2018年11月16日星期五

上午9时4分

经过一周的努力，终于抛出了原始的程序，该程序中虽然试着对每月迟到早退的用户数据进行了统计（按月统计），然而并未用于更新高危用户预测的Risk\_Ratio。

今天的主要任务是查看、分析单纯使用RLF的JSR建模方法。

分析一：基于RLV计算的JSR预测结果直接分析

不考虑用户标记的Risk\_Ratio的自动更新变化

我们先来验证检查场景二跳槽用户的情况，

首先是2010年8月离职的三个用户，需要查看2010年7月的预测结果：

VCF1602,2010-08-20,

预测：VCF1602,14.7263465016,

前100最低：AMC0854,16.561130049

CKP0630,2010-08-26,

预测：CKP0630,2.72629007281,

ZIE0741,2010-08-27,

预测：ZIE0741,12.253115567,

我们继续查看2010-9月份的Insiders2：

SIS0042,2010-09-02,

预测：SIS0042,16.6124371579,

前100最低：KCG1615,22.9780816538

本月：SIS0042,2.90294234691

本月：KCG1615,10.8445969053

TNB1616,2010-09-10,

预测：TNB1616,18.3628686663,

TRC1838,2010-09-15,

预测：TRC1838,17.1635610489,

MDS0680,2010-09-17,

**预测：MDS0680,24.2026684626, 命中第58位**

WDT1634,2010-09-20,

预测：WDT1634,3.87513518858,

OSS1463,2010-09-21,

预测：OSS1463,8.06596705673,

CIF1430,2010-09-23,

预测：CIF1430,5.28141842377,

初步查看了2010年8月与9月两个月的Insiders-2离职情况，发现，其中只有一个用户正确识别到，而其他用户却都排序较靠后。

为了进一步分析，我们查看某几个用户在2010年7月的当月RLV特征

对于CIF1480用户而言，在2010年7月时，没有离职的邮件联系人

我们试着查看下2010年12月份离职的Insiders\_2：

为此，需要分析2010年11月的预测文件

RRS0056,2010-12-10,

预测：RRS0056,32.7934833797,

ICB1354,2010-12-15,

预测：ICB1354,19.254197899,

BYO1846,2010-12-15,

**预测：BYO1846,46.0780045872**

**前100：MGS1944,45.9394562009**

HXP0976,2010-12-20,

HXP0976,19.4707901928,

HMS1658,2010-12-30,

HMS1658,24.9066235757,

HIS1394,2010-12-30,

**预测HIS1394,49.240394831,**

**我们分析下**

**RRS0056与HIS1394的不同**

**再分析RRS0056与MGS1944的不同**

**从中提取Insiders\_2的相似性，提取insiders与一般用户的差异性**

2018年11月16日星期五

下午1时56分

下午准备重新跑一遍CERT5.2所有用户的Logon迟到早退分析程序；

另一方面，自己在高危用户预测模块中设立了每个月分析完毕后，建立预测器后从当前全体用户集合中删除当月离职的用户，避免下个月再次分析这些离职的用户。其实自己读取的是邮件数据，因此如果该月用户离职，下月自然不会有邮件数据，应该同最后的结果影响不大。

2018年11月17日星期六

下午3时58分

周五下午一直在跑程序，终于跑完了一个没有阈值更新的JSR预测版本，然后写了个小程序简单分析三类场景中被预测中的比例，由于场景二是我们最为关心的主客观融合的攻击，因此我们重点来分析下30个Insiders\_2的情况。

在验证数据的时候，突然发现：自己的High\_Risk没有按照从大到小的顺序输出，导致输出错误，赶紧修正！

程序验证：

对于AAB用户而言；

2010-02的Accumulate\_JSR是：

AAB1302,1.14218527305,

2010-03的Current\_JSR是：

AAB1302,1.23029298274

则2010-03的Accumulate\_JSR是：

AAB1302,2.37247825579,

验证通过！

自己的程序应该没有问题，问题仅仅出现在2010-07，很有可能是自己人工修改文件名时重复了，因而，出现了错误。

后续需要重新运行一遍所有的程序，然后再分析结果。

2018年11月18日星期日

下午8时39分

VCF1602,2010-08-20,

* Current\_JSR: VCF1602,3.99814860414, (429)
* Accumulate\_JSR: VCF1602,**14.7263465016**, (264)
* The Last High-Risk: **SME1271,17.8212446305**(97)

~~CKP0630,2010-08-26,~~

* Current\_JSR: None
* Accumulate\_JSR: CKP0630,2.72629007281,(1872)

ZIE0741,2010-08-27,

* Current\_JSR:ZIE0741,3.89639078657(495)
* Accumulate\_JSR:ZIE0741,**13.4740105641**,(394)
* 看2010-08月份离职的Insiders\_2，其最终的JSR与前5%的预测比例相差不大，应该在数据上有改进的空间；

1. 直接分析前97个用户的四个维度的中间变量特征，然后与三个离职用户的中间变量特征比较，分析四个维度下哪些变量更突出？
2. 引入用户自身的CPB-O作为基础，以系数的形式对于初始JSR进行放缩，但是这需要比较计算Insiders\_2的CPB-O或者CPB-I是否具有区分特征（比如排位都比较靠前？）

SIS0042,2010-09-02,

* Current\_JSR: SIS0042,2.90294234691(1132)
* Accumulate\_JSR:SIS0042,16.6124371579,**(599)**
* The Last High-Risk: ONB1833,23.9557600566

TNB1616,2010-09-10,

* Current\_JSR:TNB1616,2.5493224598,(1483)
* Accumulate\_JSR:TNB1616,19.8291573278,**(304)**
* The Last High-Risk:

TRC1838,2010-09-15,

* Current\_JSR:TRC1838,2.67521850976(1354)
* Accumulate\_JSR:TRC1838,19.8508658697,**(301)**
* The Last High-Risk:

**MDS0680,2010-09-17**,

* Current\_JSR:MDS0680,9.28732503589(66)
* Accumulate\_JSR:MDS0680,25.6870261847,(53)
* The Last High-Risk:

WDT1634,2010-09-20,

* Current\_JSR:WDT1634,1.19414050789
* Accumulate\_JSR:WDT1634,3.87513518858,
* The Last High-Risk:

OSS1463,2010-09-21,

* Current\_JSR:OSS1463,2.61717538324(1431)
* Accumulate\_JSR:OSS1463,8.06596705673,(1615)
* The Last High-Risk:

CIF1430,2010-09-23,

* Current\_JSR:CIF1430,2.5343477812(1491)
* Accumulate\_JSR:CIF1430,6.60979893944,(1778)
* The Last High-Risk:

首先我们看看CIF1430在2010-08月份与离职员工的RLV特征（用来刻画二者的人际关系亲密度，或者该人对于目标用户的重要度）

* Line 998: CIF1430,PTM1432,34.0,27.0,39.0,42.0,31.0,29.0516780927,0,0,0,0,0.0,0.025641025641,20.0,993758.05,23.0,19.0,290405.736842,7.0,17,19,
* Line 999: CIF1430,PBC0077,19.0,22.0,15.0,20.0,35.0,30.6920185064,0,1,1,1,14.0,-1.0,0,0,0,1.0,22943.0,0.0,1,1,
* 在检查该用户的RLV计算时，发现了错误：

错误的公式：rlv\_line = math.log(math.e + rlv\_dis\_ocean + rlv\_dis\_os + rlv\_edays + rlv\_einfo, math.e)

上述公式中的各个变量应该同rlv是正相关，其中rlv\_einfo的确是按照邮件信息量的角度来进行的刻画，而OCEAN距离以及OS距离却未进行反置，应取

MCP0611,2010-10-06,

CHP1711,2010-10-13,

GWG0497,2010-10-15,

KSS1005,2010-10-16,

NAH1366,2010-11-17,

RRS0056,2010-12-10,

ICB1354,2010-12-15,

BYO1846,2010-12-15,

HXP0976,2010-12-20,

HMS1658,2010-12-30,

HIS1394,2010-12-30,

LVF1626,2011-01-14,

MGB1235,2011-01-21,

DCC1119,2011-01-26,

SNK1280,2011-02-11,

ITA0159,2011-02-17,

JAL0811,2011-02-25,

OKM1092,2011-04-29,

HSN0675,2011-04-29,

TMT0851,2011-05-11,

2018年11月19日星期一

上午11时34分

由于之前计算RLV时距离指标计算失误，故需要重新计算，这次我们仅重新处理RLV与JSR文件即可，即依据每个月提取的process文件，更新Current\_JSR与Accumulate\_JSR即可。

写了一个初步的程序，开始稍微进行验证

2010-02：

Current: JOE1672,5.78396847146,

Accu：JOE1672,5.78396847146,

2010-03：

Current: JOE1672,1.30141911093,

Accu: JOE1672,7.08538758239,

2018年11月19日星期一

下午4时39分

下午回来终于写好了调整后的程序，重新验证下各个用户的结果：

VCF1602,2010-08-20,

* Current\_JSR: VCF1602,4.09582899728,(387)
* Accumulate\_JSR:VCF1602,15.8665124582,(205)
* Top\_5% High-Risk:

CKP0630,2010-08-26,

* Current\_JSR:
* Accumulate\_JSR:CKP0630,2.69351553135,(1882)
* Top\_5% High-Risk:

ZIE0741,2010-08-27,

* Current\_JSR:ZIE0741,4.15667932666,(326)
* Accumulate\_JSR:ZIE0741,13.6829631935,(403)
* Top\_5% High-Risk: RWJ1403,18.1748958787,(100)

SIS0042,2010-09-02,

* Current\_JSR:SIS0042,2.60520051973,(1517)
* Accumulate\_JSR:SIS0042,16.2067383814,(675)
* Top\_5% High-Risk: RRC0891,24.5808143172,(100)

TNB1616,2010-09-10,

* Current\_JSR:TNB1616,3.00842287976,(1104)
* Accumulate\_JSR:TNB1616,21.6602938862,(216)
* Top\_5% High-Risk: RRC0891,24.5808143172,(100)

TRC1838,2010-09-15,

* Current\_JSR:TRC1838,2.71276625203,(1838)
* Accumulate\_JSR:TRC1838,20.9071681848,(256)
* Top\_5% High-Risk:

MDS0680,2010-09-17,

* Current\_JSR:MDS0680,9.5576646684,(58)
* Accumulate\_JSR:MDS0680,25.7257622891,(69)
* Top\_5% High-Risk:

WDT1634,2010-09-20,

* Current\_JSR:WDT1634,1.44711529062,(1583)
* Accumulate\_JSR:WDT1634,4.22416417435,(1879)
* Top\_5% High-Risk:
* 突发奇想
* 同样是离职联系人的影响，但是也不能单纯计算一个用户的RLV吧？比如一个用户交友广泛，有100个好友，那么偶尔1-2个离职不会对该用户有太大影响；而如果该用户只有10个常联系的好友，其中1-2个离职那么比例就十分高了；应该再RLV前乘以各自的一个系数

OSS1463,2010-09-21,

* Current\_JSR:OSS1463,2.77549854171,（1221）
* Accumulate\_JSR:OSS1463,8.1597938501,（1668）
* Top\_5% High-Risk:
* **2010-07：出勤OSS1463,9.0,18.0,15.0,1.0,25,**
* **2010-08：出勤OSS1463,9.0,18.0,13.0,0.0,26,**
* **2010-07：只有298个联系用户**

CIF1430,2010-09-23,

* Current\_JSR:CIF1430,2.85702937458,（1144）
* Accumulate\_JSR:CIF1430,6.89144712295,（1752）
* Top\_5% High-Risk:

MCP0611,2010-10-06,

CHP1711,2010-10-13,

GWG0497,2010-10-15,

KSS1005,2010-10-16,

NAH1366,2010-11-17,

RRS0056,2010-12-10,

ICB1354,2010-12-15,

BYO1846,2010-12-15,

HXP0976,2010-12-20,

HMS1658,2010-12-30,

HIS1394,2010-12-30,

LVF1626,2011-01-14,

MGB1235,2011-01-21,

DCC1119,2011-01-26,

SNK1280,2011-02-11,

ITA0159,2011-02-17,

JAL0811,2011-02-25,

OKM1092,2011-04-29,

HSN0675,2011-04-29,

TMT0851,2011-05-11,

分析2010-08月

RWJ1403,8.0,18.0,17.0,6.0,22,

2010-07月份，有4个leave\_contacts，但是有705个一般联系用户；

该用户在2010-07被预测为100号高危用户，确实在8月份出现较多的迟到17天，早退7天，可以说非常严重的LJS表现；

但是排名第一的2010-07月/08月均如此  
HBP0009,7.5,16.5,0.0,0.0,27,

27天中，没有一天迟到早退，这是不对的！并且该用户也没有离职！

再看看排名第二的

SFJ0856,39.1067258933,

2010-08：SFJ0856,8.5,19.0,12.0,0.0,28,

2010-07：SFJ0856,8.5,19.0,13.0,0.0,27,

再看一个排名第三的：

BRM1080,35.7212167045, 虽然没离职，但是出勤表现极差

2010-07：出勤BRM1080,8.0,16.0,13.0,1.0,21,

2010-08：出勤BRM1080,8.0,16.0,14.0,6.0,22,

再看排名第四

LKM0230,34.9770479291,

2010-07：出勤LKM0230,8.0,18.0,12.0,3.0,27,

2010-08：出勤LKM0230,8.0,18.0,8.0,0.0,29,

这个是有迟到，但是没有早退

第一种情况，是引入一个迟到早退的分类器，作为区分；

第二个，则是从邮件联系人角度出发，即

2018年11月19日星期一

下午5时37分

现在，有两个可能的思路：

1. 从用户的出勤角度出发，作为一个子分类器，筛选出出勤表现差的一半作为LJSR表现；
2. 从用户的月度联系用户出发，分析leave\_contact所占所有联系人的比重（人数比重、发送的邮件数比重）从而为不同的人际中相等的离职RLV建立权值

2018年11月20日星期二

上午9时16分

我们先已缺勤达到一定比例（默认取一半）的标准筛选用户，出于预测考虑，每个月均列出缺勤高的一半人，可以使用迟到天与早退天比例和的中位数。

1. 统计每个月中用户的迟到与早退天，计算所占该月工作天的比例
2. 计算每个用户每个月迟到早退比例之和，然后该月所有用户MinMax；
3. 取该月中高于均值/高于中位数的一半用户，组成WorkTime\_RiskUsers，保存
4. 检查每个月份中，Insiders\_2用户如果存在，则其所处位置是否位于前一半（或者更少）？

问题1：

2010-09月份

HXP0976,7.5,16.5,7.0, 数据不全，遗失了late\_days和all\_work\_days

问题2：

每月比较，应只比较下月离职的insiders

先手动比较2010-07，预测2010-08月份离职的insiders

VCF1602,0.6481481481481483,877

**CKP0630,0.4365079365079365,1187**

ZIE0741,0.8253968253968254,405

均值是1082， 中位数是936， 当月原有用户1940