Classical Breast Cancer Screening Database

University Mohamed 6 Polytechnic

Dr.IDRI Ali

Phd.ZIZAAN Asmaa **Phd.ZEROUAOUI** Hasnae

IHBACH Mohamed Yassine **KHOUDRAJI** Wissale

Departement AL KHWARIZMI

Master MSD

2020/2021

Plan

1. Extracted Datasets after Preprocessing
2. Binary Classification
   - Data Balancing
   - Feature selection
3. Multiclass Classification
4. Building models
5. Hyper parameters Tuning
   - SVM
6. Metrics and Scores
   - Accuracy
   - The other metrics
7. Training

Summary

After imputing the BMI with the following imputers :

- SVR
- Random Forest
- KNN

3 Datasets : d_svr , d_knn , d_rf

Summary

Binary classification

- "Cancer" as target
- "radiologist_birads" as feature

Under, Over and hybrid sampling

- **Under sampling** : Take a random sample from the majority class as sized as the minority class.
- **Over sampling** : generate synthetic observations from the minority class.
- **Hybrid sampling** : under sampling and over sampling at the same time.

Univariate filters

3 methods :

- Mutual information
- CHI_2
- Variance

We took 40% , 60% and 80% most relevant features

Multivariate filters

The best K features are not the K best features 2 methods :

- CFS : Correlation-based Feature selection (we don't need to specify k)
- MRMR : Maximum Relevance and Minimum Redundancy

We took 40% , 60% and 80% most relevant features

0	d_knn_under_chi2_40
1	d_knn_under_chi2_60
2	d_knn_under_chi2_80
3	d_knn_under_mi_40
4	d_knn_under_mi_60
5	d_knn_under_mi_80
6	d_knn_under_var_40
7	d_knn_under_var_60
8	d_knn_under_var_80
9	d_knn_hybrid_chi2_40

Figure – examples of the extracted datasets

Summary

Multiclass classification

- "radiologist_birads" as target
- "Cancer" as feature

Target Values

From

```
0 = Needs additional imaging
1 = Negative
2 = Benign finding(s)
3 = Probably benign
4 = Suspicious abnormality
5 = Highly suggestive of malignancy
```

to

```
0 = Needs additional imaging
1 = Negative (1 & 2)
2 = Positive (3 , 4 & 5)
```

Data Balancing

Same as Binary classification

- Under sampling
- Over sampling
- Hybrid sampling

Feature selection

Same as Binary classification

- Chi_2
- Mutual information
- Variance

- CFS
- MRMR

Summary

10 models

- Decision Tree
- Logistic regression
- KNN
- SVM
- Random Forest
- XGBoost , ADABoost, CATBoost , LightGBM
- MLP

Summary

Decision Tree

- criterion : ['gini','entropy']
- splitter : ["best", "random"]
- max depth : range(5,20) + None
- min samples leaf : range(5,20)

Logistic Regression

- penalty : "l1", "l2"
- C (Inverse of regularization strength) : [0.01,0.1,1,10] :
- Decision Threshold : [0.1,0.2,...,0.9]

KNN

- nbr neighbors = range(2,15)
- weights ['uniform','distance']
- algorithm = ['auto', 'ball tree', 'kd tree', 'brute']
- metric : ['hamming','canberra','braycurtis']

SVM

- C = [0.1, 1, 10, 100, 1000]
- gamma =[1, 0.1, 0.01, 0.001, 0.0001]
- kernel = ['rbf', 'linear', 'poly']

Random Forest

- nbr estimators : [10,100, 200,500]
- criterion : ['gini','entropy']
- max depth : [5,10,15,20,None]
- min samples leaf : [5,10,15,20]

ADABoost

- base estimator : Decision Tree
- nbr estimators : [10, 100, 200,500]
- learning rate : [0.001,0.01,0.1]
- algorithm : ['SAMME', 'SAMME.R']

SAMME.R uses the probability estimates to update the additive model, while **SAMME** uses the classifications only

XGBoost

- base estimator : Decision Tree
- nbr estimators : [10, 100, 200]
- learning rate : [0.001,0.01,0.1]
- subsample : [0.4,0.6,0.8]
- max epth : [5,10,15,None],
- booster : ['gbtree', 'gblinear']

CAToost

- iterations : [10, 100, 200,500]
- learning rate : [0.001,0.01,0.1]

MLP

- learning rate : ["constant","adaptive"]
- alpha : [0.1,0.0001]
- solver : ['adam']
- hidden layer sizes : [(100,),(200,)]
- batch size : [30,60]
- activation : ["relu"]

Summary

- Since the datasets are balanced we consider the accuracy as the main metric
- 5-fold : 5 values of accuracies

- Recall
- Precision
- F1-score
- ROC-AUC score

We save only the average values

Summary

Loading ...

The Training is launched ...