July 2, 2020

[61]:
```
import glob
import pandas as pd
import pyod
import os
from pyod.utils.data import generate_data, get_outliers_inliers
from pyod.models.abod import ABOD
from pyod.models.knn import KNN
from pyod.models.abod import ABOD
from pyod.models.cblof import CBLOF
from pyod.models.feature_bagging import FeatureBagging
from pyod.models.hbos import HBOS
from pyod.models.iforest import IForest
from pyod.models.knn import KNN
from pyod.models.lof import LOF
from pyod.models.pca import PCA
import numpy as np
from sklearn.model_selection import train_test_split
import tqdm
import matplotlib.pyplot as plt
from math import *
```

1 Benchmarks

[6]:
```
path = r'abalone\benchmarks'
path_list = glob.glob(os.path.join(path,'*.csv'))
```

[11]:
```
len(path_list)
```

[11]: 1725

[90]:
```
res= []
classifiers = {
     'ABOD'    : ABOD,
     'KNN' :  KNN,
    'PCA':LOF,
    'HBOS':HBOS,
    'IForest':IForest
```

```
}
scores = {
     'ABOD'    : [],
     'KNN' :   [],
    'PCA':[],
    'HBOS':[],
    'IForest':[]
}

for p in tqdm.tqdm(path_list):
    df = pd.read_csv(p)
    data_x  =df.iloc[:,6:13]
    data_y = df['ground.truth'].map({'nominal':0,'anomaly':1})
    train_x,test_x,train_y,test_y = train_test_split(data_x,data_y,test_size =␣
 ↪0.2)
    contamination = np.mean(train_y)
    for key , value in classifiers.items():
        clf = value()
        clf.fit(train_x)
        #scores[key].append(pyod.utils.precision_n_scores(test_y.values,clf.
 ↪predict_proba(test_x)[:,1],n=contamination))
        scores[key].append(np.mean(test_y.values == clf.predict(test_x)))
```

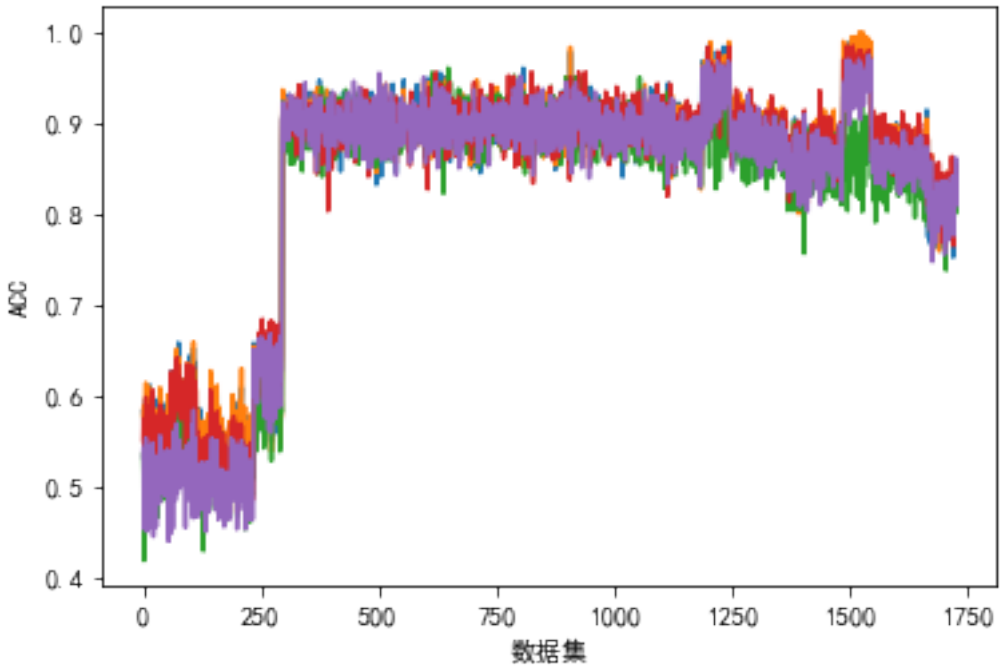
[90]:
```
1
```

[20]:
```
res =[]
for k,v in scores.items():
    res.append(scores[k])
res = np.array(res)
res = np.max(res,axis =0)
res.shape
```

[20]:
```
(1725,)
```

[49]:
```
for k,v in scores.items():
    plt.plot(scores[k])
plt.xlabel('  ')
plt.ylabel('ACC')
```

[49]:
```
Text(0, 0.5, 'ACC')
```



250 1500

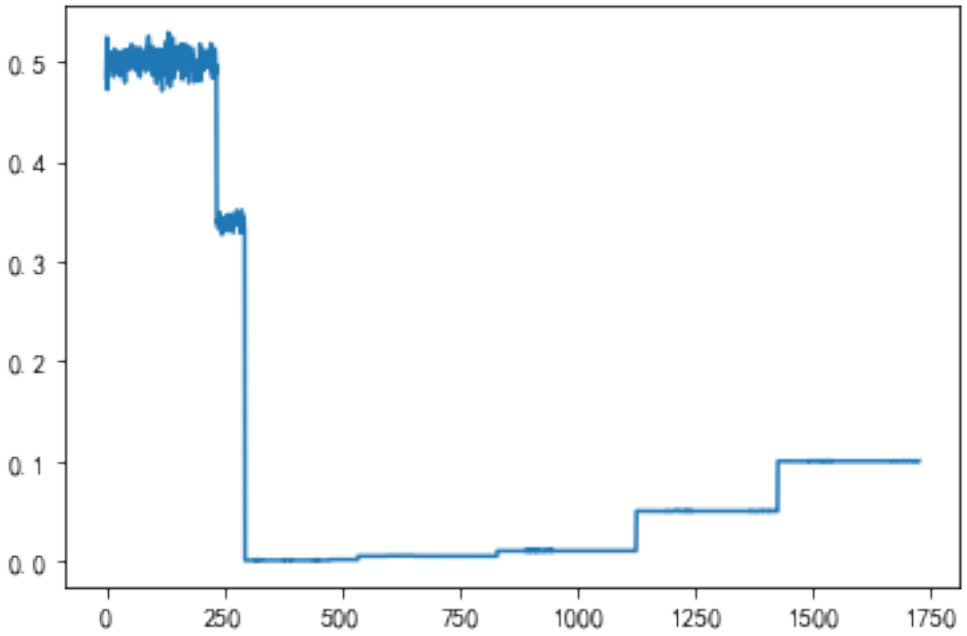
[54]:
```
n_diff =[]
for p in tqdm.tqdm(path_list):
    df = pd.read_csv(p)
    #data_x  =df.iloc[:,6:13]
    data_y = df['ground.truth'].map({'nominal':0,'anomaly':1})
    n_diff.append(data_y.mean())
```

```
100%|
  | 1725/1725 [00:45<00:00, 37.88it/s]
```

250 1500

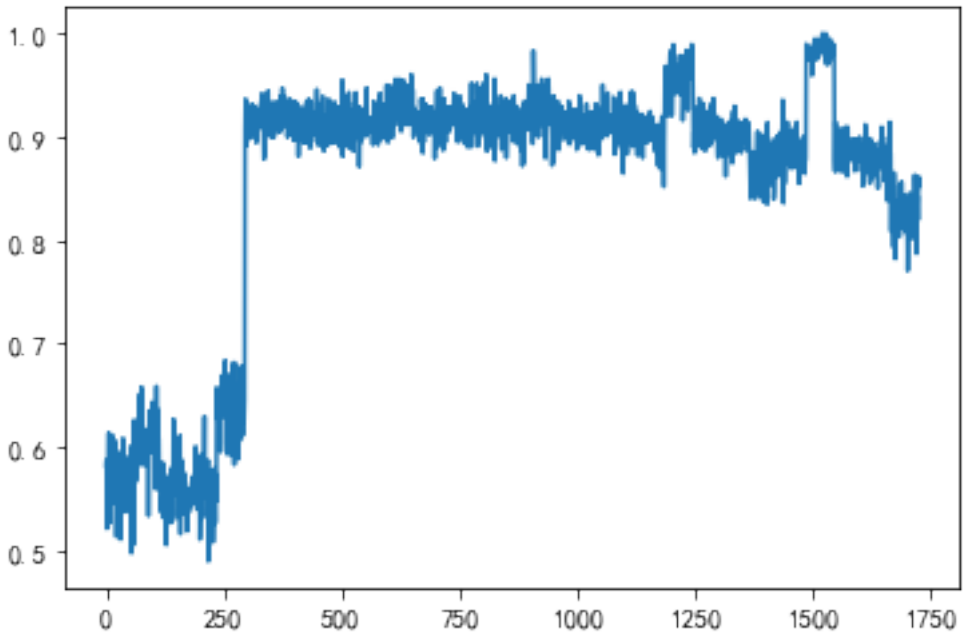
[56]:
```
plt.plot(n_diff)
```

[56]:
```
[<matplotlib.lines.Line2D at 0x1190e5760c8>]
```



[21]:
```
plt.plot(res)
```

[21]: [<matplotlib.lines.Line2D at 0x11905def388>]



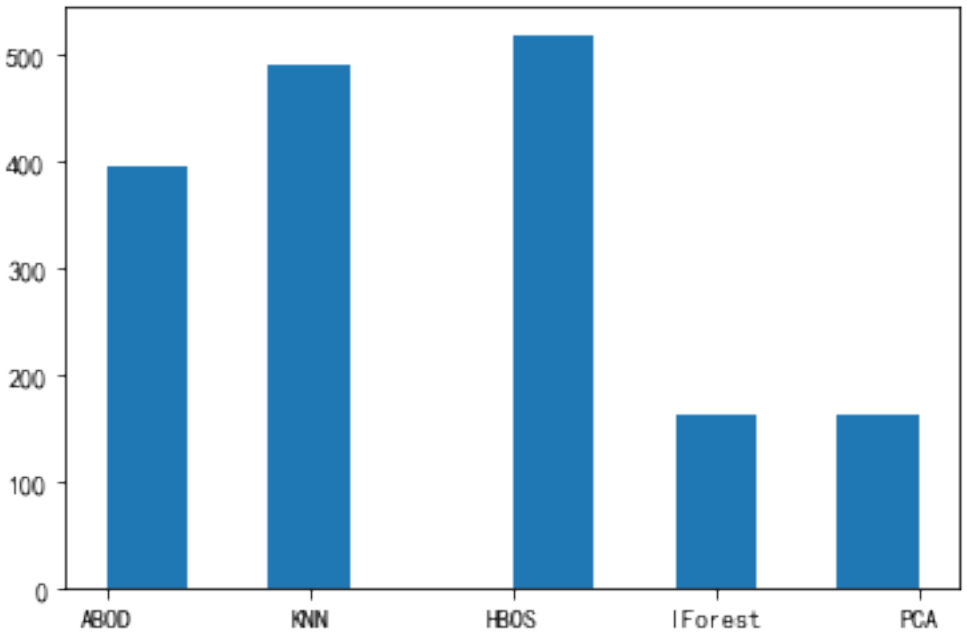
[22]:
```
res =[]
for k,v in scores.items():
    res.append(scores[k])
res = np.array(res)
res =np.argmax(res,axis=0)
```

[63]:
```
for k,v in scores.items():
    print(np.mean(scores[k]))
```

```
0.8382392977552872
0.8410883694311483
0.8211602344962222
0.8395361355371275
0.8270250285372013
```

[48]:
```
plt.hist(pd.Series(res).map({i:k for i ,k in enumerate(scores.
 ↪keys())}),label='ls')
```

[48]:
```
(array([394.,   0., 491.,   0.,   0., 518.,   0., 161.,   0., 161.]),
 array([0. , 0.4, 0.8, 1.2, 1.6, 2. , 2.4, 2.8, 3.2, 3.6, 4. ]),
 <a list of 10 Patch objects>)
```



2 Wine

[57]:
```
path = r'wine\benchmarks'
path_list = glob.glob(os.path.join(path,'*.csv'))
```

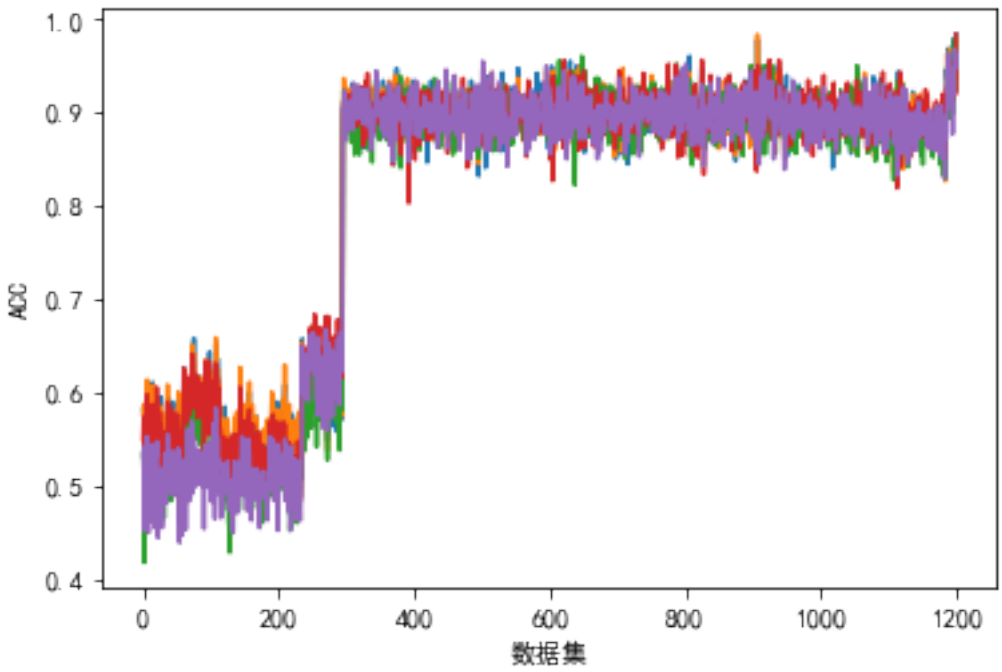
[ ]:
```
res2= []
classifiers = {
     'ABOD'   : ABOD,
     'KNN' :  KNN,
    'PCA':LOF,
    'HBOS':HBOS,
    'IForest':IForest
}
scores2 = {
     'ABOD'   : [],
     'KNN' :  [],
    'PCA':[],
    'HBOS':[],
    'IForest':[]
}

for p in tqdm.tqdm(path_list):
    df = pd.read_csv(p)
    data_x  =df.iloc[:,6:13]
    data_y = df['ground.truth'].map({'nominal':0,'anomaly':1})
    train_x,test_x,train_y,test_y = train_test_split(data_x,data_y,test_size =␣
 ↪0.2)
    contamination = np.mean(train_y)
    for key , value in classifiers.items():
        clf = value()
        clf.fit(train_x)
        #scores2[key].append(pyod.utils.precision_n_scores(test_y.values,clf.
 ↪predict_proba(test_x)[:,1],n=contamination))
        scores2[key].append(np.mean(test_y.values == clf.predict(test_x)))
```

[84]:
```
res =[]
for k,v in scores2.items():
    res.append(scores2[k])
res = np.array(res)
res = np.max(res,axis =0)
res.shape
```

[84]:
```
(1725,)
```

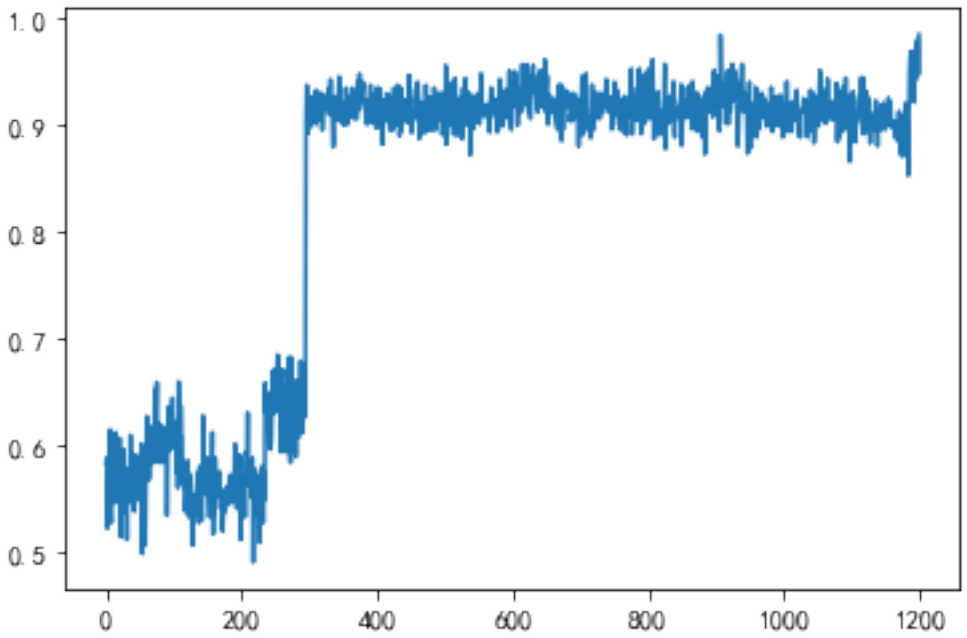
[88]:
```
for k,v in scores2.items():
    plt.plot(scores2[k])
plt.xlabel('  ')
plt.ylabel('ACC')
```

[88]: Text(0, 0.5, 'ACC')



[86]:
```
plt.plot(res[:1200])
```

[86]: [<matplotlib.lines.Line2D at 0x11909785288>]

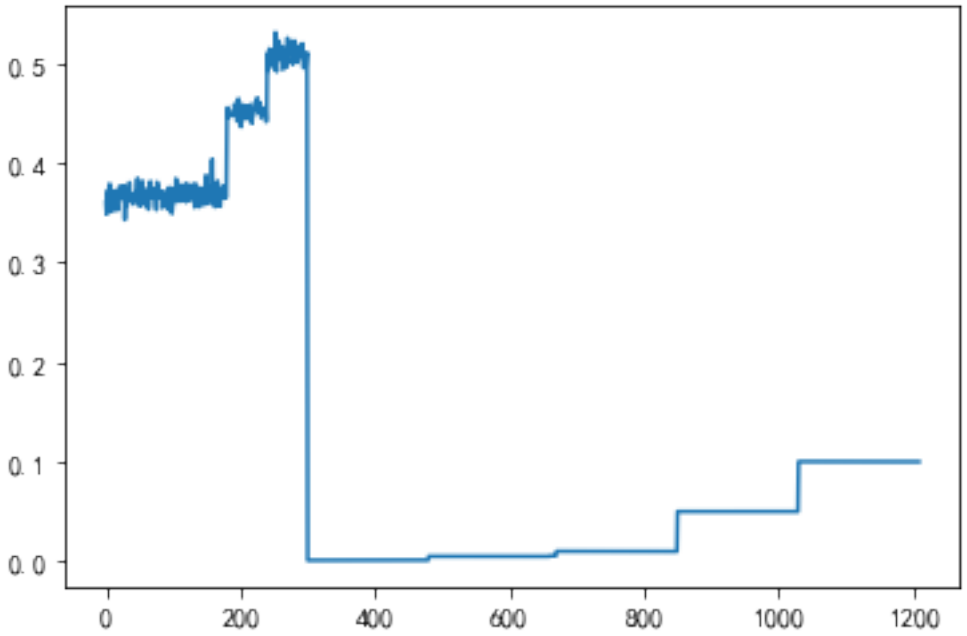


[74]:
```
n_diff =[]
for p in tqdm.tqdm(path_list):
    df = pd.read_csv(p)
    #data_x  =df.iloc[:,6:13]
    data_y = df['ground.truth'].map({'nominal':0,'anomaly':1})
    n_diff.append(data_y.mean())
```

```
100%|
  | 1210/1210 [01:08<00:00, 17.64it/s]
```

[75]:
```
plt.plot(n_diff)
```

[75]:
```
[<matplotlib.lines.Line2D at 0x1190e172c48>]
```



Benchmarks 250 0.5 0.05 wine 200 0.4 0.05

HDOS KNN

0.1 Benchmarks 1500 wine 1000 0.9 250/200 0.56

[ ]: