ATM 交易状态特征分析与异常检测

张海翔 庄煜洲 孙浩轩 仇嘉泽

2017 年 8 月 12 日

摘要

本文主要利用某商业银行总行数据中心监控系统所提供的数据进行了特征分析, 并且建立了相关的数学模型, 用以检测 ATM 系统可能存在的异常.

首先我们分析了交易量数据, 我们发现了其明显的天周期. 我们考虑将数据以天为间隔进行重新组织, 利用了主成份分析 (PCA) 与奇异值分解 (SVD) 对数据进行了降噪处理, 从而发现了其主要由最大的三个主成份决定. 因此我们对三个主成份的系数进行了分析, 发现短假期 (双休日) 对其没有显著影响, 而长假期 (春节) 对于其系数有较大的影响. 随后我们利用由 PCA 得到的特征进行异常检测. 通过最小二乘法我们实时对主成份的系数进行估计, 并对下一时刻的交易数据量进行了预测. 通过泊松分布我们算出了实际值的合理区间 (置信区间), 从而确定了异常检测的方案. 最后我们对于现有的数据进行了异常检测, 证明了此方法的高效性和可行性

第二部分我们分析了交易成功率的数据. 我们首先分析了交易成功率与时间的关系, 发现并没有明显的分布规律. 随后我们分析其与交易量之间可能存在的联系. 我们将每次交易是否成功作为随机变量考虑. 利用随机变量平均值的分布规律 (期望和方差), 我们将这些交易成功率数据进行了规范化, 从而发现了一些特征显著不同的点, 我们考虑这些点为异常点. 并且讲这些特征作为异常检测的依据. 最后我们对可能的异常点做了具体的分析, 发现确实为异常点, 从而证明了我们检测方法的可行性.

对于交易响应时间, 我们同样分析了其余时间时间可能存在的联系, 但是特征非常地不明显. 因此我们同样考虑其与交易量之间的联系. 首先我们发现绝大部分响应时间都在 600 毫秒以下, 因此我们默认响应时间不超过 600 毫秒的为正常交易. 最后我们根据数据可视化的结果, 猜测交易响应时间与交易量近似符合反比关系. 于是我们利用一个反比函数作为异常与非异常的分界线, 得到了若干可能的异常点. 最后我们对可能的异常点进行了分析, 发现其中一部分确实为系统故障, 而另一部分为网络波动. 由此我们说明了此异常检测方案的可行性.

最后我们讨论了可能的增加采集的数据: 每笔交易的金额, 每笔交易的操作时间和 ATM 机的使用率. 并对增加这些数据后可能的改进检测方案进行了讨论.

关键字: 异常检测　银行交易数据分析　主成分分析　泊松分布　大数据

目录

1 问题重述

1.1 引言

银行的 ATM 应用系统包括前端和后端两个部分. 前端是部署在银行营业部和各自助服务点的 ATM 机（系统）, 后端是总行数据中心的处理系统. 前端的主要功能是和客户直接交互, 采集客户请求信息, 然后通过网络传输到后端, 再进行数据和账务处理. 持卡人从前端设备提交查询或转账或取现等业务请求, 到后台处理完毕, 并将处理结果返回到前端, 通知持卡人业务处理最终状态, 这样完整的一个流程被称为一笔交易.

某商业银行总行数据中心监控系统为了实时掌握全行的业务状态, 每分钟对各分行的交易信息进行汇总统计. 汇总信息包括业务量、交易成功率、交易响应时间三个指标. 通过数据中心监控系统的数据, 总行可以对每家分行的汇总统计信息做数据分析, 来捕捉整个前端和后端整体应用系统运行情况以及时发现异常或故障.

1.2 问题的提出

作者围绕 ATM 机交易状态异常检测方案进行了探索, 分析了现有的数据, 并建立了数学模型研究以下几个问题:

1. 提取并分析了 ATM 交易状态的特征参数;
2. 设计了一套交易状态异常检测方案, 对该交易系统的应用可用性异常情况下做到及时报警, 同时减少虚警误报;
3. 设想了通过增加采集的数据, 提升检测系统的效率与准确性.

2 问题分析

2.1 问题一的分析

问题一要求我们提取并分析 ATM 交易状态的特征参数. 首先我们整理了数据缺失的点, 然后分三类数据分别进行了分析.

交易量数据 由于交易量数据量较大且其关于天为周期有近似的周期, 所以我们首先依据已有的 ATM 交易状态数据 (共 131013 条) 进行了重新排列, 使得其排列为一个 91×1440 的矩阵, 其中任意一个 (i, j) 元对应于第 i 天第 j 分钟的交易量数据, 记此矩阵为 N. 我们对 N 进行主成份分析 (PCA), 因此我们对 N 进行了奇异值分解, 得到奇异值矩阵 Σ, 对 Σ 进行分析后发现其主成份主要为其前三个元素. 因此我们将其前三个元素提取出来, 而将其余 88 个元素当作噪声处理. 再将处理后的数据与原数据进行对比, 发现拟合度较好.

交易成功率数据 首先我们假设了每次交易是否成功为独立同分布的事件, 然后在交易次数为 $n(n \in N_+)$ 的条件下计算了试验成功次数的置信水平为 0.95 的置信区间 (小于 1 的方向). 然后统计了在置信区间以外的数据.

交易响应时间数据 首先我们计算了交易响应时间乘以交易量的数据, 对其进行分析后发现大部分点都在某条水平直线之下. 随后我们对直线高度进行了测试, 发现存在某一数值, 明显地将点集分成了两部分, 因此我们将其作为异常点的判别准则. 之后我们还讨论了导致这一现象可能的原因.

2.2 问题二的分析

问题二要求我们设计一套交易状态异常检测方案. 我们利用第一问中所选择的特征参数和特征参数的分析结果, 给出了可行的异常检测方案.

交易量异常 我们结合 PCA 所得结果进行即时拟合, 并利用泊松分布模型得到异常点判别准则（其置信区间的大小与交易量的拟合预测值正相关）. 最后对于现有的数据进行了异常分析.

交易成功率异常 我们利用问题一中得到的判别准则 (置信区间), 首先对已有数据进行了分析, 得到了 4 处可能的异常点. 随后进行了逐点查看后发现确实存在异常, 因此验证了此方法的可行性.

交易响应时间异常 同样, 我们利用问题一中发现的规律对已有数据进行分析, 得到了 10 处可能存在异常的时间. 进行逐点分析后发现有 4 个点确实存在异常, 其余 6 个点亦为网络异常波动. 从而验证了此方法的可行性.

2.3 问题三的分析

问题三要求我们提出对于提升问题二中异常检测系统性能有帮助的扩展数据, 并且针对采集了这些扩展数据的情况, 对问题二中的检测系统进行改造. 我们认为需要增加采集的数据为每笔交易的金额, 每笔交易的操作时间和 ATM 机的使用率.

3 模型假设

- 我们不考虑无法对交易数目, 交易成功率和交易响应时间造成影响的异常
- 未考虑数据采集系统可能出现的异常
- 每次交易是否成功为独立同分布的随机事件

4 符号说明

符号	意义
U, V, Σ	奇异值分解矩阵,U,V 为正交矩阵,Σ 为对角矩阵
$\Sigma_1, \Sigma_2, \Sigma_3$	最大的三个奇异值
f_1, f_2, f_3	最大的三个奇异值对应的主成分
a_n, b_n, c_n	第 n 天 f_1, f_2, f_3 前系数
X_i	描述每次交易是否成功的随机变量
η, η_n	随机变量的期望
σ, σ_n	随机变量的方差
N	交易量
ρ	相对误差关于交易量的函数
β	置信水平
M	每笔交易金额 (元)
T	每笔交易操作时间 (秒)
R	ATM 机的使用率

5 数据分析及特征参数的选取

5.1 交易量

首先我们提取出原数据数组的日期参数 (第一列)、时间参数 (第二列) 和交易量数据 (第三列). 为了获得对数据直观的了解, 我们首先将数据进行了可视化处理 (见图1). 从图中我们可以发现交易量数据大致以一天为周期, 因此我们考虑对交易量数据进行变形, 使之成为一个 91×1440 的矩阵, 称其为 $DayData$(其中 91 对应于数据覆盖的总天数,1440 对应于一天的 1440 分钟). 因此矩阵的每一行就对应了某一天内交易量数据的变化情况.

我们考虑对矩阵进行主成份分析 (PCA)[1]. 由于 $DayData$ 非方阵, 因此我们考虑使用奇异值分解 (SVD)[2], 即计算出正交矩阵 U, V 以及对角矩阵 Σ, 使得

$$DayData_{91\times1440} = U_{91\times91} \times \Sigma_{91\times1440} \times V^T_{1440\times1440}. \tag{1}$$

我们作出了 Σ 中 91 个奇异值由大到小的排列图像 (见图2). 我们发现其主成份主要为前三个特征值 ($\sigma_1 = 279320.0, \sigma_2 = 21890.0, \sigma_3 = 13378.7$) 占了主导地位, 因此我们将前三个成分作为主成份, 而其余成分作为噪声. 我们随机抽取了几天观察其近似情况, 发现其去噪后结果与原数据拟合的较好 (见图3), 因此我们利用去噪后的结果以及三个主成份的系数来进行数据分析.

首先我们分析交易量数据随天数变化的规律. 我们作出了三个主成份对应的系数随天数变化的曲线 (见图4). 我们发现其奇异值对应的主成份的系数在前 20 天产生了较大的波动. 分析实际数据, 我们发现其对应于 1 月 23 日至 2 月 11 日, 恰好处于农历春节期间, 因此对于银行交易量造成了影响. 在第 20-91 天中, 三项系数随天数变化的规律并不明显, 其中最大成分的系数基本保持稳定, 另外两个成分的系数有近似周期 7 的波动, 但是规律不明显. 因此交易量与周中和周末并没有明显的联系, 仅仅因为的长假 (春节) 到来而产生波动.

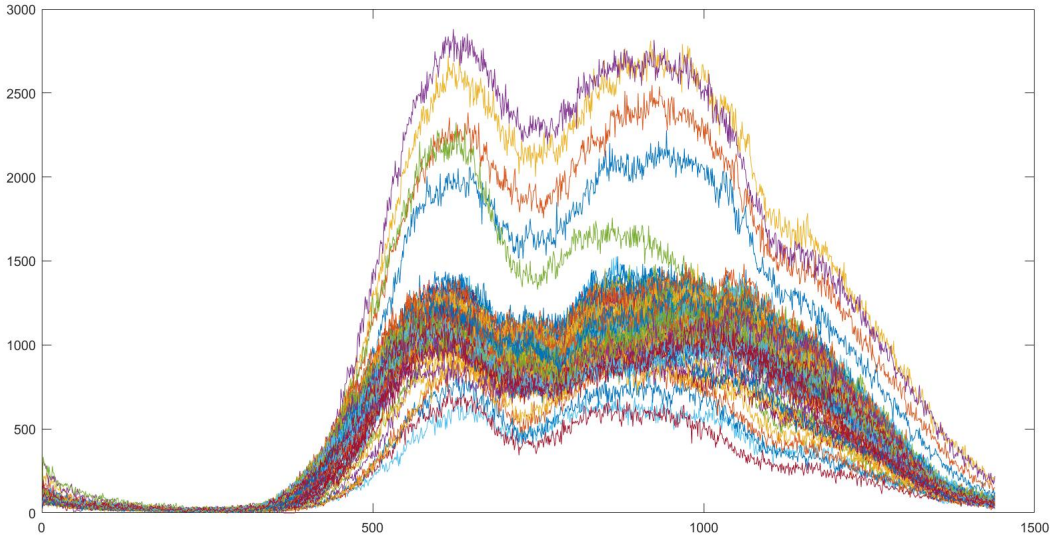

图 1: 交易量数据随时间分布曲线

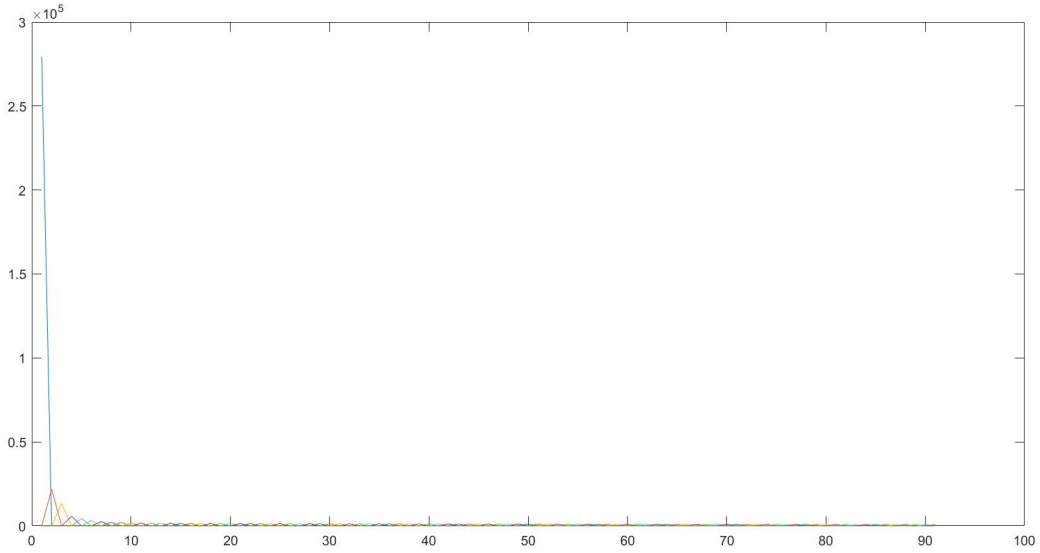

图 2: 奇异值大小分布

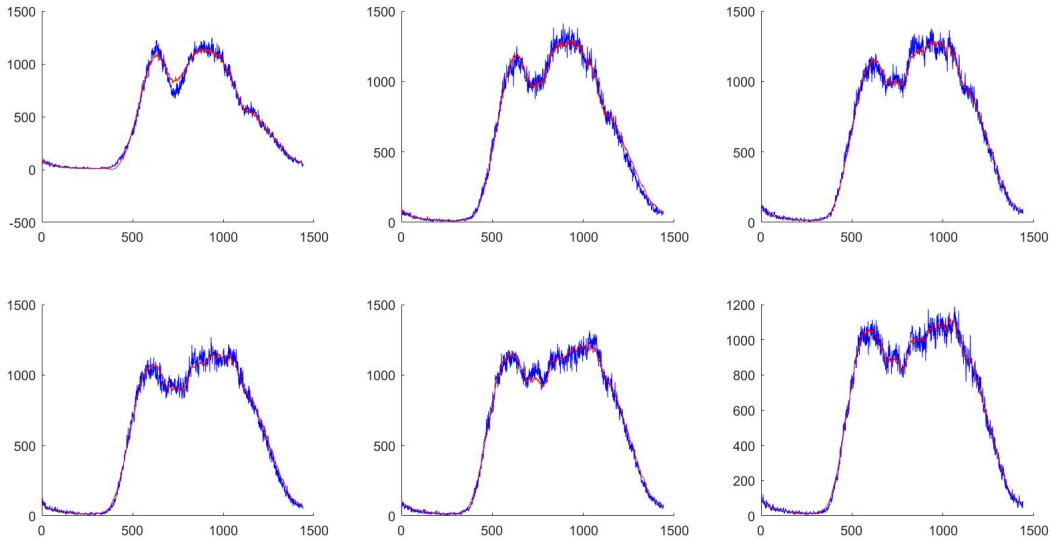

图 3: 去噪前后数据对比 (红色为去噪后)

随后我们分析数据在一天内的分布情况. 我们作出三个主成份对应的数值在一天内分布的曲线 (见图5). 其中每天的交易量数据分布近似由最大的成分决定, 因此其形状也与其非常相似. 我们发现其图像中存在两个高峰段, 其分别分布于 9 点-11 点和 14 点-18 点. 而在中午时段 (11 点-14 点) 处出现了一个局部低谷段. 在 17 点之后, 数据量逐渐变小, 在凌晨 4 点左右达到极小, 而后逐渐增大. 分析其背后的原因, 我们认为上午及下午为大部分人办理业务的主要时间, 而中午时段大部分人选择休息, 而在晚上交易量自然处于低谷期.

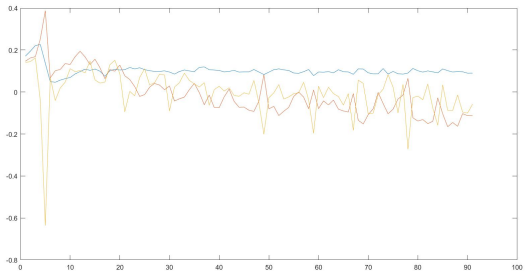

图 4: 主成份系数随天数的变化曲线

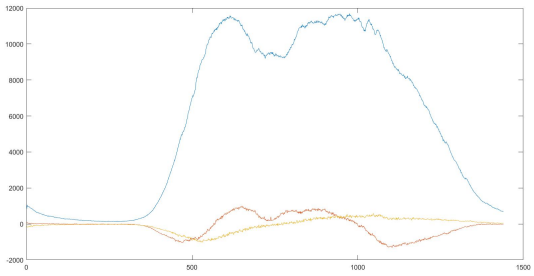

图 5: 主成份数据在一天内的变化曲线

5.2 交易成功率

首先我们作出了交易成功率随交易量变化的散点图 (见图6), 发现在小交易量处交易成功率波动较大, 而在大交易量处较为稳定, 其取值大约在 0.96 附近. 因此我们不能仅仅利用交易成功率的大小来衡量其是否出现异常. 因此我们需要制定一个合理的特征来刻画成功率的大小.

首先我们假设每次交易是否成功是一个随机变量 X_i. 考虑到正常情况下各次交易 X_i 是否成功是独

立同分布的事件, 设其期望值为 η, 方差为 σ^2. 我们取 η 的估计值为交易成功率的平均值 0.958561. 所以我们统计 n 次交易成功率时计算的 $Rate_n = \frac{1}{n}(\sum_{i=1}^{n} X_i)$ 的平均值 $\eta_n = 0.958561$, 方差 $\sigma_n^2 = \frac{\sigma^2}{n}$. 我们先将各个 $Rate_n$ 规范化, 即考虑各个 $Rate_n$ 和 η 的差值, 再将其乘以 $\sqrt{n}$, 得到期望为 0, 方差为 σ^2 的随机变量.

$$Normalized\ Rate = (Rate_n - 0.958561) \times \sqrt{n} \tag{2}$$

作出各个时间的规范化后的成功率, 结果如图7. 我们发现仅有四个时间段此变量的值大于 2, 我们推断这些时刻为异常时刻. 详细的讨论见6.2节.

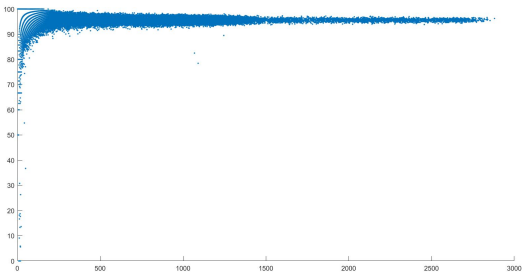

图 6: 交易成功率随交易量的分布

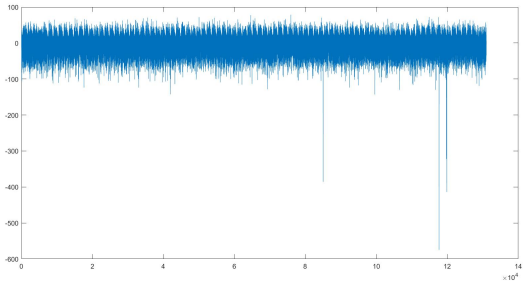

图 7: 规范化后交易成功率随时间的分布

5.3 交易响应时间

首先我们作出了交易响应时间与交易量的关系 (见图8). 我们发现在交易量较大的时候其值基本都在 600 毫秒以下, 而在交易量较小时其交易响应时间有较大的波动. 因此我们默认 600 毫秒为系统正常的响应时间. 我们还发现在小交易量位置有一段点集近似于反比函数的轨迹, 且有仅有少量点位于轨迹之上. 我们对此作如下可能的解释: 其中数据传输的耗时大约为 600 毫秒且每个任务处理所需要的时间近似相等, 而系统中同时运行的任务 (线程) 数近似与总任务数成正比, 因此耗时减去 600 毫秒后的平均耗时近似与总交易量成反比.

我们考虑这些位于反比函数之上的点为可能的异常点. 对于响应时间不超过 600 毫秒的点, 我们已经假设了其响应时间不属于异常. 对于超过 600 毫秒的点, 我们计算 $(Time-600)\times Number$, 其中 $Time$ 为交易响应时间,$Number$ 为交易数目. 我们发现若取 55000 作为阈值, 所有可能的异常点 (即图9中的红点) 对应的值都在其以上. 因此我们利用 $(Time - 600) \times Number > 55000$ 作为判别系统是否出现了异常的标准.

6 异常检测方案

6.1 交易量异常的检测

利用在5.1节中得到的结果, 我们发现每天的交易量数据分布大致都由三个最大成分前的系数所决定. 记这三个主成份依奇异值大小排列为 f_1, f_2, f_3. 因此每天的数据分布大致可以表示为 $Data_n(t) = a_n f_1(t) + b_n f_2(t) + c_n f_3(t)$, 其中 $a_n, b_n, c_n \in R$, $1 \le t \le 1440$. 我们只要对 a_n, b_n, c_n 作出尽可能精确的估计, 就可以得到对一天交易量分布更加精确的估计. 因此如果我们利用当天当前时刻前已有的数据对

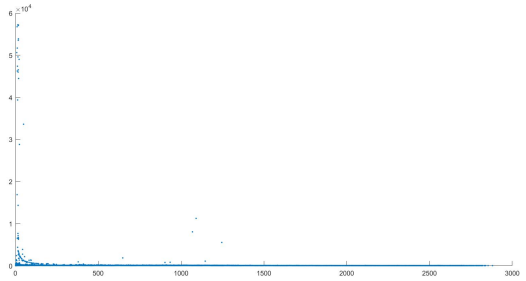

图 8: 交易响应时间随交易量的分布

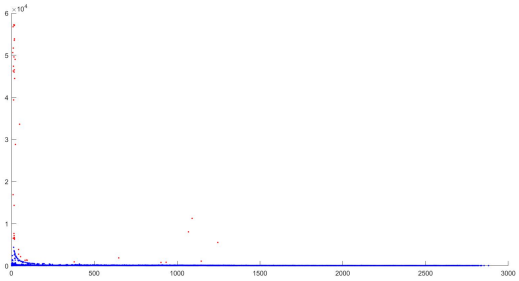

图 9: 判定可能的异常点 (Red)

a_n, b_n, c_n 进行估计, 就可以预测后一时间段可能的交易量大小. 再利用这一预测值与实际值进行比较, 如果超出了合理的噪声大小 (置信区间), 那么就可以说明产生了异常. 图10为我们随机选取若干天进行预测得到的结果. 而且此预测方法对于一天交易数据的 1439 次预测总耗时平均为 0.0385 秒. 因此我们的预测方法是精确且高效的.

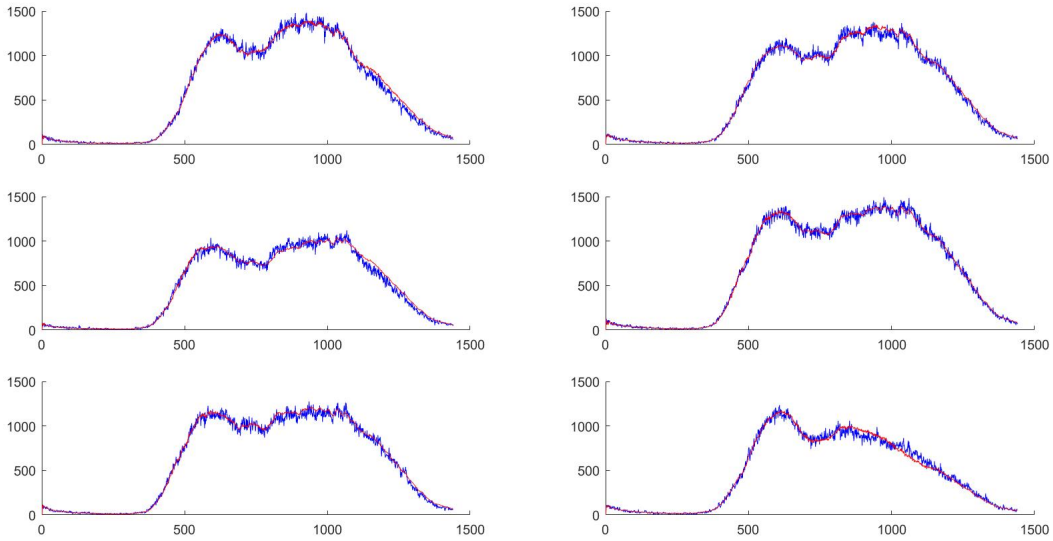

图 10: 随机若干天预测数据 (Red) 与实际数据 (Bule) 对比

首先我们给出置信区间的确定方法. 首先我们对图1进行观察, 发现在交易量较高的时候其交易量的相对波动较大, 而交易量较小时相对波动较小, 因此采用一个与交易量无关的阈值来判断异常是不合适的. 因此我们对于每一个交易量 N 计算允许的相对误差 $\rho(N)$(主要计算减小的相对误差).

考虑到我们的交易量数据满足以下两个特征:

1. 在任意一个非常短的时间段 δt 内至多只有一个人前往 ATM 机;
2. 各个时间段内前往 ATM 机的人数仅与时间长度有关.

因此我们考虑实际值为一个服从泊松分布的随机变量, 其参数为我们的预测值, 即

$$Data(t) \sim P(Expectation\ Data(t)) \tag{3}$$

我们只要计算出实际值满足置信水平 β 的置信区间, 只要实际值超出了这个置信区间, 我们即认为其可能存在异常. 通过计算泊松分布的置信区间, 我们得到了交易量为 1-1900 之间允许的波动率 (见图11), 而对于大于 1900 的交易量, 我们直接取 0.08 作为置信区间. 随后我们对现有的数据进行了统计, 没有发现现有数据中存在异常点.

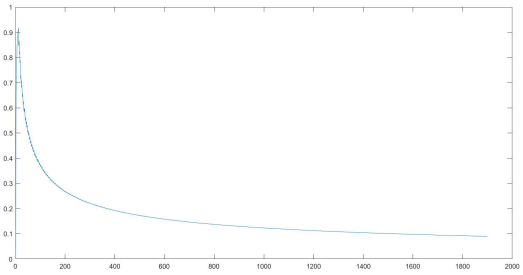

图 11: 不同交易量数据下允许的相对波动大小

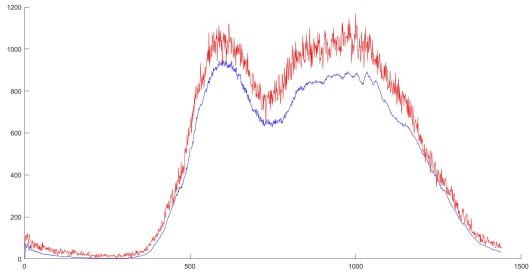

图 12: 某天预测值 (Red) 与异常阈值 (Bule) 之间的关系

6.2 交易成功率异常的检测

在5.2节中我们发现有四处规范化后的成功率指标远大于其他值, 实际查看发现对应时间段分别为 3 月 23 日 0 点 48 分至 1 点 03 分;4 月 14 日 17 点 33 分至 17 点 35 分;4 月 16 日 4 点 01 分;4 月 16 日 6 点 00 分至 6 点 03 分, 逐项分析发现这些时间段如下:

3 月 23 日 0 点 在原数据中找到这些数据, 发现从 0 点 48 分开始, 交易成功率突然降至 18.18%, 交易时间增至 46256 毫秒. 而到 1 时 02 分, 虽然成功率升至 79%, 但是相对于其交易数目而言仍然是小概率事件, 而且参考其响应时间 (40000 毫秒左右), 我们认为直至 1 点 02 分异常依然存在. 而到了 1 点 03 分, 异常基本解除.

4 月 14 日 17 点 在 17 点 33 分, 交易成功率突然降至 82.51%, 虽然不是很低, 但是参照其交易次数仍是小概率事件, 而且交易时间非常长 (8072 毫秒). 直到 17 点 36 分, 数据才恢复正常, 因此我们认为 17 点 33 分至 17 点 35 分之间系统发生了故障.

4 月 16 日 4 点 在 4 点 01 分与 4 点 02 分处, 交易成功率突然降至 17.65% 和 62.5%, 且交易响应时间增至 10000 毫秒量级. 因此这段时间系统也出现了异常.

4 月 16 日 6 点 在 6 点 00 分至 6 点 03 分, 交易成功率降至 36.73% 与 0%, 响应时间也增至 30000 毫秒以上. 因此这段时间出现了异常.

综上所述, 我们计算规范化后的交易成功率得到的异常点确实发生了异常. 因此我们认为此种异常检测方法是合理并且有效的, 即当 $Normalized\ Rate_n = (Rate_n - 0.958561) \times \sqrt{n} > 2$ 时, 我们认为系统出现了异常.

6.3 交易响应时间异常的检测

在5.3节中, 我们提出了一个特征参数 $(Time - 600) \times Number$. 我们取阈值 55000 得到了所有可能的异常时间点. 以下我们对这些可能的异常点做分析:

单点波动 在可能的异常点中, 有 6 处非连续的异常 (1 月 28 日,2 月 25 日,3 月 9 日,3 月 28 日,4 月 2 日,4 月 3 日), 我们将其归做网络波动, 因此不属于系统异常.

实际异常 统计到的异常中有 2 月 9 日 2 点 17 分,3 月 23 日 0 点 48 分,4 月 14 日 17 点 32 分及 4 月 16 日 4 点 01 分出现了持续的异常, 我们将其归做系统异常. 其中 3 月 23 日,4 月 14 日 4 月 16 日在6.2中已经讨论过了. 我们分析 2 月 9 日 2 点 17 分的数据发现出现了持续 15 分钟的系统延迟 (响应时间为 7000-10000 毫秒), 因此此时刻系统出现了异常.

综上所述, 我们找到了 6 个系统波动和 4 个系统异常, 因此我们的异常检测方案是可行的.

6.4 节总结

结合以上三节的讨论, 我们找到了 5 个异常点. 其中 3 月 23 日 0 点,4 月 14 日 17 点与 4 月 16 日 6 点为交易成功率和交易响应时间同时出现异常, 我们分析其为数据中心后端处理系统应用进程异常. 而 4 月 16 日 4 点仅有交易成功率出现了异常, 我们分析其为分行侧参数数据变更或者配置错误. 对于 2 月 9 日 2 点, 仅有交易响应时间异常, 我们分析其为数据中心后端处理系统异常.

7 扩展数据及可能的检测方法改进

因此我们考虑增加采集如下的数据:

1. 每笔交易的金额 M（或者每分钟内所有交易的平均金额）;
2. 每笔交易的操作时间 T（或者每分钟内所有交易的平均操作时间）;
3. ATM 机的使用率 R, 即被使用时间和总时间的比例.

无论是分行侧网络传输节点故障、分行侧参数数据变更或者配置错误还是总部数据中心后端处理系统异常, 都将影响用户实际使用 ATM 机进行交易的交易时间 (T) 和 ATM 机的使用率 (R), 所以我们通过对 T 和 R 进行类似的异常分析, 进而交叉对比多种方式所得的评估结果, 可进一步改进已有的交易状态异常检测方案的可行性和准确性, 同时还可能可以分析出一些仅依据原有的三种数据无法检测出的异常交易进程, 如分行侧所属的 ATM 机局部故障（而非分行侧整体故障）、ATM 机与分行之间的网络传输故障. 特别是在检测 ATM 侧与分行端之间的异常情况时,T 和 R 有较大的参考价值.

而通过对 M 的分析则可以获知 ATM 系统的整体现金流和资金流是否出现异常, 进而可能检测出 ATM 现金交易部件异常或故障、ATM 现金储量异常、不正常取现行为等情况, 进一步完善和提升监控系统对各类异常和故障情形的检测能力.

8 模型的评价与改进

8.1 模型的优点

1. 利用奇异值分解, 较好的拟合出了一天内的交易量变化规律, 进而提取出交易量变化特征, 并能较好地预测未来一段时间内交易量的变化规律
2. 利用随机变量的泊松分布模型, 给出了一种基于交易量预测数值计算对应的交易量置信区间的具体方法. 这一模型的假设较为合理, 且计算代价较低, 能对异常情况做出及时预警, 故能为故障预警提供一种有效的判别方法
3. 虽然春节前后交易量波动变化较大, 该模型仍能较好的拟合及预测其变化趋势, 可见模型对于各类情景的普遍适用性
4. 通过大量数据的计算与分析, 发掘出了交易量与交易响应时间的数值关系所蕴含的实际原理

8.2 模型的不足

1. 模型以天为周期进行预测评估, 对每天较早时刻的预测样本较少, 有潜在的误差较大的风险
2. 模型建立仅基于已有数据, 故特征参数计算可能存在一定误差, 可能对预估数据及故障预警存在一定程度的影响

8.3 模型的改进与推广

1. 基于更多数据样本对模型加以修正, 减小误差, 削弱误差影响
2. 基于交易金额、操作时间、ATM 使用率等扩展数据, 拓宽交易异常的判别依据, 进一步完善故障预警方案
3. 节假日、经济政策调整等因素会很大程度上影响个别时段的交易量, 可对模型进一步优化, 使其能够适应更多社会因素对数据规律的影响

参考文献

[1] Wikipedia. Principal component analysis. https://en.wikipedia.org/wiki/Principal_component_analysis

[2] Wikipedia. Singular value decomposition. https://en.wikipedia.org/wiki/Singular_value_decomposition