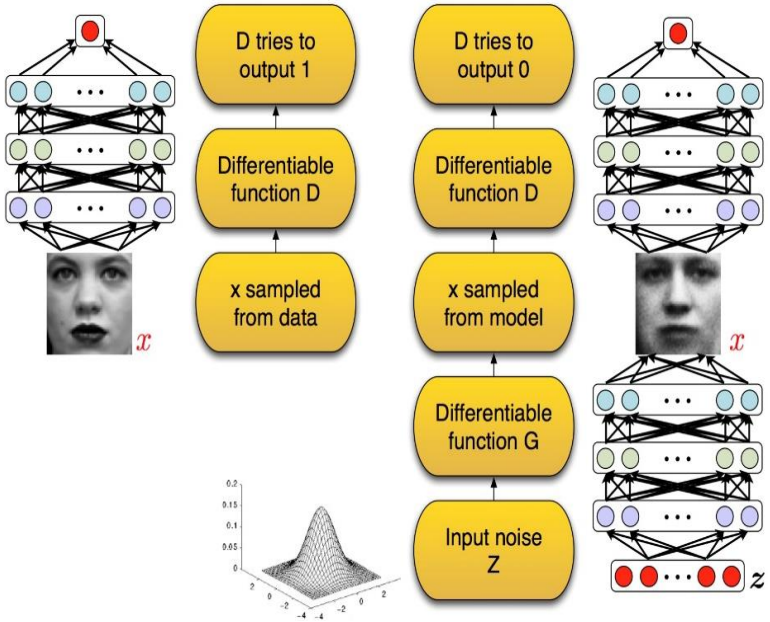
- Edited Nearest Neighbor（ENN）
  - Undersampling enhanced data separability



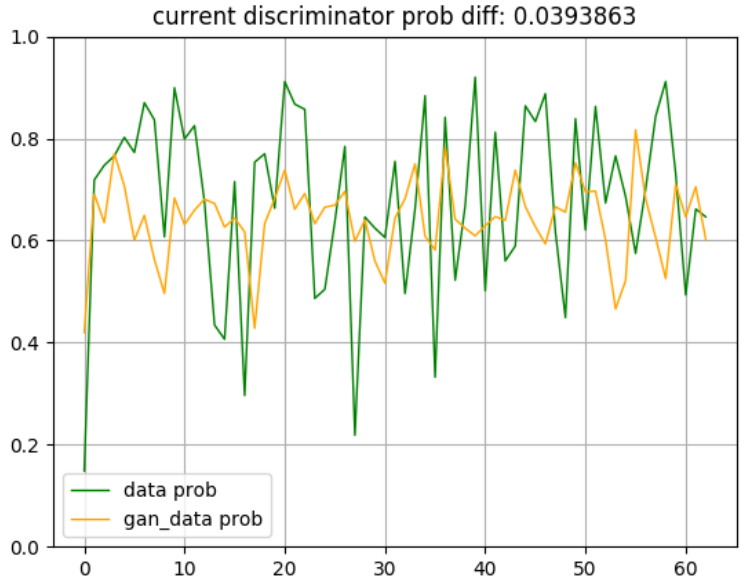
Detailed explanation of key parts

➢ Balanced

- Generated GAN



GAN frame

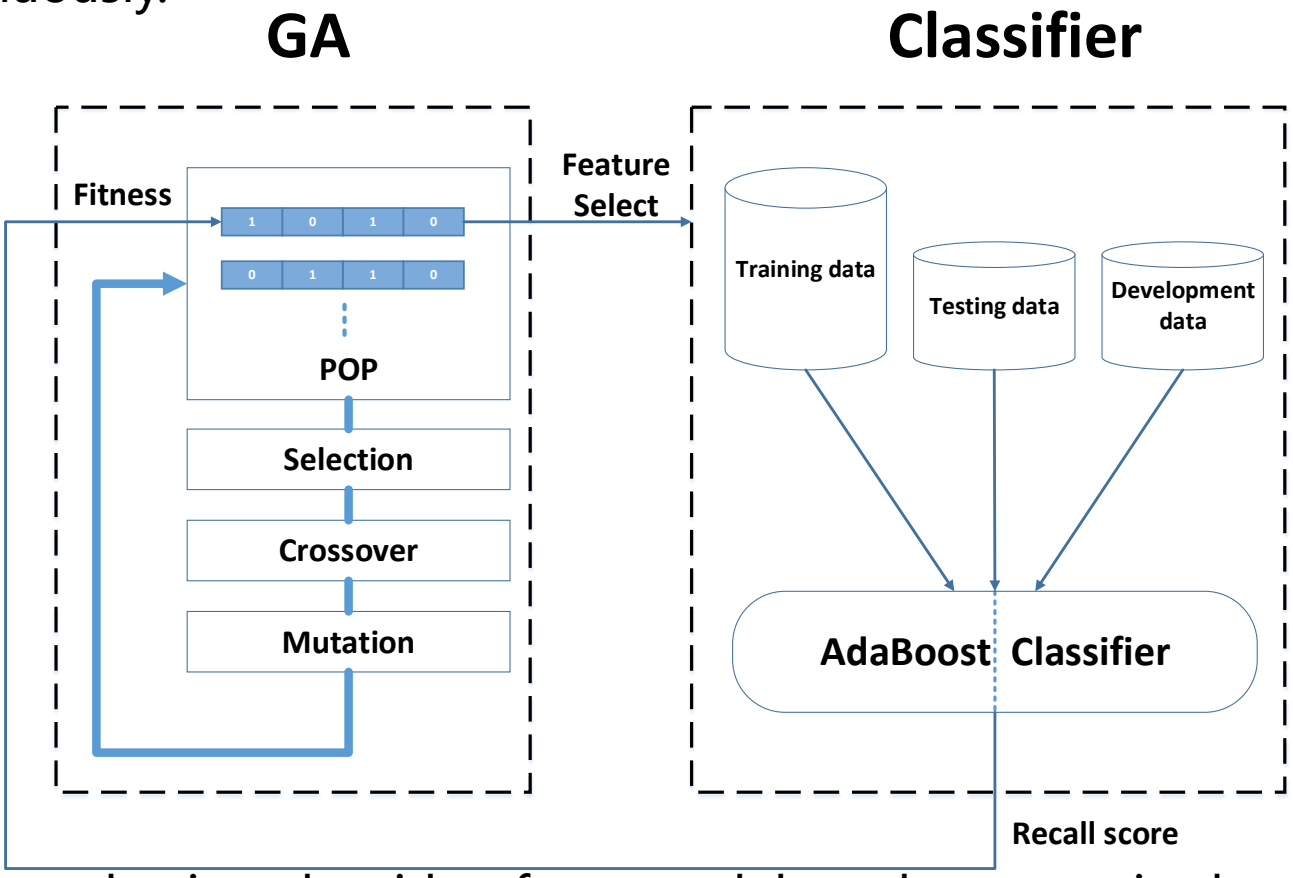


Simulation data and real data Discriminator probability result output

Detailed explanation of key parts

Feature selection based on genetic algorithm

The features are selected by 0-1 coding, the selected features are learned by the AdaBoost classifier, and the recall rate of the test set is returned as the fitness value, and the population is iterated continuously.

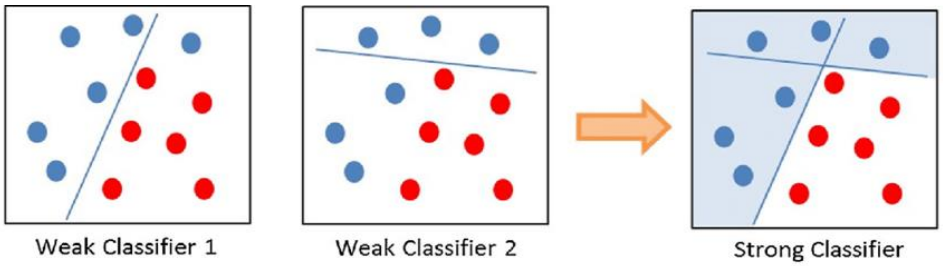


Feature selection algorithm framework based on genetic algorithm

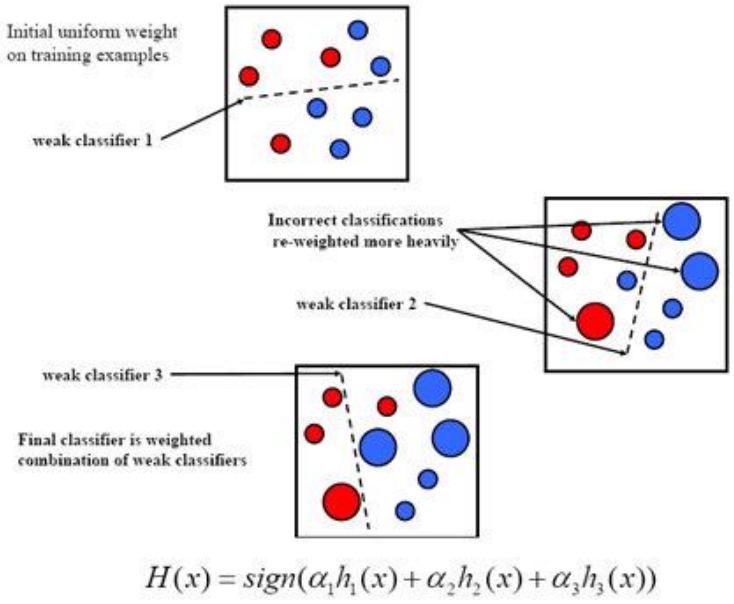
Detailed explanation of key parts

Integrated learning algorithm AdaBoost for data classification

The AdaBoosting algorithm iterates on a base learner, and each adjustment focuses on weight updates for the current misclassified sample.



AdaBoosting integrated learning effect



AdaBoosting integrated learning principle

東華大學
DONGHUA UNIVERSITY

Contents

Comparison and analysis

Different classifiers (uncharacterized screening)

The type of classifier	Recall	Processing time（s）
naive_bayes_classifier	0.80	0.005
knn_classifier	0.75	0.007
logistic_regression_classifier	0.75	0.03
decision_tree_classifier	0.86	0.03
svm_classifier	0.94	0.7
Adaboosting	0.87	0.4

Comparison and analysis

Different crossover operators

crossover	mutate	selection	Recall
cxOnePoint	mutFlipBit	selTournament	0.857
cxTwoPoint	mutFlipBit	selTournament	0.75
cxUniform	mutFlipBit	selTournament	0.8125
cxPartialyMatched	mutFlipBit	selTournament	0.85
cxOrdered	mutFlipBit	selTournament	0.83

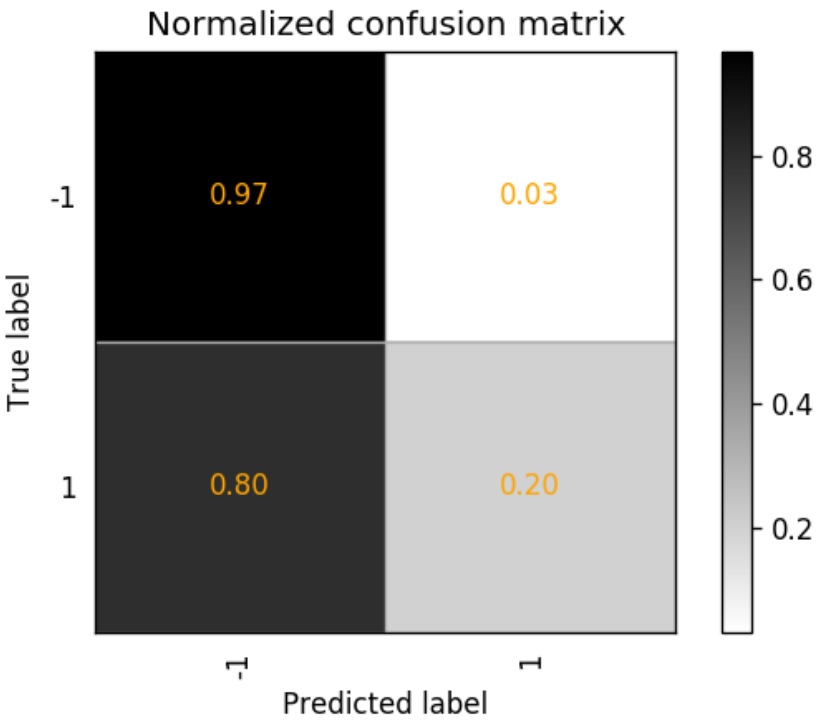
Comparison and analysis

Different ways of variation

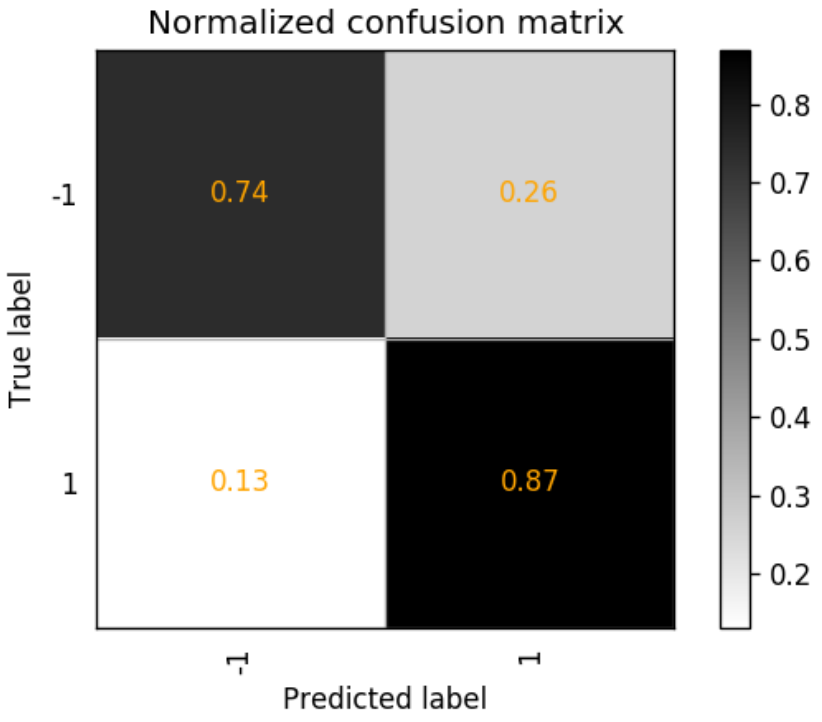
crossover	mutate	selection	Recall
cxOnePoint	mutFlipBit	selTournament	0.857
cxOnePoint	mutShuffleIndexes	selTournament	0.7
cxOnePoint	mutGaussian	selTournament	0.714
cxOnePoint	mutFlipBit	selTournament	0.857
cxOnePoint	mutFlipBit	selRoulette	0.625
cxOnePoint	mutFlipBit	selStochasticUniversalSampling	0.73

Comparison and analysis

➢ Different data balancing methods



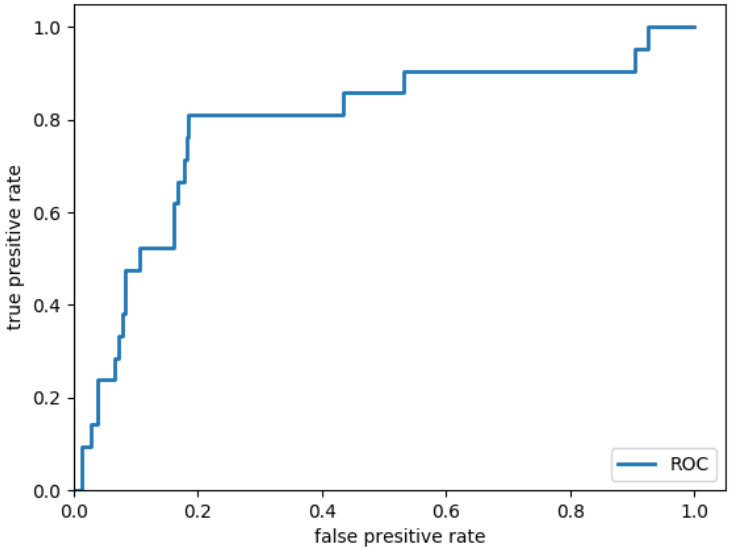
The result of Gan



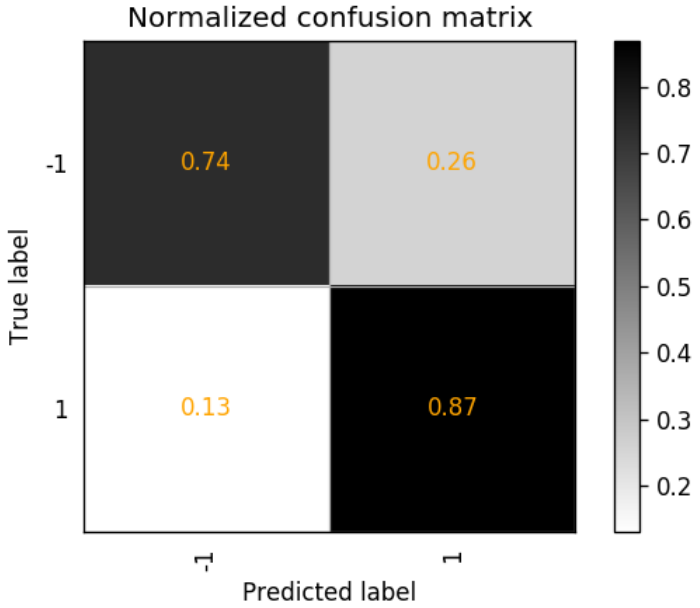
The result of Smote&Enn

Comparison and analysis

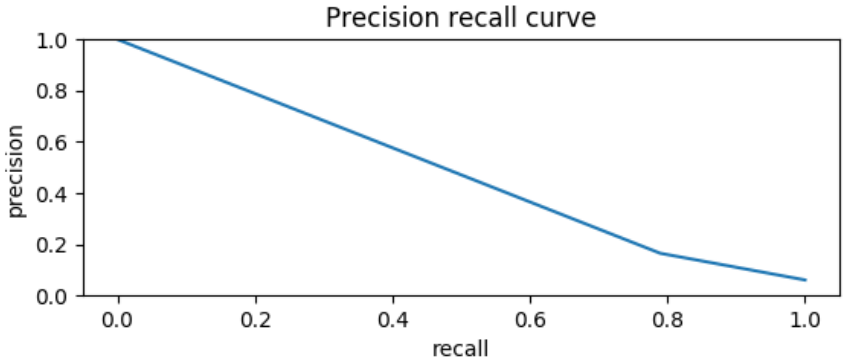
Experimental result



ROC curve



Confusion matrix



PR curve

Comparison and analysis

Experimental result

- **Best individual is**

[1, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1]

- **The best recall is:**

0.7561538461538461

- **The last features that we choose is**

[15, 18, 46, 50, 51, 59, 60, 65, 71, 90, 115, 117, 148, 154, 203, 208, 252, 271, 283, 285, 289, 318, 319, 321, 324, 333, 339, 341, 388, 410, 417, 421, 423, 424, 430, 439, 440, 452, 456, 457, 460, 469, 472, 474, 477, 490, 496, 510, 520, 522, 527]

Contents

Summary and outlook

➢ Summary

- Familiar with a variety of data imbalance processing methods: Gan、Smote
- Skilled in using genetic algorithms for feature selection
- Skilled in using a variety of libraries, toolboxes for algorithm implementation, visualization, etc.
- The team has both division of labor and collaboration, and progress is faster.

➢ Outlook

- Exploring the quality evaluation criteria of a Gan method to generate one-dimensional data
- The preliminary data analysis work needs to be further

THANKS