4 附录

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import pywt
%matplotlib inline
```

In [2]:
```
sns.set_style("darkgrid", {"grid.color": ".6", "grid.linestyle": ":"})
sns.set_theme(font='Times New Roman', font_scale=1.2)
plt.rc("figure", autolayout=True)
# Chinese support
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
```

In [3]:
```
col_names = ["duration","protocol_type","service","flag","src_bytes",
    "dst_bytes","land","wrong_fragment","urgent","hot","num_failed_logins",
    "logged_in","num_compromised","root_shell","su_attempted","num_root",
    "num_file_creations","num_shells","num_access_files","num_outbound_cmds",
    "is_host_login","is_guest_login","count","srv_count","serror_rate",
    "srv_serror_rate","rerror_rate","srv_rerror_rate","same_srv_rate",
    "diff_srv_rate","srv_diff_host_rate","dst_host_count","dst_host_srv_count",
    "dst_host_same_srv_rate","dst_host_diff_srv_rate","dst_host_same_src_port_rate",
    "dst_host_srv_diff_host_rate","dst_host_serror_rate","dst_host_srv_serror_rate",
    "dst_host_rerror_rate","dst_host_srv_rerror_rate","label"]
df = pd.read_csv('./kddcup.data_10_percent_corrected', header=None, names=col_names)
print(df.shape)
# df.describe() # 结果太长不添加在附录中
```

```
(494021, 42)
```

In [4]:
```
df.head()
```

Out[4]:

	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	\
0	0	tcp	http	SF	181	5450	0	
1	0	tcp	http	SF	239	486	0	
2	0	tcp	http	SF	235	1337	0	

```
3         0           tcp    http   SF        219       1337     0
4         0           tcp    http   SF        217       2032     0

   wrong_fragment  urgent  hot  ...  dst_host_srv_count  \
0               0       0    0  ...                   9
1               0       0    0  ...                  19
2               0       0    0  ...                  29
3               0       0    0  ...                  39
4               0       0    0  ...                  49

   dst_host_same_srv_rate  dst_host_diff_srv_rate  \
0                     1.0                     0.0
1                     1.0                     0.0
2                     1.0                     0.0
3                     1.0                     0.0
4                     1.0                     0.0

   dst_host_same_src_port_rate  dst_host_srv_diff_host_rate  \
0                         0.11                          0.0
1                         0.05                          0.0
2                         0.03                          0.0
3                         0.03                          0.0
4                         0.02                          0.0

   dst_host_serror_rate  dst_host_srv_serror_rate  dst_host_rerror_rate  \
0                   0.0                       0.0                   0.0
1                   0.0                       0.0                   0.0
2                   0.0                       0.0                   0.0
3                   0.0                       0.0                   0.0
4                   0.0                       0.0                   0.0

   dst_host_srv_rerror_rate    label
0                       0.0  normal.
1                       0.0  normal.
2                       0.0  normal.
3                       0.0  normal.
4                       0.0  normal.
```

```
        [5 rows x 42 columns]
```

```
In [5]: nums_data = df._get_numeric_data() # 获取定量型数据
        nums_data.values
```

```
Out[5]: array([[0.000e+00, 1.810e+02, 5.450e+03, ..., 0.000e+00, 0.000e+00,
                0.000e+00],
               [0.000e+00, 2.390e+02, 4.860e+02, ..., 0.000e+00, 0.000e+00,
                0.000e+00],
               [0.000e+00, 2.350e+02, 1.337e+03, ..., 0.000e+00, 0.000e+00,
                0.000e+00],
               ...,
               [0.000e+00, 2.030e+02, 1.200e+03, ..., 1.000e-02, 0.000e+00,
                0.000e+00],
               [0.000e+00, 2.910e+02, 1.200e+03, ..., 1.000e-02, 0.000e+00,
                0.000e+00],
               [0.000e+00, 2.190e+02, 1.234e+03, ..., 1.000e-02, 0.000e+00,
                0.000e+00]])
```

```
In [6]: # cA: 近似系数 cD: 细节系数
        # 近似系数: 低频信息 细节系数: 高频信息
        # 低频信息: 整段信号的整体特征 高频信息: 信号中的细节特征
        (cA, cD) = pywt.dwt(nums_data, 'db1', axis=0) # 离散小波变换
```

```
In [7]: cA
```

```
Out[7]: array([[0.00000000e+00, 2.96984848e+02, 4.19738585e+03, ...,
                0.00000000e+00, 0.00000000e+00, 0.00000000e+00],
               [0.00000000e+00, 3.21026479e+02, 1.89080353e+03, ...,
                0.00000000e+00, 0.00000000e+00, 0.00000000e+00],
               [0.00000000e+00, 3.06884343e+02, 2.87368196e+03, ...,
                0.00000000e+00, 0.00000000e+00, 0.00000000e+00],
               ...,
               [0.00000000e+00, 4.18607214e+02, 2.94651396e+03, ...,
                1.41421356e-02, 0.00000000e+00, 0.00000000e+00],
               [0.00000000e+00, 3.49310750e+02, 1.69705627e+03, ...,
                1.41421356e-02, 0.00000000e+00, 0.00000000e+00],
               [0.00000000e+00, 3.09712770e+02, 1.74513954e+03, ...,
                1.41421356e-02, 0.00000000e+00, 0.00000000e+00]])
```

```
In [8]: cD
```

```
Out[8]: array([[   0.        ,  -41.01219331, 3510.07806181, ...,    0.        ,
                   0.        ,    0.        ],
               [   0.        ,   11.3137085 ,    0.        , ...,    0.        ,
                   0.        ,    0.        ],
               [   0.        ,    0.        ,    0.        , ...,    0.        ,
                   0.        ,    0.        ],
               ...,
               [   0.        ,   19.79898987, -286.37824638, ...,    0.        ,
                   0.        ,    0.        ],
               [   0.        ,  -62.22539674,    0.        , ...,    0.        ,
                   0.        ,    0.        ],
               [   0.        ,    0.        ,    0.        , ...,    0.        ,
                   0.        ,    0.        ]])
```