大数据比赛-恶意代码识别分赛第一题说明文档

1. 第一题题目设计说明

面对大量的 PC 恶意病毒的动态行为进行识别，进行分类判别，即判断一个 exe 文件经过沙箱运行后输出的 xml 文件来判断该文件是否为恶意程序。

分析程序的种类，需要完成：根据训练集样本，对测试集样本进行是否是恶意样本的判断。

Safe_type: 安全结果，恶意程序为 1，正常程序为 0

2. 题目数据的下发形式及文件大小（初赛数据已完成，决赛数据还在进行筛选）

（1） 数据集说明（第一题）：

约 45000 条样本数据，压缩包 md5 值为 210d2b05a3a5afc3f1458d408f802820；训练数据集约为 30000 条（其中黑样本 10000 条，白样本 20000 条，并包含样本 ground truth 标签）测试数据集为 15000 条，文件类型为.xml。

3. 总体题目结果提交形式及文件大小

（1） 第一题比赛期间提交识别结果 CSV 文件（CSV 文件见示例），最后提交算法程序

（2） 最终比赛成绩结果按 ScoreF= 0.5*Score1 题目一 + 0.5*Score2 题目二进行排名，并提交第一题及第二题的算法设计说明 PPT

（3） 经评委对决赛成绩进行审核后，确认无作弊及其他异常问题后，最终颁布决赛排名成绩

4. 题目评判规则

（1） 评分规则

a. **第一题**

$$Score1 = \frac{1}{N}\sum_{k=1}^{N}([ST_k = TrueST_k] - (ST_k - TrueST_k)\cdot ST_k)$$

N: 测试集样本数量

ST_k：第 k 个样本的 Safe_type 预测值

$TrueST_k$：第 k 个样本的 Safe_type 标准值

[ ]：判定成立为 1，判定不成立为 0

Safe_type 判断正确，可得 1 分；

Safe_type 判断不正确时。若属于误报，即白样本判断为恶意样本，扣 1 分。若属于漏报，即恶意样本判断为白样本，不得分也不扣分。

备注：更看重对误报的惩罚，参赛者提交的是非 0 即 1 的硬分类结果。

（2） 提交说明：第一题比赛每日可提交一次，取截至日期前得分最高成绩； 决赛成绩按截至日期前最优成绩排名确定（进入前三名的队伍最终的名次确定会参考一定比例的算法设计思路）；

5. 备注

（1） 第一题文件样例：

```
<report report_id="EAE38ADA" report_version="2">
  <file_list file_uid="784413" file_name="0a0a237585c5b0ca7a258f1a6e1495488e89f77b88b59ce7879516c45aec3b99.exe" file_error="0">
    <file>
      <start_boot>
        <field value="start_boot"/>
        <action_list>
          <action name="AnalyzeStart" api_name="AnalyzeStart" ret_value="0" call_pid="0" call_name="" call_time="19:49:49.000">
            <apiArg_list count="0"/>
            <exInfo_list count="0"/>
          </action>
          <action name="VasInfo" api_name="VasInfo" ret_value="0" call_pid="0" call_name="" call_time="19:49:49.001">
            <apiArg_list count="0"/>
            <exInfo_list count="3">
              <exInfo value="18.2"/>
              <exInfo value="60"/>
              <exInfo value="2019-02-19_19:49:47==2019-02-19_19:49:49"/>
            </exInfo_list>
          </action>
          +<action name="BeCreated" api_name="Fake_BeCreated" ret_value="0" call_pid="212" call_name="vaslauncher.exe" call_time="19:49:50.002"></action>
          +<action name="BeCreatedEx" api_name="Fake_BeCreatedEx" ret_value="0" call_pid="212" call_name="vaslauncher.exe" call_time="19:49:50.003"></action>
          +<action name="BeCreated" api_name="Fake_BeCreated" ret_value="0" call_pid="396" call_name="1.exe" call_time="19:49:54.004"></action>
          +<action name="BeCreatedEx" api_name="Fake_BeCreatedEx" ret_value="0" call_pid="396" call_name="1.exe" call_time="19:49:54.005"></action>
          +<action name="TryToAnalyze" api_name="TryToAnalyze" ret_value="0" call_pid="396" call_name="1.exe" call_time="19:49:54.006"></action>
```

（2） 第一题提交文件样例（csv 文件）：

id	safe_type
53f448f9db9dc1ffb7f55d2b3a189609fd58ce80ed7958890d3d871565fa2be7	0
80f5672e7a581b44d2cce92d1c27e9e4ad81c42e0a42e11c0f17e5e51669b660	0
439626eab2205758bf1a72e7623cee250b796a554535bd7fc27e6f1ccca1e16a	0
1e108525d48c999db9fce05bc24b8047ac28e73ef0adb4348dd55fe16b308274	0
82a966ad38c24f2fc66cb8ed2c163b7390b6e04077e2e7dee45693366854f177	0
5ea41aefce38908ad3ab2a4e5ad95a5b525a3d6448aaf1a24e1115e04c205aee	0
3f214eb4549c6c983374d4c96dd0ea25ee0469523e8bc8be75218b8aecaafe91	0
fffebbb09aeb89235b9704f335d3d3e6fce0db4f31121430e408355da2f9ad34	0
094ac45236fc0267c087cb9eb1a6d8851f5863facb5126d0c2a87f503a2d18d9	0
28b64cb73d08acd0735146b7665d8e0ef7668be014941e6c59c4d6b7ea5e0b38	0