二〇一九年一月十日星期四

下午3时51分

由于自己之前的客观行为检测实验使用的是CERT4.2数据集的1000个用户，而18年9月份以来做主观检测实验不知道为什么用的竟然是CERT5.2数据集的2000个用户，结果就是，现在需要融合两类结果，首要任务是将主观检测的实验移植到CERT4.2上，并且希望结果一样有效。

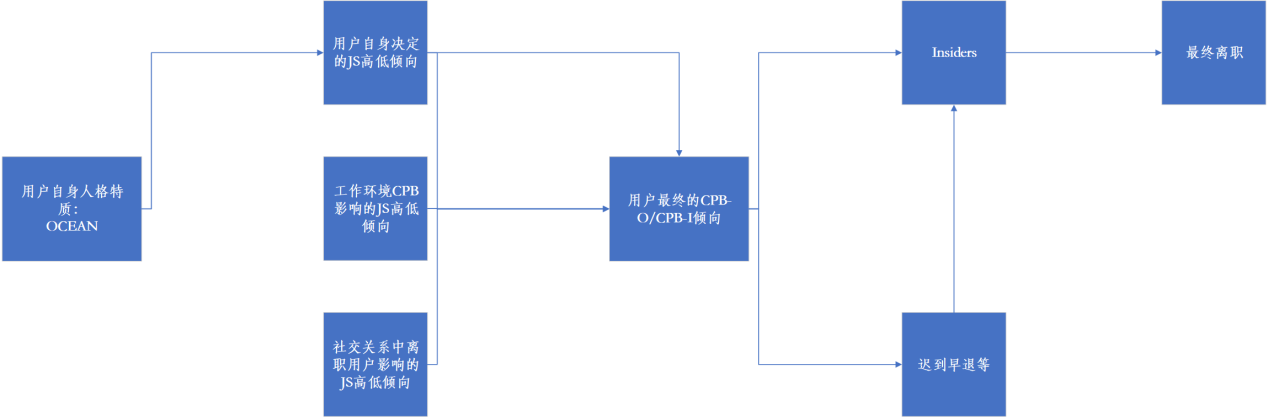
借着这个机会，我们将整体的主观检测实验过程过一遍，从而为日后融合实验与写作论文打好基础。

1. 心理模型

整体的心理模型基于几个重要关联：

1. 用户反生产行为与Insiders攻击的直接联系；
2. 用户的大五人格与反生产行为倾向的联系（个体人格直接影响+工作满意度的间接影响）

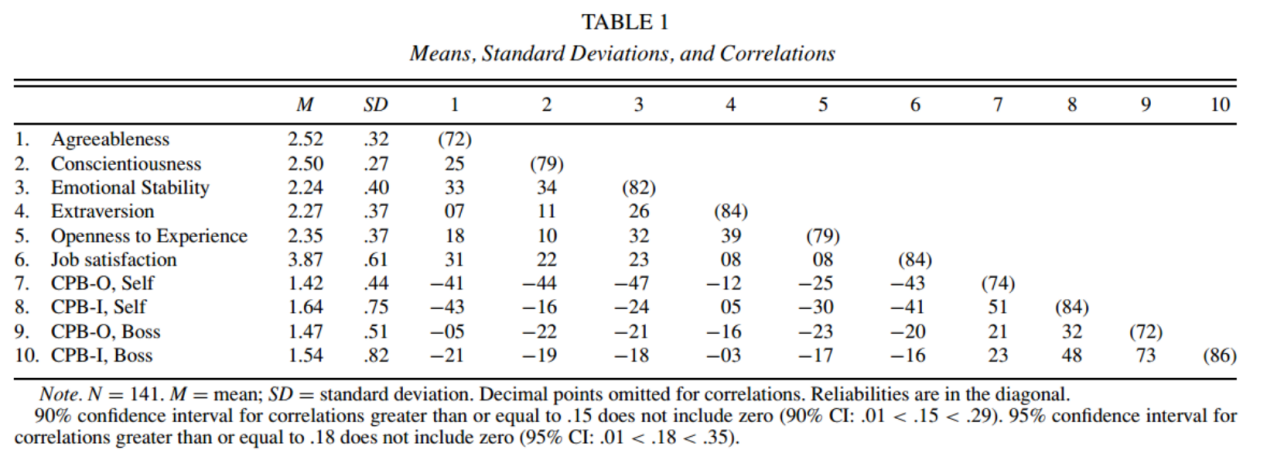
考虑内部攻击者攻击倾向的刻画，一种来自于自身人格因素决定的反生产行为倾向（CPB-I/CPB-O），一种来自于工作满意度的侧写；工作满意度的侧写又需要从自身人格决定的基本满意度状态（OCEAN:JS）、团队环境积极友善程度（团队虚拟人格的CPBs）、直接领导者的CPBs、缺勤情况统计以及与离职用户的通信联系情况（虚拟离职联系人的Email\_Info）来共同表述。



训练OCSVM遍历参数空间以确定最优参数空间（召回率优先，其次是误报率，或使用公式，上述公式综合考虑了召回率与误报率的重要性，默认x=0.9，实际实验可以依据结果修正）。实际实验中选择了x=1.0的Recall最高原则筛选；

需要声明的一点，最初的想法是在训练第N个月的OCSVM时，参数选择过程的验证集包含既往1:N-1个月离职用户的ATF特征，如此需要ATF中涉及数量的尽量转换成均值；而结果中由于设定了以召回率优先，会导致得到的OCSVM中FP过高；因此，暂时不考虑将验证过程引入上个月之前的离职用户ATF，但是ATF中涉及到虚拟人格的暂时全部均值化以统一处理方法。

此外，此次计算ATF由于需要计算JS分数，因而建议使用参考文献[10]中的完备结果，其中与CPB的关系建议采用用户自我评价，原因是自己更清楚真实的感受，而满意度无疑是决定于自我感觉的；其次可以发现关于CPB-O与C的关联性个体评价更显著，**因此使用个体评价取代Boss评价。**

为此，可能需要重新计算CERT5.2中各个用户团队样本中的CPB分数，并且以此计算JS分数；  


最终，我们的ATF特征应具有以下字段

|  |  |
| --- | --- |
| 自身人格特质因子 | OCEAN五个分数(5)[实际使用中删除了] |
| 反生产行为倾向CPB-I/O (2) |
| 工作满意度因子 | 自身OCEAN映射的JS分数（1） |
| Team中CPBs状态：(8)  Team\_CPB-I-mean,Team\_CPB-O-mean,  Users-less-mean-A,Users-less-mean-A and C,  Users-less-mean-C,Users-High-mean-N,  Team\_CPB-I-median,Team\_CPB-O-median |
| Leader的CPBs(2) |
| Leave\_Contacts\_EmailInfo:(9)  需全部转变为均值/len(Leave\_Contacts)  dis\_ocean, avg\_dis\_ocean, dis\_os, avg\_dis\_os, email\_ratio, cnt\_send/recv, cnt\_s/r\_size, cnt\_s/r\_attach, cnt\_s/r\_days, cnt\_email\_days |
| 缺勤特征（3）  cnt\_late\_days, cnt\_early\_days, //转变为比例  month\_work\_days, |

1. 实验数据准备

本部分主要说明如何一步一步构建主观检测实验所需要的用户主观特征ATF。

1. 用户工作环境因子影响的CPB因素

直接使用之前提取的CERT4.2-2009-12-New-26JS.csv文件，自动选取其中Team\_CPB与Leader\_CPB部分即可；

1. CERT4.2的基本数据集

包括Personality/LDAP-OS结构/Insiders的GT数据

由于自己编写的程序代码过于混乱，于是重新，从头对于主体实验v01-v09的全部实验文件进行注释说明

|  |  |  |
| --- | --- | --- |
| 文件名 | 模块 | 功能 |
| V09-LDAP\_AllUsers.py | 出勤率 | 获取CERT4.2中的四元部门及对应用户列表，便于进行后续的dis\_OS特征提取 |
| V09\_Extract\_Logon.py | 出勤率 | 顾名思义：将统一的Logon数据按照2010-01月份不同的用户进行梳理归类，便于计算用户的习惯上下班时间--Team上下班时间--用户迟到早退特征 |
| V09\_Extract\_WorkTime.py | 出勤率 | 提取用户的习惯上下班时间 |
| V09\_Extract\_Team\_WorkTime.py | 出勤率 | 提取用户所在Team的标准上下班时间 |
| V09\_Check\_Early\_Late.py | 出勤率 |  |
|  |  |  |
| V09-LaidOff Insiders.py | 离职用户列表 | 1. 统计CERT4.2中的Leave用户，并写入文件（需要前置获取用户最后一次登录的日期天） 2. 按用户提取各自的Logon数据，并写入Logon目录 |
| CERT4.2-Leave-Users\_OnDay\_0.9.csv | 离职用户列表 | 上述离职用户提取的结果，2010-01:2011-04的离职用户、最后一天工作日期以及其LDAP |
|  |  |  |
| AAE0190\_LCE\_Feats.csv | LCE特征 | 单个用户与Leave Contacts通信的Email四元特征 |
| CERT4.2\_Users\_LeaveContacts\_EmailFeats.csv | LCE特征 | 所有用户集合的按月份排列的LCE特征 |

二〇一九年一月十一日星期五

上午8时55分

CERT4.2复现实验的数据验证：

1. V09\_Check\_Early\_Late.py验证2010-01

*Format: User\_id, workon time, workoff time, cnt\_late, cnt\_early, all work days, the days with late and early at the same time*

NGF0157,7.0,17.0,8.0,4.0,23,3,2010-01-04,2010-01-22,2010-01-29,

验证通过！在中心台式机上开始CERT4.2用户的全面出勤率统计

1. CERT4.2用户邮件内容归档

数据格式：

Fields: id, date, user, pc, to, cc, bcc, from, size, attachment\_count, content

{R3I7-S4TX96FG-8219JWFF},01/02/2010 07:11:45,LAP0338,PC-5758,Dean.Flynn.Hines@dtaa.com;Wade\_Harrison@lockheedmartin.com,Nathaniel.Hunter.Heath@dtaa.com,,Lynn.Adena.Pratt@dtaa.com,25830,0,middle f2 systems 4 july techniques powerful destroyed who larger speeds plains part paul hold like followed over decrease actual training international addition geographically side able 34 29 such some appear prairies still 2009 succession yet 23 months mid america could most especially 34 off descend 2010 thus officially southward slope pass finland needed 2009 gulf stick possibility hall 49 montreal kick gulf

验证通过！

1. 离职用户统计

CERT4.2的用户数据相比CERT5.2不规范，尤其是2011-05月的审计数据竟然不是到2011-05-31

需要先判断一次CERT4.2的所有Insiders的最后一次离职时间是多少？

为此，决定先对所有用户的最后登录时间进行升序排序，查看最后登录的时间；

经过分析：

CERT4.2中共有154个用户的最后一次登录时间在2011-05之前，最晚的一次是在

FSB0399,2011-04-29,

NAF0033,2011-05-12,

CEL0561,2011-05-16,

而2011-05最后一次登录系统的有1000-154=846个用户。故使用前154个用户作为Leave用户，且Insiders\_1/2/3均在2011-05前离职，因此，直接只考虑2010-01:2011-04的离职情况即可；

最终得到了所需的CERT4.2用户离职文件数据：CERT4.2-Leave-Users\_OnDay\_0.9.csv

其中包括154个离职用户（不考虑2011-05月份）

二〇一九年一月十一日星期五

下午4时37分

终于静下心来写了提取四元LCE特征的程序，运行一次测试验证：

User: AAE0190 确实是两个，日期也对，验证通过，开始1000用户的提取

Lcontact:Buffy.Vivien.Nichols@dtaa.com

2010-06-08,BVN0514,2.0['2010-05-18'; '2010-06-04'],66643.0,0.0

Line 1303: 05/18/2010 15:42:23,

AAE0190,PC-8915,Buffy.Vivien.Nichols@dtaa.com,Whoopi.Britanney.Camacho@dtaa.com,,August.Armando.Evans@dtaa.com,31885,0,

Line 1478: 06/04/2010 18:22:36,

AAE0190,PC-8915,Callum.Scott.Parrish@dtaa.com,Sawyer.Abel.Turner@dtaa.com;Buffy.Vivien.Nichols@dtaa.com;Eugenia.Whilemina.Harrington@dtaa.com,,August.Armando.Evans@dtaa.com,34758,0,

1. 构建均值化的LCE特征/每个月

现在就差最后一步就可以直接计算每个月的ATF和静态ATF了！

然后就可以重新实验了！

但是前提是将每个月在职用户的LCE计算出来：

副产品：

1. 得到每个月的离职用户列表以及时间+LDAP

下午10时10分

开始改写之前的程序，由所有用户的LCE特征生成2010-01:2011-04每月的用户的ATF特征以及每月的离职用户列表。

每月的ATF中关键的LCE部分是截止到当月所有用户（包含当月的离职用户）的数据的，如此才可以实现按月的Risk预测，如知晓了第N月的离职用户的ATF，才能对其进行高危预测；

最初的CERT42\_Users中包含所有用户，进入到2月份时去掉1月份离职的用户，如此类推；

先生成了2010-02的LCE特征，验证：

CEL0561,40.0,39.0,36.0,19.0,40.0,-7.81,-20.47,45.6043474837,22.8021737418,30.0,15.0,**21.0,658953.0,24.0,16.0,**

2010-02-11,DGM0754,1.0,['2010-01-29'],20578.0,0.0

2010-02-17,LDD0560,20.0,['2010-01-11'; '2010-01-14'; '2010-01-19'; '2010-01-20'; '2010-01-22'; '2010-01-25'; '2010-01-28'; '2010-01-28'; '2010-02-01'; '2010-02-02'; '2010-02-05'; '2010-02-05'; '2010-02-09'; '2010-02-10'; '2010-02-10'; '2010-02-11'; '2010-02-15'; '2010-02-16'; '2010-02-16'; '2010-02-16'],638375.0,24.0

上述验证通过！

二〇一九年一月十二日星期六

下午7时16分

今天晚上，开始CERT4.2数据集上，全体用户的KMeans+OCSVM的复现实验。

1. 目标：验证Static方法是否可以在CERT4.2上应用；
2. 目标：验证RealTime方法是否可以在CERT4.2上应用（逐月应用KMeans+OCSVM检测当月离职用户中的Insiders）

实际构造ATF时突然发现竟然还没有计算LED特征，于是赶紧补上：

在Build\_ATF中直接添加LED模块即可，因为前期已经得到了每月的用户LED特征，只需要根据每月用户进行累加构造ATF的最后三个特征即可。

按月的累积ATF编写完毕，开始验证：

2010-03

user\_id,O\_Score,C\_Score,E\_Score,A\_Score,N\_Score,CPB-I,CPB-O,JS\_Score,Team\_CPB-I-mean,Team\_CPB-O-mean,Users-less-mean-A,Users-less-mean-A and C,Users-less-mean-C,Users-High-mean-N,Team\_CPB-I-median,Team\_CPB-O-median,leader-CPB-I,leader-CPB-O,dis\_ocean,dis\_os,cnt\_send,cnt\_send\_size,cnt\_send\_attach,cnt\_send\_days,cnt\_email\_days,cnt\_late\_days,cnt\_early\_days,month\_work\_days,

CEL0561,40.0,39.0,36.0,19.0,40.0,-7.81,-20.47,11.35,-14.0682666667,-17.6715733333,10,5,8,13,-10.164,-10.164,-9.8596,45.6043474837,46.0,21.0,658953.0,24.0,16.0,**22.0,17.0,63.0**,

单项验证：

Calvin Edan Love,CEL0561,40,39,36,19,40

CEL0561,40.0,39.0,36.0,19.0,40.0,-14.0682666667,-17.6715733333,10,5,8,13,-10.164,-10.164,-9.8596,

CEL0561,40.0,39.0,36.0,19.0,40.0,-7.81,-20.47,45.6043474837,46.0,21.0,658953.0,24.0,16.0,

2010-03：CEL0561,9.0,18.0,7.0,5.0,23,1,2010-03-16,

2010-02：CEL0561,9.0,18.0,10.0,6.0,20,5,2010-02-02,2010-02-03,2010-02-04,2010-02-09,2010-02-24,

2010-01：CEL0561,9.0,18.0,5.0,6.0,20,3,2010-01-05,2010-01-13,2010-01-15,

2010-04：CEL0561,9.0,18.0,5.0,7.0,21,4,2010-04-08,2010-04-09,2010-04-26,2010-04-29,

验证通过！

晚上回家开始验证Static模式的CERT4.2数据集的KMeans+OCSVM

在继续实验前，需要先整理出CERT4.2中Insiders\_1/2/3的用户列表即离职日期

1. 获取Insiders列表
2. 获取Leave Users列表

二〇一九年一月十三日星期日

上午12时7分

在CERT4.2上进行实验：

Miu = 1.0

PCA=1

0 [0.05, 1.0, **0.6714285714285714**, 0.6714285714285714, 0.02843601895734597, 0.05052526263131566, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **0.7666666666666667, 0.5666666666666667, 0.7**,

**如果采用PCA=1，但是nu最小0.01**

**0.01, 1.0, 0.7714285714285715, 0.7714285714285715, 0.028962611901000527, 0.054527263631815905, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',**

**max\_iter=-1, nu=0.01, random\_state=10, shrinking=True, tol=0.01,**

**verbose=False), 0.9, 0.7333333333333333, 0.5**

**使用上述最好结果：**

**场景1：召回率0.9**

**场景2：召回率0.73**

**场景3：召回率0.5**

**或者可以使用**

**[0.04, 1.0, 0.7142857142857143, 0.7142857142857143, 0.027909426013691417, 0.05152576288144072, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',**

**max\_iter=-1, nu=0.04, random\_state=10, shrinking=True, tol=0.01,**

**verbose=False), 0.8, 0.6, 0.8,**

**哈哈，重新调整了nu的值终于得到**

**[0.003, 1.0, 0.7857142857142857, 0.7857142857142857, 0.030015797788309637, 0.0560280140070035, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',**

**max\_iter=-1, nu=0.003, random\_state=10, shrinking=True, tol=0.01,**

**verbose=False), 0.9333333333333333, 0.7, 0.6,**

**保存该结果！**

**可以根据需要，灵活使用nu=0.003和nu=0.01两种情况**