DataCon
大数据安全分析比赛

DataCon
大数据安全分析比赛

之二恶意代码行为检测

队名：loom

演讲者：杨航锋

广州大学 GUANGZHOU UNIVERSITY

方滨兴班

团队介绍

DataCon
大数据安全分析比赛

李树栋
广州大学
百人计划/副教授
指导老师

杨航锋（队长）
广州大学
方滨兴院士班
研一在读

范美华
广州大学
方滨兴院士班
研一在读

陈园园
广州大学
方滨兴院士班
研一在读

付潇鹏
广州大学
方滨兴院士班
研一在读

DataCon
大数据安全分析比赛

目录
contents

1 赛题背景介绍

2 特征工程

3 算法的设计和结果展示

4 可改进的方向

DataCon
大数据安全分析比赛

01 | 赛题背景介绍

赛题背景介绍

DataCon 大数据安全分析比赛

随着大数据和人工智能时代的到来，催生了很多新的安全技术方法，包括人工智能、机器学习、可视化分析等，这些技术方法有别于传统的安全检测等技术。恶意代码行为检测方向设有两个赛题分别为：恶意代码识别、恶意代码家族聚类。


```
- <action ret_value="0" call_time="18:45:42.006" call_pid="396" call_name="1.exe" api_name="TryToAnalyze">
    <apiArg_list count="0"/>
  - <exInfo_list count="1">
      <exInfo value="C:\program\1.exe"/>
    </exInfo_list>
  </action>
- <action ret_value="0" call_time="18:45:42.007" call_pid="396" call_name="1.exe" api_name="UnpackSelf">
    <apiArg_list count="0"/>
    <exInfo_list count="0"/>
  </action>
- <action ret_value="0" call_time="18:45:42.020" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\user32.dll"/>
    </exInfo_list>
  </action>
- <action ret_value="0" call_time="18:45:42.022" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\gdi32.dll"/>
    </exInfo_list>
  </action>
```


某软件执行时产生的动作序列节选

Model

- 是否为恶意代码
- 所属恶意代码家族

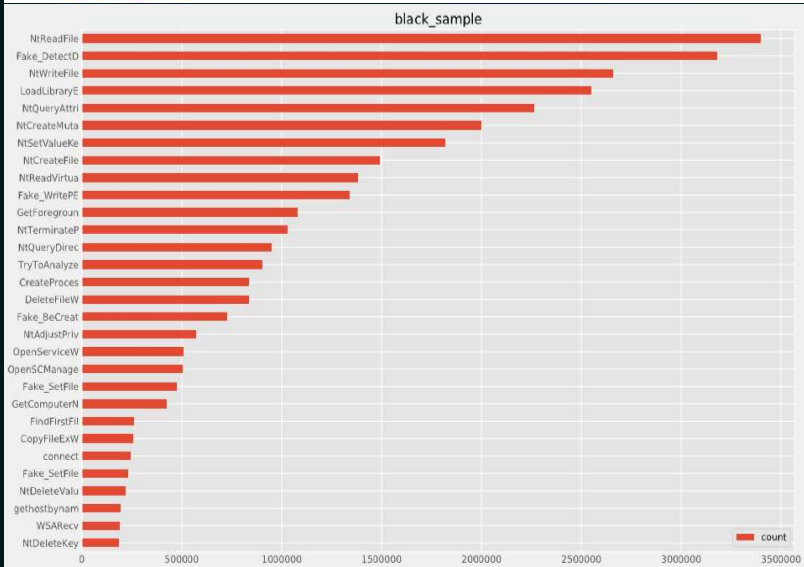
DataCon
大数据安全分析比赛

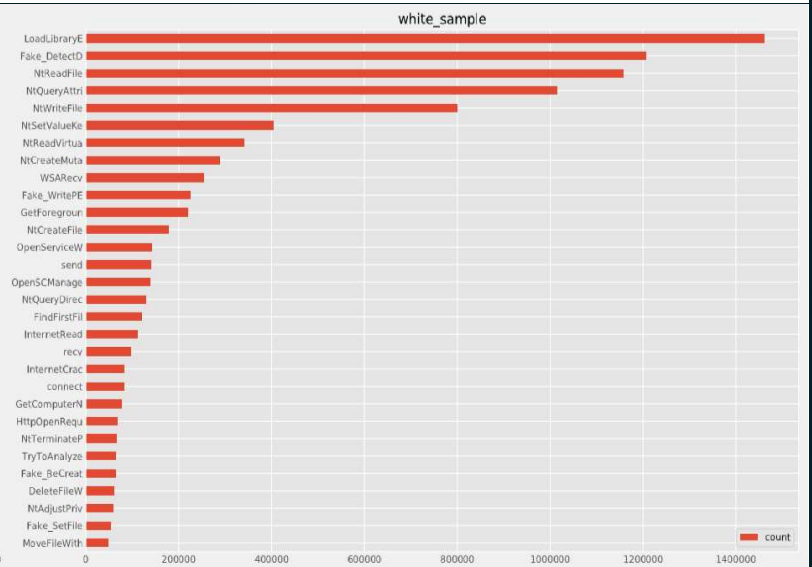
02 | 特征工程

特征工程之数据探索

DataCon
大数据安全分析比赛


```
- <action ret_value="0" call_time="18:45:42.020" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\user32.dll"/>
    </exInfo_list>
  </action>
```

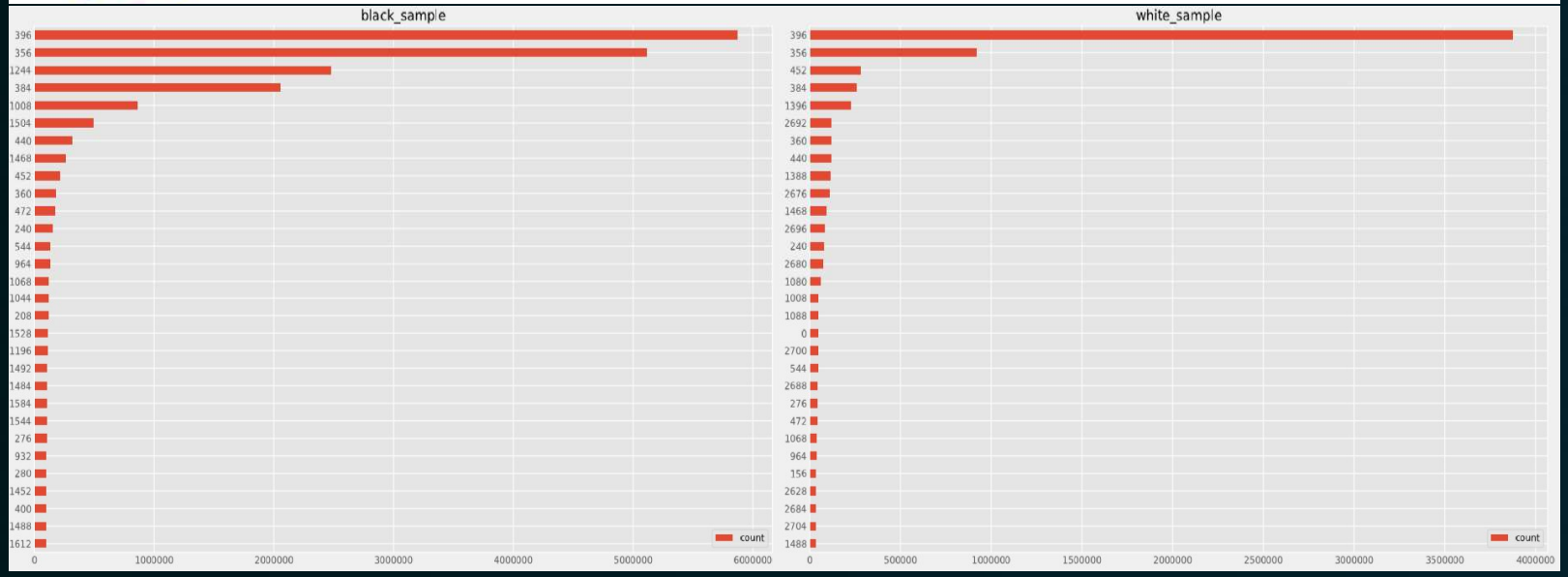


探索训练样本中黑白样本api_name字段的差异

广州大学
GUANGZHOU UNIVERSITY
方滨兴班

特征工程之数据探索


```
- <action ret_value="0" call_time="18:45:42.020" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\user32.dll"/>
    </exInfo_list>
  </action>
```

探索训练样本中黑白样本call_pid字段的差异

DataCon 大数据安全分析比赛

- 在数据探索分析阶段发现，只要程序调用了pid为1244的进程，那么该程序一定是恶意软件；
- 后续细分后发现当程序调用了pid 为1396的进程时，如果特征出现的次数为254、256、257、258、259、262时必为正常程序。

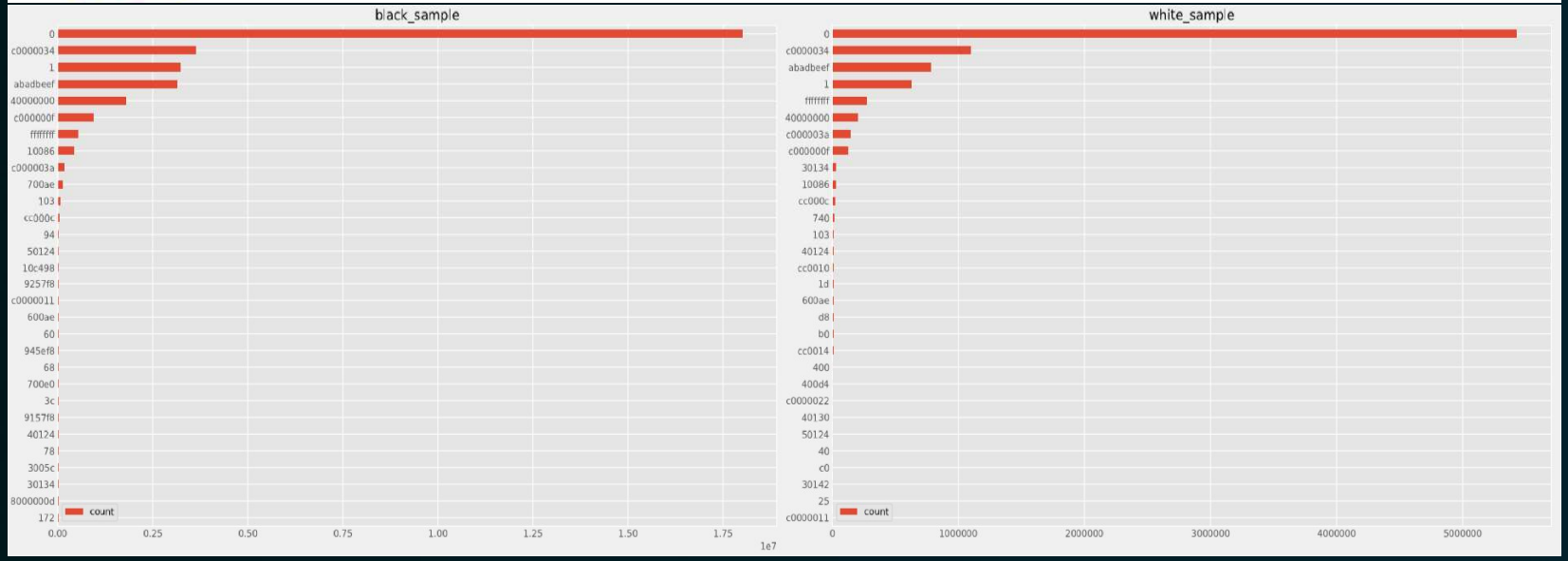
广州大学 GUANGZHOU UNIVERSITY

方滨兴班

特征工程之数据探索

DataCon 大数据安全分析比赛

```
- <action ret_value="0" call_time="18:45:42.020" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\user32.dll"/>
    </exInfo_list>
  </action>
```



探索训练样本中黑白样本ret_value字段的差异

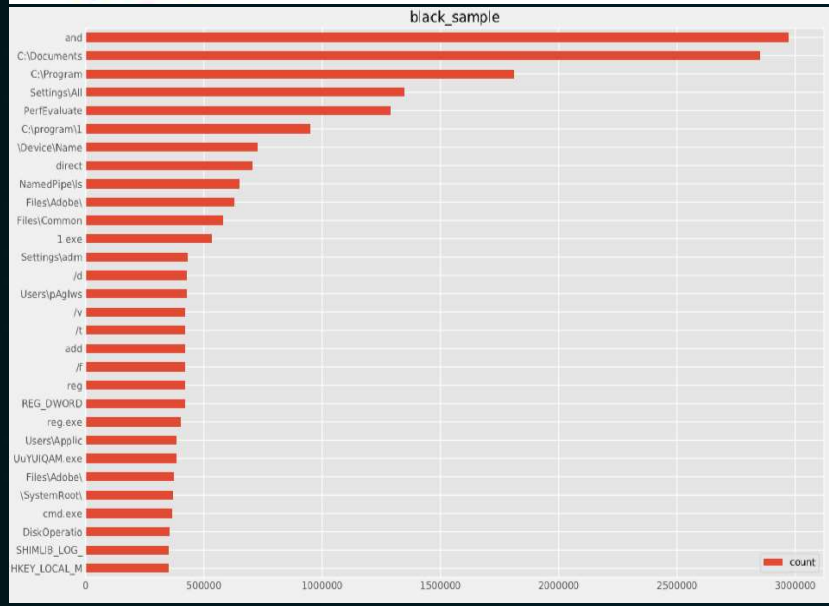
广州大学 GUANGZHOU UNIVERSITY

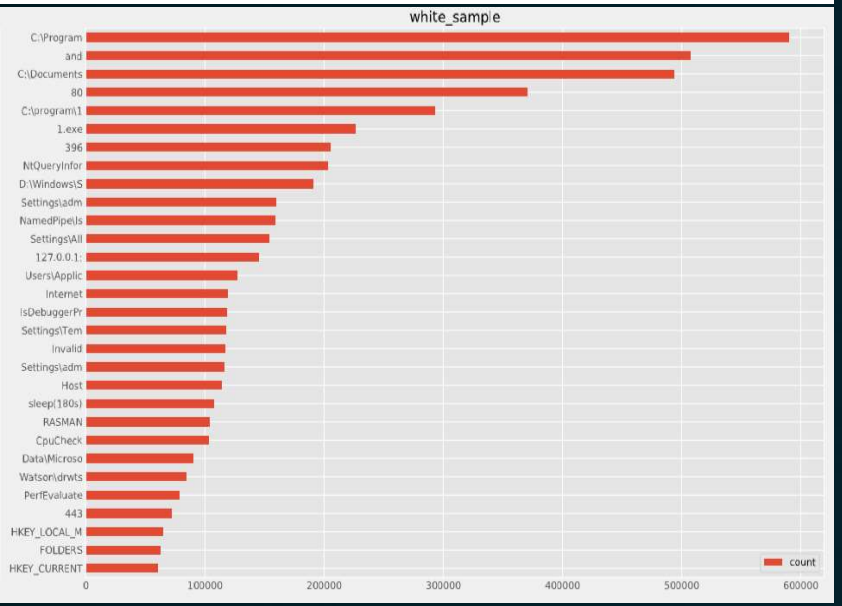
方滨兴班

特征工程之数据探索

DataCon 大数据安全分析比赛

```
- <action ret_value="0" call_time="18:45:42.020" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\user32.dll"/>
    </exInfo_list>
  </action>
```





- 在数据探索分析阶段发现exinfos字段中如果特征reg出现的次数大于等于2；又或者特征add出现的次数大于等于6时，该程序必为恶意程序。

探索训练样本中黑白样本exinfo_value字段的差异

特征工程

DataCon 大数据安全分析比赛

- 特征工程
  - TFIDF (n_gram=(1,5))
  - 统计特征
    - 文件执行了多少种不同的api
    - 文件总共执行了多少个api
    - 文件调用进程的个数
    - 文件调用了了多少种不同的进程
    - 文件进程返回值为0的个数
    - ……
    - 文件进程返回值为0的比例

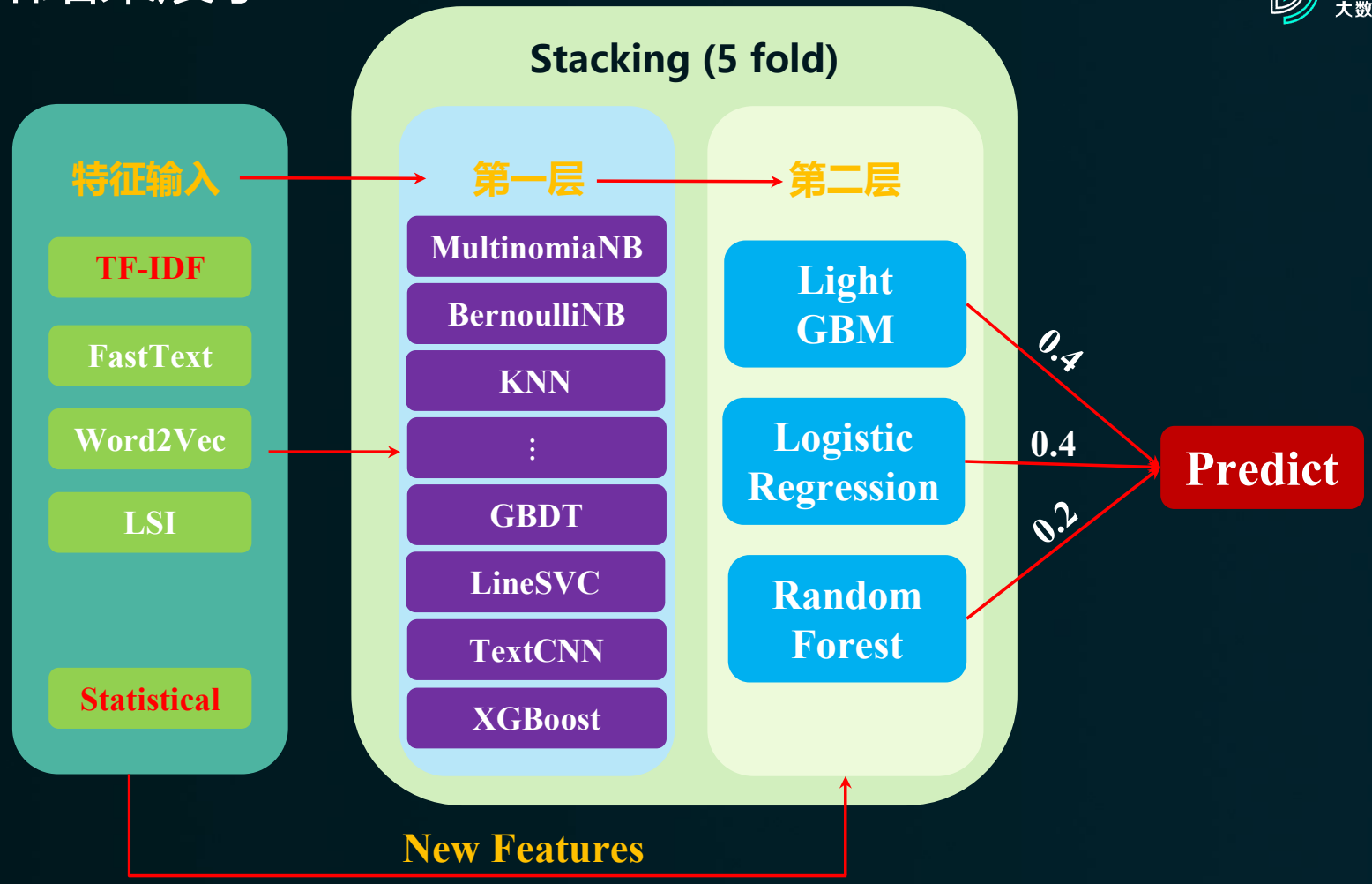
DataCon
大数据安全分析比赛

03 | 算法的设计和结果展示

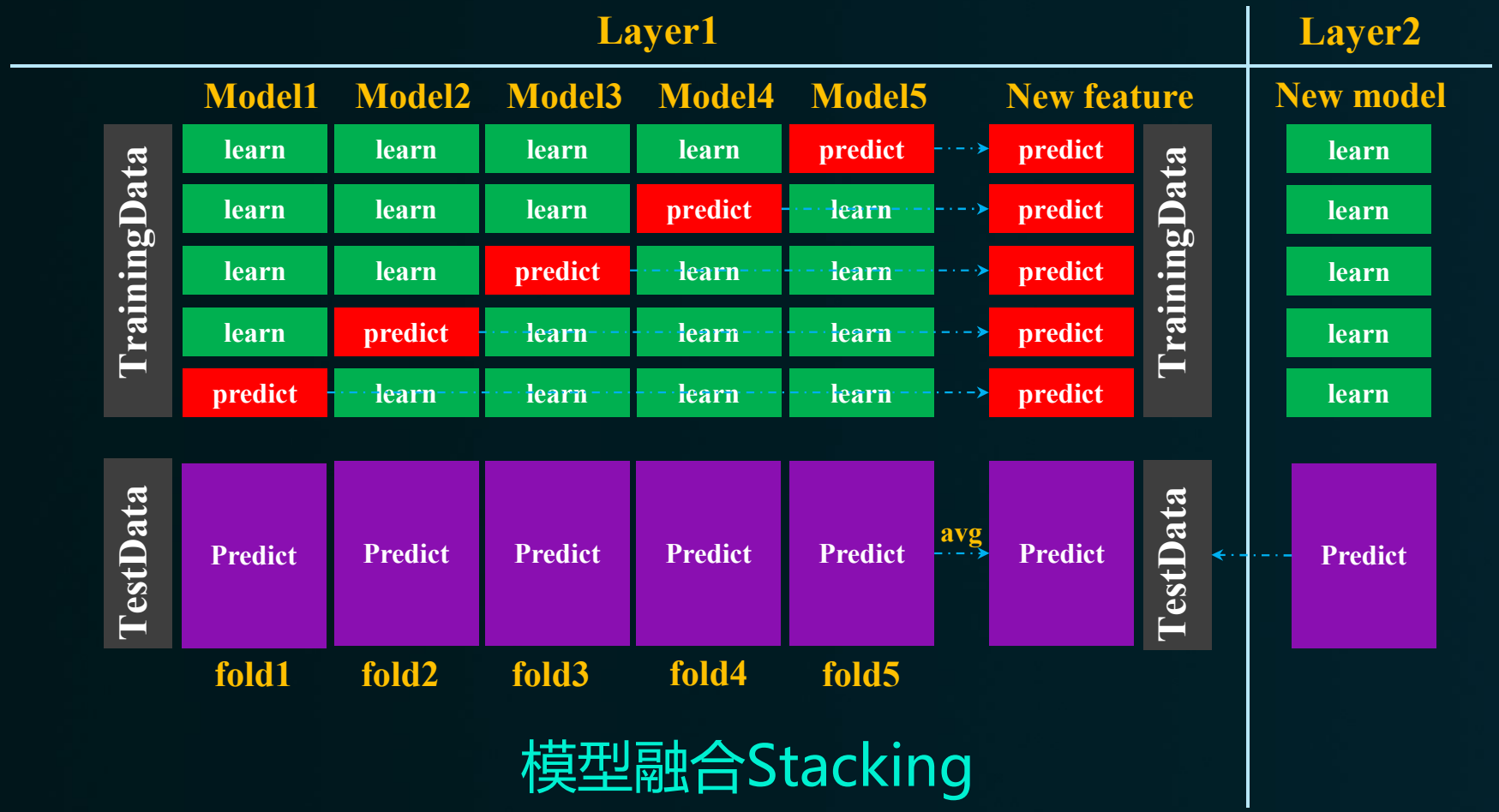
算法的设计和结果展示

DataCon 大数据安全分析比赛

恶意代码识别模型的构建



算法的设计和结果展示



模型融合Stacking

算法的设计和结果展示

DataCon 大数据安全分析比赛

算法名称	准确率	召回率	F1 Score
Random Forest	0.9800	0.9800	0.9800
XGBoost	0.9800	0.9700	0.9800
BPNN	0.9635	0.9635	0.9635
Decission Tree	0.9500	0.9600	0.9500
Logistic Regression	0.9600	0.9700	0.9400
Naive Byes[1]	0.9400	0.9500	0.9400
Naive Byes[2]	0.9484	0.9484	0.9484
GBDT	0.9700	0.9500	0.9600
Bagging	0.9700	0.9600	0.9700
AdaBoost	0.9521	0.9521	0.9521
SVM	0.9700	0.9500	0.9600
Ensemble model[1]	**0.9839**	**0.9839**	**0.9839**
Ensemble model[2]	**0.9967**	**0.9967**	**0.9967**

```
random_forest_model.fit(train_data, train_label)
y_pred = random_forest_model.predict(test_data)
```

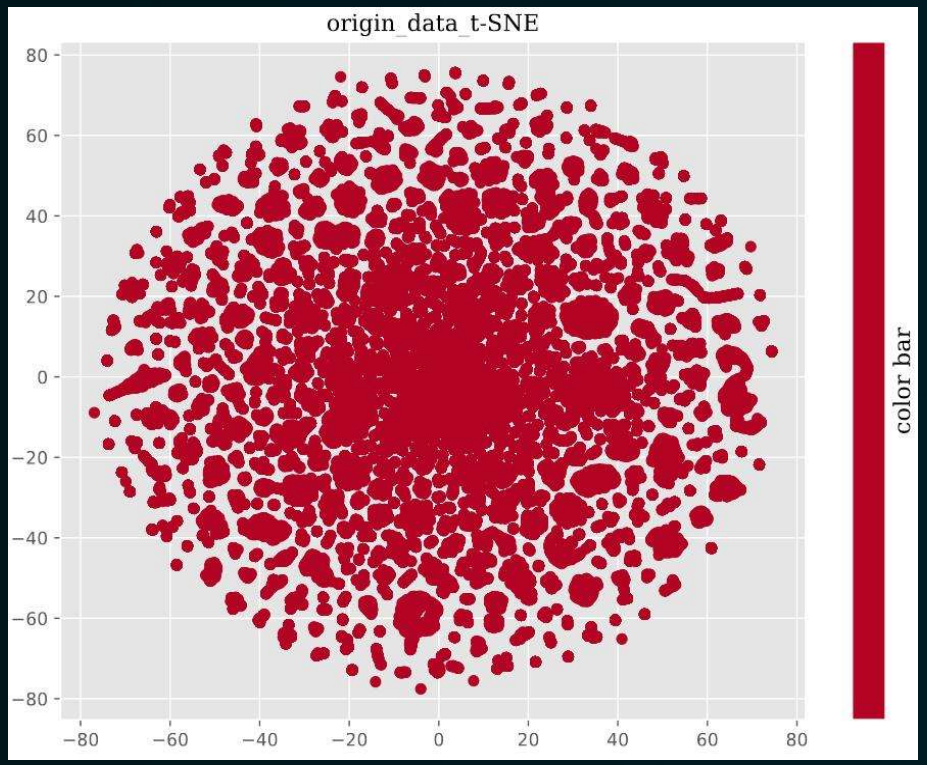
```
print(classification_report(test_label, y_pred,
                            target_names=["white", "black"]))
```

```
              precision    recall  f1-score   support

       white       0.98      1.00      0.99      4023
       black       0.99      0.95      0.97      1977

   micro avg       0.98      0.98      0.98      6000
   macro avg       0.98      0.97      0.98      6000
weighted avg       0.98      0.98      0.98      6000
```

恶意代码识别结果展示

算法的设计和结果展示

DataCon 大数据安全分析比赛



可视化调参

特征输入

TF-IDF

FastText

Word2Vec

LSI

Statistical

构建模型

K-means

层次聚类

DBSCAN

聚类结果

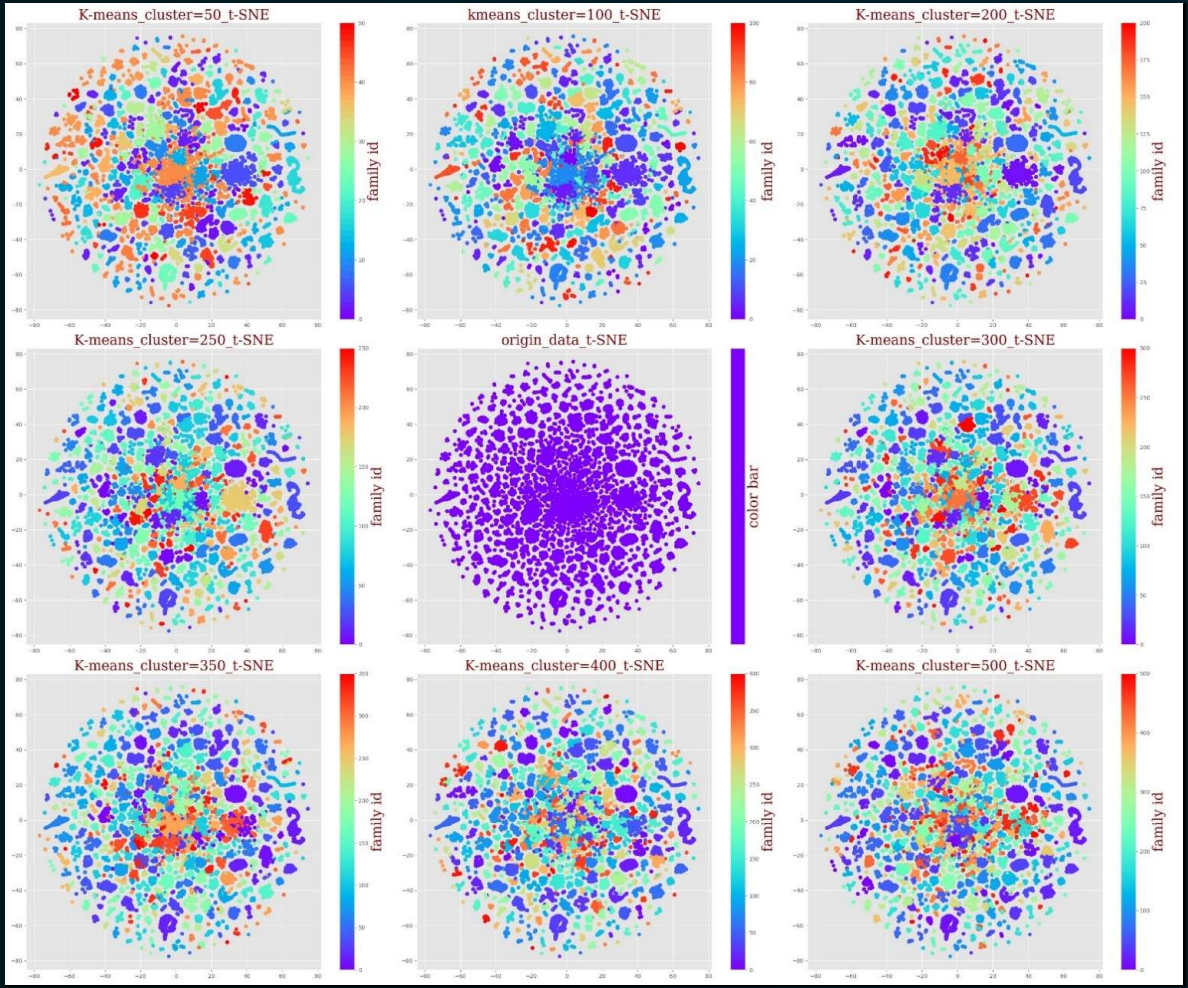
恶意代码家族聚类模型的构建

广州大学 GUANGZHOU UNIVERSITY

方滨兴班

算法的设计和结果展示

DataCon 大数据安全分析比赛



恶意代码家族聚类结果展示

DataCon
大数据安全分析比赛

04 | 可改进的方向

可改进的方向

DataCon 大数据安全分析比赛


```
- <action ret_value="0" call_time="18:45:42.006" call_pid="396" call_name="1.exe" api_name="TryToAnalyze">
    <apiArg_list count="0"/>
  - <exInfo_list count="1">
      <exInfo value="C:\program\1.exe"/>
    </exInfo_list>
  </action>
- <action ret_value="0" call_time="18:45:42.007" call_pid="396" call_name="1.exe" api_name="UnpackSelf">
    <apiArg_list count="0"/>
    <exInfo_list count="0"/>
  </action>
- <action ret_value="0" call_time="18:45:42.020" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\user32.dll"/>
    </exInfo_list>
  </action>
- <action ret_value="0" call_time="18:45:42.022" call_pid="396" call_name="1.exe" api_name="LoadLibraryExW">
  - <apiArg_list count="3">
      <apiArg value="0"/>
      <apiArg value="0"/>
      <apiArg value="0"/>
    </apiArg_list>
  - <exInfo_list count="1">
      <exInfo value="C:\WINDOWS\system32\gdi32.dll"/>
    </exInfo_list>
  </action>
```


某软件执行时产生的**动作序列**节选

1. 把xml格式的文件用二进制读取后转化成图片格式；
2. 使用卷积神经网络抽取图片特征；
3. 把图片特征与之前提取的文本特征做特征融合。

按二进制读取xml文件后转化为**图片形式**

DataCon
大数据安全分析比赛

Thank you!

广州大学
GUANGZHOU UNIVERSITY

方滨兴班