2018年12月3日星期一

下午1时40分

由于之前的实验太多太杂，这周开始正式启用新的实验记录模式：周记录。

每周会有一个独立的文件记录该周的实验分析与研究思路，不同日期的记录使用显著的日期分隔符进行标注；同一天中不同内容的研究段落采用段落分割标注，每天最后一个段落应是当日实验小结。

今天是12月第一个工作周的第一天，时间不多，务必紧张起来。

段落分隔标志

1. 改进的实验

原本想着重新开一个实验序号作为V8，不过还是放弃了；因为现在虽然提取出了CERT5.2中的每月用户与当月离职邮件联系人的联系特征，但是可能还是需要依靠OCEAN/OS特征进行辅助分析。因而，我们继续在V7的框架下进行实验，并将V8当作主客观融合的检测研究。

改进的实验思路主要如下：

1. 先仅依靠邮件联系特征作为建模离职员工对该用户JS影响的因素（暂时没有考虑OCEAN/OS）
2. 改进的关键在于：不再是使用该用户与所有离职用户的通信计算JS影响，而是仅依据“亲密性”的联系开展分析；
3. “亲密性”首先应体现在用户A/B之间的联系频率/时间，即B离职前与A联系了多少天？一般而言，突然任务造成某天联系多次，其亲密性体现远不及经常性地联系来的可靠；
4. 因此，我们每个月都从最终的离职通讯统计文件*CERT5.2\_Users\_Final\_LeaveContacts\_EmailFeats\_V0.7.csv*中提取应分析的，该用户截止到当月的所有与亲密离职用户的邮件联系特征，并以此进行后续计算；
5. 因此，我们只需要修改Predictor模块中，提取每月需分析的用户与离职用户邮件通讯部分即可，后续尚且不变；
6. **筛选的依据：A与B的通讯天的长度（不考虑重合），然后按照从大到小选取均值/中位数（含中位数）的部分，并保存到该月该用户所选择的特征，即每个月份目录下，应保存一份当月的CERT5.2用户与离职用户通信特征的特例；（格式类似: Month\_Users\_LC\_EmailFeats\_OnMean/Median.csv）**
7. **值得注意的是，上述代码不需要从2010-02开始计算，由于数据提取已经完毕，可以直接从关心的2010-07月开始分析，以预测2010-08月中Insiders\_2的离职情况；**

2018年12月3日星期一

下午10时44分

初步实验完成，全部采用了类结构来编写，果然层次感清晰了很多，就是代码量增加了。

我们先来验证下2010-08月离职的三个场景二攻击用户：

VCF1602,2010-08-20,

On\_Mean: VCF1602,3.64043229448,（254）

On\_Median: VCF1602,3.93566414623, （331）

<<VCF1602>> On\_Mean:

*NWP1609,0.138888888889,41.0,664646.195122,30.0,31.0,111969.322581,3.0,34.0,22.0,48.0,*

*ZAD1621,0.193548387097,74.0,351953.0,30.0,50.0,148037.36,5.0,57.0,37.0,77.0,*

*TAG1610,0.131034482759,82.0,456721.182927,37.0,63.0,279294.809524,14.0,61.0,45.0,86.0,*

CKP0630,2010-08-26,

On\_Mean: CKP0630,1.38629436112, (1522)

CKP0630,1.38629436112,(1833)

<<CKP0630>>

*NTV1777,1.0,1.0,24299.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,*

*RDP1751,-1.0,0.0,0.0,0.0,1.0,34067.0,0.0,0.0,1.0,1.0,*

ZIE0741,2010-08-27,

On\_Mean: ZIE0741,2.16224841453,(832)

On\_Median: ZIE0741,2.65488905589,

<<ZIE0741>>

*CDO0684,0.391304347826,16.0,27518.6875,0.0,7.0,1363520.0,10.0,10.0,7.0,14.0,*

*BNS0484,1.0,11.0,30665.6363636,0.0,0.0,0.0,0.0,9.0,0.0,9.0,*

*原有数据*

*<<ZIE0741\_start>>*

*2010-02:*

*KSW0708,-1.0,0,[],0,0,1.0,[2010-02-04],3073968.0,2,0,1,*

*2010-03:*

*CDO0684,0.391304347826,16.0,[2010-01-04; 2010-01-05; 2010-01-08; 2010-01-12; 2010-01-13; 2010-01-15; 2010-01-18; 2010-01-20; 2010-01-21; 2010-01-22],27518.6875,0.0,7.0,[2010-01-05; 2010-01-07; 2010-01-08; 2010-01-11; 2010-01-13; 2010-01-15; 2010-01-19],1363520.0,10.0,10,7,*

*NWP1609,-1.0,0,[],0,0,1.0,[2010-02-09],27703.0,0.0,0,1,*

*2010-04:*

*2010-05:*

*OCW1127,-1.0,0,[],0,0,1.0,[2010-01-08],34511.0,0.0,0,1,*

*2010-06:*

*JXH1061,1.0,1.0,[2010-03-03],41853.0,0.0,0,[],0,0,1,0,*

*TAG1610,1.0,1.0,[2010-03-15],24317.0,0.0,0,[],0,0,1,0,*

*BMR0865,1.0,1.0,[2010-05-06],40914.0,0.0,0,[],0,0,1,0,*

*2010-07:*

*BNS0484,1.0,11.0,[2010-01-06; 2010-01-08; 2010-01-12; 2010-01-13; 2010-01-14; 2010-01-18; 2010-01-19; 2010-01-22; 2010-06-22],30665.6363636,0.0,0,[],0,0,9,0,*

*DAS1320,1.0,1.0,[2010-07-30],25796.0,0.0,0,[],0,0,1,0,*

*AWW0718,-1.0,0,[],0,0,1.0,[2010-04-08],33501.0,0.0,0,1,*

*2010-08:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2010-09:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2010-10:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2010-11:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2010-12:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2011-01:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2011-02:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2011-03:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2011-04:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*2011-05:*

*ZJN1492,1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,*

*DTB0722,1.0,1.0,[2010-08-18],24682.0,0.0,0,[],0,0,1,0,*

*CKP0630,-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,*

*<<ZIE0741\_end>>*

此次排名第一的竟然是

CDB1594,15.9167359738,

<<CDB1594>>

XMG1579,-0.333333333333,2.0,29284.0,0.0,4.0,2029399.5,10.0,2.0,4.0,6.0,

初步怀疑是邮件大小造成了混淆，那么修改RLV的计算公式，不再考虑邮件size，而是直接数量的求和：

rlv\_e = math.log(1.0 + cnt\_email\_days + cnt\_send + cnt\_recv + cnt\_attach\_send + cnt\_attach\_recv, math.e)

修改后重新计算下2010-07月三个场景2攻击者的结果：

VCF1602,2010-08-20,

OnMean: VCF1602,4.62421468719,（301）

OnMedian:VCF1602,4.98784509047,(480)

小结：

最初假设：

对于一个用户而言，只有最重要的用户需要考虑，在此基础上利用On\_Mean/Median进行了筛选，但是结果十分不理想，原因在于有些用户的筛选并未严格按照自己所设想的进行，如下述用户，原始特征竟然没有进行筛选！全部参与了运算！

由于该用户均值即为1，因此高于均值时，其结果为0，而等于均值的结果为全集

有两点可以尝试：

1. 确定高于均值或者高于中位数；
2. 单个用户计算时不再进行归一化，而是直接用log的形式进行运算（e或者10为底）；
3. 最后所有用户的RLV进行归一化或者简单排序

对于

FMP0459,11.0382388878,

<<FMP0459>>

ADL1898,1.0,1.0,41673.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,

SDL0541,1.0,1.0,238282.0,1.0,0.0,0.0,0.0,1.0,0.0,1.0,

CDO0684,-1.0,0.0,0.0,0.0,1.0,34154.0,0.0,0.0,1.0,1.0,

FKH0864,1.0,1.0,1469980.0,1.0,0.0,0.0,0.0,1.0,0.0,1.0,

GMM1037,1.0,1.0,22767.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,

WBP0828,1.0,1.0,33527.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,

RFP1918,1.0,1.0,1846331.0,1.0,0.0,0.0,0.0,1.0,0.0,1.0,

MAR1075,1.0,1.0,609075.0,1.0,0.0,0.0,0.0,1.0,0.0,1.0,

RDP1751,1.0,1.0,26751.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,

BRM0126,-1.0,0.0,0.0,0.0,1.0,319949.0,1.0,0.0,1.0,1.0,

HKK0881,1.0,1.0,18708.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,

ESP1198,1.0,1.0,38369.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,

EJO0236,1.0,1.0,24591.0,0.0,0.0,0.0,0.0,1.0,0.0,1.0,

<<FMP0459\_start>>

2010-02:

ADL1898,1.0,1.0,[2010-01-04],41673.0,0.0,0,[],0,0,1,0,

SDL0541,1.0,1.0,[2010-01-08],238282.0,1,0,[],0,0,1,0,

2010-03:

CDO0684,-1.0,0,[],0,0,1.0,[2010-01-26],34154.0,0.0,0,1,

2010-04:

FKH0864,1.0,1.0,[2010-01-26],1469980.0,1,0,[],0,0,1,0,

GMM1037,1.0,1.0,[2010-01-28],22767.0,0.0,0,[],0,0,1,0,

2010-05:

2010-06:

WBP0828,1.0,1.0,[2010-01-25],33527.0,0.0,0,[],0,0,1,0,

RFP1918,1.0,1.0,[2010-03-15],1846331.0,1,0,[],0,0,1,0,

MAR1075,1.0,1.0,[2010-04-05],609075.0,1,0,[],0,0,1,0,

RDP1751,1.0,1.0,[2010-05-06],26751.0,0.0,0,[],0,0,1,0,

BRM0126,-1.0,0,[],0,0,1.0,[2010-03-24],319949.0,1,0,1,

2010-07:

HKK0881,1.0,1.0,[2010-03-05],18708.0,0.0,0,[],0,0,1,0,

ESP1198,1.0,1.0,[2010-05-05],38369.0,0.0,0,[],0,0,1,0,

EJO0236,1.0,1.0,[2010-05-18],24591.0,0.0,0,[],0,0,1,0,



2018年12月4日星期二

上午9时16分

在继续了这么多次实验后，发现不能这么靠运气实验了，必须先从数据分析入手，待找到某种特性后，再沿着该方向进行实验。

特性分析实验一

既然已经得到了CERT5.2中所有用户离职前的离职联系人的通讯列表，那么就来看看，如果直接从与离职人员发生联系的人数上，是否Insiders有明显特征呢？

或者，如果不能直接从绝对人数上发现规律，如果是比例代表的相对人数呢？即一个用户在工作周期中所有通信员工中与离职员工通信所占的比例；

感谢自己当初统计当月用户邮件联系人时，区分为了当月的离职用户通信与截止到当月的非离职用户通信，因此，如果想简单比较（比如2010-07月）CERT5.2中用户与离职用户通信的延伸影响，可以计算截止到2010-07月每个用户联系的离职用户数量占所有通讯用户的比重（反映出该用户与离职联系人联系对其自身社交关系的重要性）

首先检查了之前的文件，重新进行了更新，得到了CERT5.2中所有用户的，与离职用户的联系信息***CERT5.2\_Users\_Final\_LeaveContacts\_EmailFeats.csv***

这次如果用户离职，则离职后的联系人信息不再记入。

这次我们的实验简单实现两个功能，目标时探索联系离职用户的人数是否具有区分性呢？

针对Insiders\_2的用户，统计分析：

1. 全生命周长度内：CERT5.2所有用户分别联系过的离职用户的个数，按照从大到小排列，保存到结果文件；（user\_id, cnt\_lc\_0, sequence\_0）
2. 单独分析Insiders\_2离职前一个月为止，离职用户联系数量特征（cnt\_lc\_1, sequnce\_1）
3. 单独分析Insiders\_2离职前一个月为止，但是将所有可能的变量作为特征输入，采用随机森林标记计算各个特征的重要性；

结果一：

统计全生命周期的Insiders\_2联系的离职用户的个数，且返回其大小排序索引号

TMT0851 59 469

HSN0675 53 685

LVF1626 46 1007

HIS1394 43 1096

SNK1280 40 1206

BYO1846 38 1271

JAL0811 37 1312

DCC1119 36 1335

OKM1092 32 1543

。。。上述结果说明，单从全生命周期来看，肯定不具有区分性！

那么如果我们稍作修改，即考虑到全生命周期，所有用户总共联系的联系人中，离职用户所占的比例呢？

下午3时53分

中间运行出现了一点小错误，原因在于自己统计用户邮件联系时，忽略了CEO，因为CEO没有离职，也不是Insiders，因此不做考虑。

统计出勤时考虑了CEO，一共2000个用户；

而统计邮件的Leave\_Contacts时没有考虑CEO，因此只有1999个用户；

AEH0001 2011-05-31 不存在contacts记录...

此时的结果依旧十分分散：整体效果不如单独使用离职通信个数

ICB1354 0.0480225988701 1103

ITA0159 0.0478359908884 1124

OKM1092 0.0473372781065 1171

HMS1658 0.045738045738 1316

HSN0675 0.0448772226926 1398

JAL0811 0.0428240740741 1552

TMT0851 0.0423546302943 1573

LVF1626 0.0417422867514 1612

HIS1394 0.0412272291467 1630

DCC1119 0.0404494382022 1676

GWG0497 0.0392441860465 1726

BYO1846 0.0353817504655 1818

SNK1280 0.0344827586207 1828

RRS0056 0.0337381916329 1836

LC\_Statis\_0\_sort 0 ['SAA0374', 0.08444444444444445]

LC\_Statis\_0\_sort 1 ['QRC1676', 0.0796812749003984]

LC\_Statis\_0\_sort 2 ['RRG1253', 0.07962962962962963]

LC\_Statis\_0\_sort 3 ['SRC1468', 0.0777988614800759]

LC\_Statis\_0\_sort 4 ['WGF1260', 0.07707509881422925]

前五名都没有离职，因此直接计算离职联系人/总联系人比重的方法行不通，说明这里面无效联系人太多，影响了判断

Leave\_Insiders\_Earliest\_Time:

Scene\_1: KEW0198,2010-07-29,

Scene\_2: VCF1602,2010-08-20,

Scene\_3: MPF0690,2010-06-18,

**方案备选：**

**一种可用的备选方案：**

1. **结合最初工作的CPB指标，考虑用户自身的CPB指数，以及工作团队的CPB指数，然后结合上出勤率（迟到次数+早退次数）以及和离职人邮件联系的社交属性，共同得到用户的JSR指标；**
2. **选择2010-01：2010-05作为训练集，采用有标记的SVM训练，其中，目标是识别出未来离职的用户；**
3. **所有离职用户的共同点就是：Low Job Satisfactory，LJS-->Low Work-->Laid off & LJS-->Look for new job**
4. **识别出的结果中，再使用出勤率分析，以说明LJS的刻画一致性；**

**下午4时11分**

接下来，不再眉毛胡子一把抓，也不能再像没头的苍蝇乱飞，我们回归到一个真实的Insider与一个普通用户，需要注意的是：

LJS：可能表现出跳槽或迟到早退；

HJS：一定没有跳槽或者迟到早退；

2010-08：

Insiders\_2: MDS0680,8.0,17.5,0.0,0.0,22,-1

LDAP：['2 - Executive': '1 - Adminstration': '6 - Security': '1 - BuildingSecurity']

跳槽离职，却没有出勤问题

<<MDS0680\_start>>:2010-09-17

2010-02:

RMB1821,-1.0,0,[],0,0,**1.0**,[2010-01-08],22681.0,0.0,0,1,

['1 - Executive': '5 - SalesAndMarketing': '2 - Sales': '5 - RegionalSales']

2010-03:

NWP1609,1.0**,1.0**,[2010-03-01],40087.0,0.0,0,[],0,0,1,0,

['1 - Executive': '5 - SalesAndMarketing': '2 - Sales': '4 - RegionalSales']

2010-04:

CLL0306,1.0,**1.0**,[2010-01-08],17361.0,0.0,0,[],0,0,1,0,

['2 - Executive': '2 - ResearchAndEngineering': '3 - SoftwareManagement': '2 - DesktopSoftware']

2010-05:

2010-06:

RFP1918,1.0,1.0,[2010-01-11],416485.0,1,0,[],0,0,1,0,

['2 - Executive': '3 - Manufacturing': '3 - Assembly': '5 - AssemblyDept']

NWH0960,1.0,1.0,[2010-03-10],29739.0,0.0,0,[],0,0,1,0,

['1 - Executive': '2 - ResearchAndEngineering': '3 - SoftwareManagement': '6 - EmbeddedSoftware']

HFF0560,1.0,1.0,[2010-03-24],724551.0,1,0,[],0,0,1,0,

['1 - Executive': '2 - ResearchAndEngineering': '4 - Engineering': '3 - ElectricalEngineering']

WBP0828,1.0,1.0,[2010-03-26],18393.0,0.0,0,[],0,0,1,0,

['1 - Executive': '3 - Manufacturing': '3 - Assembly': '3 - Stockroom']

MAR1075,-1.0,0,[],0,0,1.0,[2010-01-12],29864.0,0.0,0,1,

['1 - Executive': '5 - SalesAndMarketing': '2 - Sales': '2 - RegionalSales']

MPF0690,-1.0,0,[],0,0,1.0,[2010-02-17],28634.0,0.0,0,1,

['2 - Executive': '1 - Adminstration': '6 - Security': '2 - ElectronicSecurity']

RDP1751,-1.0,0,[],0,0,1.0,[2010-04-08],3061064.0,5,0,1,

['2 - Executive': '5 - SalesAndMarketing': '3 - FieldService': '4 - RegionalFieldService']

2010-07:

CRD0272,-1.0,0,[],0,0,1.0,[2010-06-04],34377.0,0.0,0,1,

['2 - Executive': '1 - Adminstration': '6 - Security': '2 - ElectronicSecurity']

KEW0198,-1.0,0,[],0,0,1.0,[2010-06-21],38655.0,0.0,0,1,

['1 - Executive': '5 - SalesAndMarketing': '2 - Sales': '1 - RegionalSales']

2010-08:

DTB0722,1.0,1.0,[2010-03-09],19181.0,0.0,0,[],0,0,1,0,

['2 - Executive': '2 - ResearchAndEngineering': '3 - SoftwareManagement': '3 - EmbeddedSoftware']

ZJN1492,1.0,1.0,[2010-03-26],26265.0,0.0,0,[],0,0,1,0,

['2 - Executive': '5 - SalesAndMarketing': '2 - Sales': '3 - RegionalSales']

SAF1942,1.0,1.0,[2010-04-15],33175.0,0.0,0,[],0,0,1,0,

['2 - Executive': '5 - SalesAndMarketing': '2 - Sales': '5 - RegionalSales']

CKP0630,1.0,1.0,[2010-04-22],31909.0,0.0,0,[],0,0,1,0,

['1 - Executive': '3 - Manufacturing': '3 - Assembly': '2 - AssemblyDept']

PBC0077,1.0,1.0,[2010-05-11],28780.0,0.0,0,[],0,0,1,0,

['1 - Executive': '2 - ResearchAndEngineering': '3 - SoftwareManagement': '1 - SQA']

ZIE0741,1.0,1.0,[2010-08-12],23978.0,0.0,0,[],0,0,1,0,

['2 - Executive': '2 - ResearchAndEngineering': '3 - SoftwareManagement': '4 - WebSoftware']

EPG1196,-1.0,0,[],0,0,1.0,[2010-07-08],108829.0,1,0,1,

['2 - Executive': '2 - ResearchAndEngineering': '3 - SoftwareManagement': '7 - WebSoftware']

2010-09:

<<MDS0680\_end>>

**我们再来选择一个正常用户（无离职，无迟到）**

**XDG1256,8.0,16.5,0.0,0.0,22,-1 （没有与离职用户的联系）**

**Line 28639: <<XDG1256\_start>>:2011-06-30**

**Line 28687: <<XDG1256\_end>>**

**HJB1444**

**HJB1444,7.0,16.5,0.0,0.0,28,-1**

**Line 44556: <<HJB1444\_start>>:2011-06-30**

**Line 44620: <<HJB1444\_end>>**

**可以这样推测：**

**如果一个用户没有离职，没有缺勤，那么该用户认为是High JS的，而相应的，也与leave\_contacts没有联系。**

**经过分析，一共有21个用户没有LC联系。**

**以XDG为例无迟到，无离职，但是有LC**

**<<XDG1256\_start>>:2011-06-30**

**2010-02:**

**2010-03:**

**MIB1265,0.2,3.0,[2010-01-12; 2010-01-14; 2010-03-12],25829.0,0.0,2.0,[2010-03-12; 2010-03-16],32233.5,0.0,3,2,**

**2010-04:**

**2010-05:**

**2010-06:**

**WBP0828,1.0,1.0,[2010-04-23],32415.0,0.0,0,[],0,0,1,0,**

**JXH1061,-1.0,0,[],0,0,1.0,[2010-05-07],15650.0,0.0,0,1,**

**RKW1936,-1.0,0,[],0,0,1.0,[2010-05-24],49941.0,0.0,0,1,**

**2010-07:**

**LSM1382,-1.0,0,[],0,0,1.0,[2010-02-23],997954.0,1,0,1,**

**2010-08:**

**JIB1258,-0.333333333333,5**

**.0,[2010-01-19; 2010-01-27; 2010-06-09; 2010-06-22; 2010-07-19],25623.6,0.0,10.0,[2010-01-11; 2010-01-13; 2010-02-09; 2010-05-03; 2010-05-04; 2010-05-25; 2010-06-01; 2010-06-02; 2010-06-25; 2010-07-15],300219.0,3.0,5,10,**

**PBC0077,-1.0,0,[],0,0,1.0,[2010-03-31],242806.0,1,0,1,**

**2010-09:**

**MDS0680,-1.0,0,[],0,0,1.0,[2010-05-21],28580.0,0.0,0,1,**

**TNB1616,-1.0,0,[],0,0,1.0,[2010-05-24],42228.0,0.0,0,1,**

**NBL1190,-1.0,0,[],0,0,1.0,[2010-06-29],33017.0,0.0,0,1,**

**2010-10:**

**MTD0971,1.0,1.0,[2010-05-10],23744.0,0.0,0,[],0,0,1,0,**

**RPJ1159,-1.0,0,[],0,0,1.0,[2010-06-18],55231.0,0.0,0,1,**

**2010-11:**

**HMK0653,-1.0,0,[],0,0,1.0,[2010-05-12],27596.0,0.0,0,1,**

**2010-12:**

**ZVW1475,1.0,1.0,[2010-01-06],13583.0,0.0,0,[],0,0,1,0,**

**HXP0976,1.0,1.0,[2010-05-28],33753.0,0.0,0,[],0,0,1,0,**

**PCK0271,-1.0,0,[],0,0,1.0,[2010-06-30],19460.0,0.0,0,1,**

**2011-01:**

**LVF1626,-1.0,0,[],0,0,1.0,[2010-06-30],11274.0,0.0,0,1,**

**DPK0954,-1.0,0,[],0,0,1.0,[2010-09-08],27316.0,0.0,0,1,**

**NEG0281,-1.0,0,[],0,0,1.0,[2010-12-22],32478.0,0.0,0,1,**

**2011-02:**

**SNK1280,0.0,1.0,[2010-11-12],32888.0,0.0,1.0,[2010-11-08],32328.0,0.0,1,1,**

**ZKP0542,-1.0,0,[],0,0,1.0,[2010-05-21],47222.0,0.0,0,1,**

**2011-03:**

**WHB1247,0.2,12.0,[2010-01-11; 2010-06-01; 2010-06-24; 2010-07-30; 2010-08-20; 2010-10-20; 2010-11-04; 2010-12-30; 2011-01-05; 2011-02-02; 2011-02-11; 2011-02-16],27516.8333333,0.0,8.0,[2010-02-09; 2010-02-22; 2010-03-02; 2010-03-09; 2010-03-24; 2010-04-21; 2010-09-29; 2011-03-02],1266844.0,9.0,12,8,**

**GTN1021,-1.0,0,[],0,0,1.0,[2010-02-16],25877.0,0.0,0,1,**

**MBW1149,-1.0,0,[],0,0,1.0,[2010-07-13],35187.0,0.0,0,1,**

**NIV1608,-1.0,0,[],0,0,1.0,[2010-07-20],30816.0,0.0,0,1,**

**SQC1072,-1.0,0,[],0,0,1.0,[2010-10-15],27844.0,0.0,0,1,**

**2011-04:**

**DCA0857,-1.0,0,[],0,0,2.0,[2010-12-17; 2011-04-12],26841.5,0.0,0,2,**

**2011-05:**

**JIP1503,-1.0,0,[],0,0,1.0,[2010-01-13],31034.0,0.0,0,1,**

**LAH0463,-1.0,0,[],0,0,1.0,[2010-02-26],26570.0,0.0,0,1,**

**JKB1843,-1.0,0,[],0,0,1.0,[2010-05-14],22386.0,0.0,0,1,**

**MTP1582,-1.0,0,[],0,0,1.0,[2011-01-17],447256.0,1,0,1,**

**<<XDG1256\_end>>**

以MMK1532为例，其没有跳槽离职，但是却在2010-08月份有着出勤率问题（13次迟到，12次早退）

MMK1532,7.5,18.5,13.0,12.0,22,9,2010-08-05,2010-08-06,2010-08-09,2010-08-10,2010-08-17,2010-08-24,2010-08-27,2010-08-30,2010-08-31,

<<MMK1532\_start>>:2011-06-30

2010-02:

WMH1300,1.0,1.0,[2010-01-25],38275.0,0.0,0,[],0,0,1,0,

2010-03:

MIB1265,1.0,1.0,[2010-03-05],26650.0,0.0,0,[],0,0,1,0,

2010-04:

2010-05:

2010-06:

GWH0961,1.0,1.0,[2010-01-28],34386.0,0.0,0,[],0,0,1,0,

BRM0126,-1.0,0,[],0,0,1.0,[2010-02-05],30144.0,0.0,0,1,

OCD1985,-1.0,0,[],0,0,1.0,[2010-02-05],22704.0,0.0,0,1,

MAR1075,-1.0,0,[],0,0,1.0,[2010-03-05],16372.0,0.0,0,1,

2010-07:

ESP1198,-1.0,0,[],0,0,1.0,[2010-04-20],26469.0,0.0,0,1,

2010-08:

CKP0630,-1.0,0,[],0,0,1.0,[2010-03-10],37179.0,0.0,0,1,

VRP0267,-1.0,0,[],0,0,1.0,[2010-04-30],40770.0,0.0,0,1,

ZJN1492,-1.0,0,[],0,0,1.0,[2010-08-05],33719.0,0.0,0,1,

2010-09:

DMP0344,1.0,1.0,[2010-02-26],21582.0,0.0,0,[],0,0,1,0,

TRC1838,-1.0,0,[],0,0,1.0,[2010-06-02],37426.0,0.0,0,1,

KLB0918,-1.0,0,[],0,0,1.0,[2010-08-05],38257.0,0.0,0,1,

2010-10:

JKB0287,-1.0,0,[],0,0,1.0,[2010-09-23],27548.0,0.0,0,1,

DNJ0740,-1.0,0,[],0,0,1.0,[2010-10-25],21706.0,0.0,0,1,

2010-11:

DXF1662,-1.0,0,[],0,0,1.0,[2010-02-15],1565990.0,2,0,1,

CTT0639,-1.0,0,[],0,0,1.0,[2010-03-17],24830.0,0.0,0,1,

ISW0738,-1.0,0,[],0,0,1.0,[2010-08-03],35961.0,0.0,0,1,

RAT0514,-1.0,0,[],0,0,1.0,[2010-08-31],36964.0,0.0,0,1,

2010-12:

MJA1784,-1.0,0,[],0,0,1.0,[2010-01-13],1680821.0,2,0,1,

KJG1121,-1.0,0,[],0,0,1.0,[2010-02-12],22063.0,0.0,0,1,

HMS1658,-1.0,0,[],0,0,1.0,[2010-02-19],30893.0,0.0,0,1,

KVF1143,-1.0,0,[],0,0,1.0,[2010-05-10],32429.0,0.0,0,1,

RRS0056,-1.0,0,[],0,0,1.0,[2010-05-28],25067.0,0.0,0,1,

GCB0118,-1.0,0,[],0,0,1.0,[2010-06-11],28456.0,0.0,0,1,

JOE1672,-1.0,0,[],0,0,1.0,[2010-08-10],33154.0,0.0,0,1,

HIS1394,-1.0,0,[],0,0,1.0,[2010-09-27],1192950.0,1,0,1,

2011-01:

KRC1348,1.0,1.0,[2010-02-11],39327.0,0.0,0,[],0,0,1,0,

DPK0954,-1.0,0,[],0,0,1.0,[2010-08-27],30732.0,0.0,0,1,

KHW0289,-1.0,0,[],0,0,1.0,[2010-09-16],27792.0,0.0,0,1,

RBC1624,-1.0,0,[],0,0,1.0,[2010-10-22],1496105.0,1,0,1,

2011-02:

ZKP0542,1.0,1.0,[2011-01-20],41568.0,0.0,0,[],0,0,1,0,

2011-03:

TMC0934,1.0,1.0,[2010-02-24],31889.0,0.0,0,[],0,0,1,0,

IAJ1729,-1.0,0,[],0,0,1.0,[2011-03-04],2149293.0,2,0,1,

2011-04:

XAM0376,1.0,1.0,[2010-10-08],18282.0,0.0,0,[],0,0,1,0,

DCA0857,-1.0,0,[],0,0,1.0,[2010-01-18],24213.0,0.0,0,1,

XBK0246,-1.0,0,[],0,0,1.0,[2010-04-09],37363.0,0.0,0,1,

HSN0675,-1.0,0,[],0,0,1.0,[2011-01-17],24694.0,0.0,0,1,

WWW0701,-1.0,0,[],0,0,1.0,[2011-03-18],17013.0,0.0,0,1,

2011-05:

CWW1120,-1.0,0,[],0,0,2.0,[2010-01-28; 2010-06-09],30235.5,0.0,0,2,

LAH0463,-1.0,0,[],0,0,1.0,[2010-05-19],23628.0,0.0,0,1,

VPA0974,-1.0,0,[],0,0,2.0,[2010-08-20; 2010-12-01],25243.5,0.0,0,2,

TMT0851,-1.0,0,[],0,0,1.0,[2011-01-26],21809.0,0.0,0,1,

<<MMK1532\_end>>



二〇一八年十二月五日星期三

上午9时26分

CERT5.2中三类攻击场景离职的最早时间：

Leave\_Insiders\_Earliest\_Time:

Scene\_1: KEW0198,2010-07-29,

Scene\_2: VCF1602,2010-08-20,

Scene\_3: MPF0690,2010-06-18,

我们看最早的2010-06，主观分析是否可以从2010-06开始预测？

经过验证，MPF0690的用户恶意行为最早开始于2010-06-17

经过验证，KEW0198的用户恶意行为最早开始于2010-07-07

经过验证，VCF1620的用户恶意行为最早开始于2010-06-23

因此，我们可以将原有数据按照上述事件截断分为两段：

1. 训练集时间段：2010-01：2010-05，该时间段内将在职用户标记为+1，离职用户标记为-1，并由此提取特征训练SVM，并遍历参数空间得到最优参数，作为之后预测的分类器；抑或，建立一个分类器参数按照月份结果更新的机制，后续不断自我更新；
2. 使用训练得到的分类器对于2010-06开始预测谁会离职？重要的指标有两个（预测结果中包含当月离职Insiders的召回率；以及当月预测结果中，真正离职用户的召回率与误报率）
3. 简单的情况下，分类器依次对逐个月进行检测，输出结果；后续可以考虑对分类器进行实时进化调整，依据三个重要指标进行计算更新
4. 今天周三，我们先来实现一个朴素的主观SVM预测器；建立实验八，主体核心模块全部采用面向对象实现；

朴素的主观SVM预测器特征说明：

对于CERT5.2的每个用户而言，

|  |  |
| --- | --- |
| 字段名称 | 注释 |
| User\_id |  |
| Personality | Scores of O,C,E,A,N |
| Relationship with LC | 同离职者的人际影响因子 |
| Dis\_OCEAN | 同虚拟离职综合体的人格偏差 |
| Dis\_OS | 同虚拟离职综合体的组织距离 |
| Cnt\_Send/Cnt\_Recv | 同虚拟离职综合体的收发邮件总数（before预测月份） |
| Cnt\_Send\_Size/Cnt\_Recv\_Size | 同虚拟离职综合体收发邮件总字数大小（before预测月份） |