ATM 机交易异常时的预警告警机制及优化方案

B 题

摘要

随着金融电子化的发展，ATM 机在金融行业的应用越来越广泛。然而传统的 ATM 监测系统缺少能够快速主动识别并判断异常种类的功能，所以在交易系统发生故障时，难以做到准确报警和及时预警。为了解决这一问题，我们建立了基于数据的统计模型。在分析了系统发生异常的原因之后，找到报警的产生模式，并将这些模式运用的预警当中，建立快速高效的故障响应机制。

针对问题一，由于工作日和非工作日的交易量存在差别，因此需将已有数据按过年前、过年期间、休息日（双休日和小长假）和正常工作日分别进行分析。定义交易量峰谷值这一特征参数，之后运用统计方法筛选出交易量骤降的异常数据点；通过分析响应时间与交易量之间的函数关系构造 CPU 载荷这一特征参数；利用方差分析法中的双次 N 值比较法，得出后端 CPU 载荷正常范围的阈值为 8.32，筛选出后端 CPU 载荷过大所导致的异常数据点；通过构建自组织映射模型（SOM），将数据聚类后筛选出离群数据点。通过三种方式筛选，共得到已给数据中的 10 条异常交易记录，如 4 月 16 日出现的数据中心后端处理系统应用进程异常、4 月 14 日发生的后端操作系统 CPU 载荷过大。

针对问题二，建立交易量的差分自回归模型，定义数据下降指标，观察其历史分布，得到交易量骤降异常的监测判据。分析四个特征参数（CPU 载荷、交易量、响应时间、成功率）间的相关性。利用问题一中提取出的异常数据点，分工作日和非工作日构造朴素贝叶斯分类器，定义四种不同的交易状态。对于某一时刻，求出对应不同交易状态时的先验概率。模型根据先验概率的大小可以精确地判断系统的状况，异常发生时，模型能实时给出异常类型及位置。定义正常状态的先验概率比例作为 ATM 系统的健康度，并且定义健康度下降指标来实时分析 ATM 系统交易状态。经过统计筛选，确定健康度下降指标大于 0.2，且分类器显示状态为正常时，进行预警；而一旦分类器显示状态为异常时，认为系统处于异常状态，进行报警。在代入部分真实数据对模型进行验证后，发现能在异常发生前1～2 分钟提出预警。通过改变系统参数的方法对模型进行灵敏度分析，发现系统灵敏度合适，同时为不同情况下下降指标的选取提供依据。

针对问题三，如果提供更多关于ATM机交易情况的指标，如交易种类、每条交易的记录（每笔交易时间、ATM机编号、是否成功、响应时间），就能构造更精确的特征参数，实现对系统实现局部监测，判断出具体发生故障位置；若增加数据为全年甚至连续几年的交易情况，则可以分析有无年周期性或季节性因素影响，建立更为精确的梳系数ARIMA时间序列模型。同时，基于更大的训练

集，能够提高贝叶斯分类器的精度，使得分类结果更为准确。通过大量数据的实验与拟合，验证了所建模型的正确性，并可将其推广应用到解决信息传输的监测、信号编码和传输的监测等数据流问题，从而帮助服务商和用户及时发现和解决问题。

关键词：自组织映射神经网络（*SOM*） 方差分析 *ARIMA*时间序列 贝叶斯分类器

目录

一、问题重述

1.1 问题背景

随着中国金融电子化建设的深入发展和银行客户对金融服务质量要求的提高，ATM 机在金融行业的应用越来越广泛。然而，很多 ATM 机并没有充分发挥其先进的作用，有些 ATM 机经常是故障连连，无法提供服务；有些 ATM 机则经常发生错账或吞卡现象。除了 ATM 机自身会出现机器故障，ATM 交易系统也常会因为软件故障或传输网络故障，造成 ATM 机无法正常使用。因此如何对于银行 ATM 交易系统的发生的异常进行预报和告警，从而对 ATM 交易系统进行优化，使得其能够更好的服务大众，成为了银行在优化服务时的关注点之一。

银行的ATM交易系统包括前端和后端两个部分。银行总行数据中心监控系统通过汇总统计每家分行的业务量、交易成功率、交易响应时间，来做出数据分析，从而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。

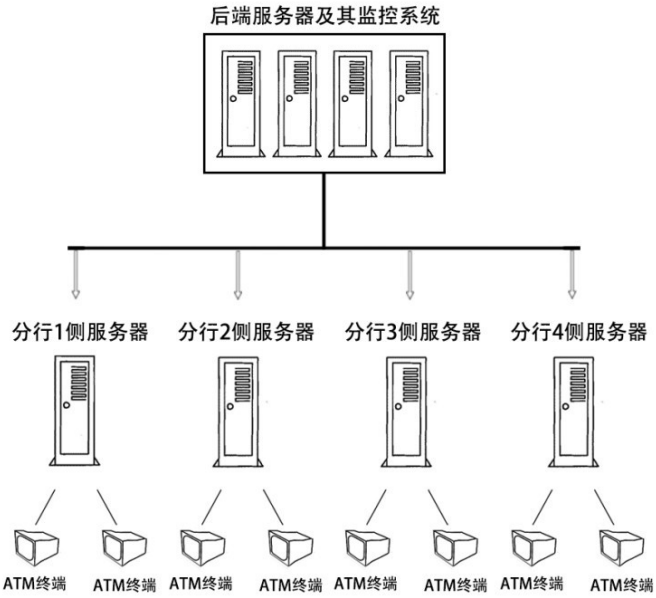

图1：ATM系统拓扑网络结构图

ATM常见的故障场景包括但不限于如下情形：

（1）分行侧网络传输节点故障，前端交易异常，导致业务量陡降；

（2）分行侧参数数据变更或者配置错误，前端交易异常，影响交易成功率指标；

（3）数据中心后端处理系统异常（如操作系统CPU载荷过大），影响交易响应时间指标；

（4）数据中心后端处理系统应用进程异常，导致交易失败或响应缓慢。

1.2 待解决问题

附件中给出了某商业银行ATM 应用系统某分行的交易统计数据。需要建立数学模型解决的如下问题：

（1）选择、提取和分析ATM 交易状态的特征参数；

（2）设计一套交易状态异常检测方案，在对该交易系统的应用可用性异常情况下能做到及时报警，同时尽量减少虚警误报；

（3）若可增加采集的数据，基于需扩展的数据，寻找达到任务（1）（2）中目标的方法。

1.3 研究现状

近十几年来，伴随着信息化的推进，基于数据分析的预警和报警系统越来越受到人们的重视，成为数据科学研究的重点。现有的主要手段包括以下几种：

（1）统计分析。通过大量的历史数据生成系统的正常行为轮廓，自适应的学习系统的正常行为模式。五个经典的异常检测的统计模型：操作模型、平均值和标准差模型、多元模型、马尔科夫过程和时间序模型[1]。

（2）智能算法。采用仿生的神经网络模型，可以有效的对系统进行有监督或无监督学习，归纳出型的输入输出关系，由此进行判断。同时神经网络可以和一些智能算法如遗传算法、退火算法结合起来，提高有效性[2][3]。

（3）贝叶斯技术。贝叶斯技术是一种概率论的推理技术。它将时间的先验概率与后验概率联系起来，利用先验信息和样本数据确定时间的后验概率。

（4）模式预测。模式预测的目标则是在样本特征和样本标签之间建立起有一种有效的映射关系。Teng 将时间序列和模式预测结合起来，提出基于时间的推理方法，将偏离预测的行为看成是异常[4]。

二、问题分析

2.1 问题一分析

由于工作日和非工作日的交易量存在差别，因此首先将已有数据分为过年前、过年期间、休息日（双休日和小长假）和正常工作日四类。题目中要求提取并分析特征参数，通过观察和分析附件中给出的数据，发现若仅选择已交易量、成功率、响应时间作为特征参数，不能够全面的反映导致交易异常的原因。因此在分析三者之间的相互关系后，还需提取出新的指标作为特征参数。之后对四类交易日期分别进行假设检验，分析参数变化与对应交易发生时间的关系以及参数间是否相互独立。综合上述分析结果后，对1～4 的数据进行处理，筛选出异常点。

2.2 问题二分析

题目中的给出四种故障最终都通过业务量、成功率、响应时间以及三者间的关系反应出来，所以以问题一中提取的特征参数为依据，对这些异常进行刻画。为了及时报警同时减少虚警误报，运用数理统计中的贝叶斯定理，根据问题一中已筛选出的异常点计算出每种特征参数可能出现异常的先验概率，并通过选择合适的分类器，对各个异常点进行分类报警。构建趋势型指标可以提前对系统出现的异常进行预警。

2.3 问题三分析

考虑基于横向的数据拓展，通过建立新的参数，对新信息进行分析建模，可以细分系统的交易状态，以进一步剔除随机因素以及操作人的影响。考虑基于纵向数据，可为问题二模型提供更多样本，整体提高系统的精度。

三、符号说明

s	步长指数
t	训练样本的指数
$D(t)$	输入向量
u	$D(t)$ 的 BMU 指数
$\alpha(s)$	学习系数
p	自回归项数
q	滑动平均项数
d	成为平稳序列所做的差分次数
$\hat{X}_k(q)$	预报向量
$\hat{X}_k(m)$	时间序列
U	实时的数据下降指标
C	独立的类别变量
$F_1,\ldots,F_N$	特征变量
n	特征数量
Z	证据因子
$p(C)$	类先验概率
$p(F_i \mid C)$	独立概率分布
S	交易状态，$S \in \{R, E_1, E_2, E_3, E_4\}$
r_{ij}	相关系数矩阵
λ	特征值
X_I	特征参数

	$X_I \in \{$交易量，响应时间，成功率，CPU载荷，预测交易量$\}$
b_j	信息贡献率
$s=\frac{p(R)}{\sum p(S)}$	系统健康度，$S \in \{R, E_1, E_2, E_3\}$
$D(i)$	W 时间窗口内递减的分钟数
W	最小时间窗口
η_i	第 i 天的健康度下降指标
S_l	第 l 天的健康度
$T=\frac{\sum_{n=i-W}^{i} D(n)}{W+1}$	健康度下降指标

四、模型假设

1、ATM 系统的数据仅与系统状态有关，而与持卡人的操作水平无关。
2、假设题目中所给数据是 2017 年，工作日和节假日按照 2017 年划分。
3、假设日交易总量差异仅与工作日、节假日有关，不存在洗钱等行为。
4、假设仅考虑该银行 ATM 机前后端系统应用程序故障。
5、假设所给数据均为真实可靠的。

五、模型的建立与求解

5.1 数据的预处理

（1）数据分类

首先，对整体数据进行宏观上的探索。根据数据交易量的时序变化以及银行业务的特点，发现双休日的日交易量略多于工作日，而节假日（即春节前后和清明节三天小长假）的日交易量则明显多于工作日和双休日。于是根据日交易量将交易日期分为工作日、双休日和节假日，便于后续对于数据的准确使用。

（2）数据清洗

在数据来源可靠的条件下，进行数据挖掘之前，还必须对数据的质量进行评估。因为已有数据大多来自计算机的自动获取，所以结果有一定几率出现异常，因此在对原始数据进行缺失值、重复值的筛选与处理后，发现原始数据中没有重复，但有 27 组数据缺失。

表 1：1~4 月缺失的数据

日期	时间	日期	时间

1 月 28 日	07：31	1 月 31 日	07：21
1 月 28 日	08：22	3 月 19 日	07：18
1 月 29 日	07：24	3 月 19 日	07：19
1 月 29 日	08：21	3 月 30 日	05：28
1 月 30 日	07：28	4 月 16 日	10：04～ 10：21

在处理非连续缺失数据时，选择利用均值填补遗漏值，如：1 月 31 日缺失的数据采用上下相邻的两分钟数据平均值补齐；在处理连续缺失数据时，则采用同类别均值填补遗漏值，如 4 月 16 日缺失的数据采用前一天的数据补齐[5]。

5.2 问题一模型的建立与求解

5.2.1 交易量骤降

（1）模型分析

由于交易量在一天中的变化趋势相似，但统计指标的数值相差大，和具体的日期相关性强，且数值易受到多种随机因素的影响，噪声较大。所以通过构造峰-谷值（ω）这一特征参数并结合统计分析，准确的筛选出交易量异常值。

对于第i分钟的交易量d_i，其峰-谷值（w）为

$$\omega=\frac{d_{i-3}+d_{i-2}+d_{i-1}+d_{i+1}+d_{i+2}+d_{i+3}}{6}-d_i$$

通过对第i分钟前后三分钟取平均，来准确地描述骤降的含义，同时减小随机因素对结果的影响，提高结果的可靠性。

考察ω的统计分布如下

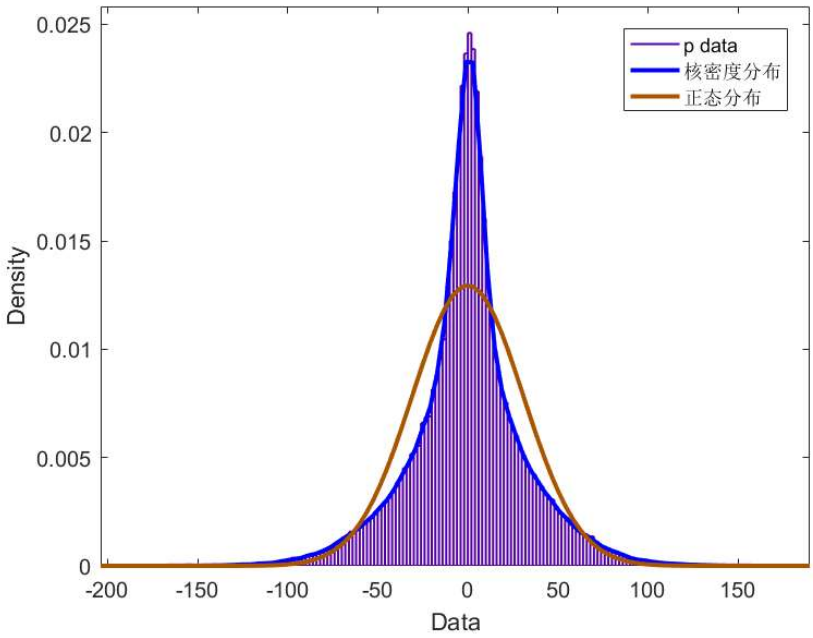

图 2：峰—谷值ω的分布情况

通过分析图 2 知，相较于正态分布，ω分布峰度更高、尾部更肥，统计学上将这种分布特点称为“尖峰肥尾”，其中“肥尾”解释为信息偶尔以成堆的方式

出现，而不是以平滑连续的方式出现。

为了合理检测出数据中不符合统计规律的点，采用与正态分布中3δ分位数一致的点作为异常检测的阈值，即认为数据中有 99.74%的数据为正常值[6]，其余的为异常值。

由此得到峰-谷值（ω）的阈值为ω=-131，此时有

$$\begin{cases}\omega \le -131 & \text{异常}\\ \omega > -131 & \text{正常}\end{cases}$$

根据ω的阈值确定出原始数据中的 65 组异常数据（已经剔除掉了数据量突增的点）。同时因为故障在出现后会持续一段时间，所以在去除某一分钟突然出现交易量下降而下一分钟交易量又恢复正常的数据时间点后，最终得到以下两个交易异常情况

1 月 25 日 16:04~16:06
1 月 26 日 13:26~13:27
2 月 10 日 16:28~16:32

5.2.2 基于自组织特征映射神经网络的异常值筛选模型（*SOM*）

（1）模型分析

相比于交易量这一特征参数，响应时间和成功率不具有明显的分布特征。因此在分析响应时间和成功率时，采用机器学习这一方式，通过聚类来判断数据的分布情况。由于维数和样本量庞大，而有监督的机器学习在训练过程中，需要预先给网络提供期望输出，因此对于本题中无期望输出、无监督的情况下，选择无监督学习中自组织特征映射网络（*SOM*）来解决问题。

（2）模型简介

1981 年芬兰*Helsink*大学的*T.Kohonen*教授提出一种自组织特征映射网，简称*SOM*网，又称*Kohonen*网。*Kohonen*认为：一个神经网络接受外界输入模式时，将会分为不同的对应区域，各区域对输入模式具有不同的响应特征，而且这个过程是自动完成的，是一种典型的无监督学习。自组织特征映射正是根据这一看法提出来的，其特点与人脑的自组织特性相类似。

通过训练，建立起这样一种布局：它使得每个权值向量都位于输入向量聚类的中心。一旦*SOM*网完成训练，就可以用于对训练数据或其他数据进行聚类[7]。

（3）模型的建立

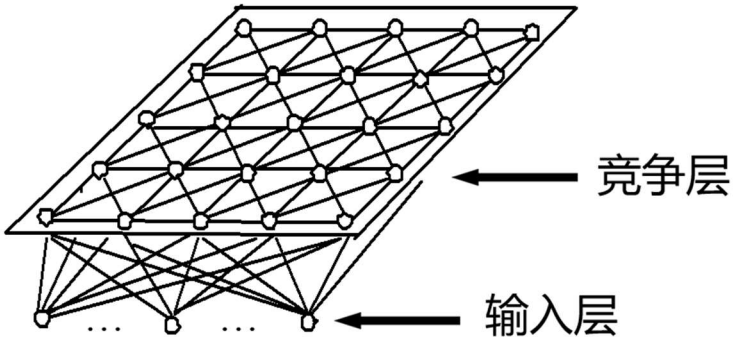

图 3：竞争学习过程

因为 *SOM* 训练采用的是竞争性学习的方式，所以当训练样本提供给网络的时候，就会计算它与每个权重之间的欧氏距离。将权重向量与输入最相似的神经元称为最佳匹配单元（*BMU*）。通过改变 *SOM* 栅格中 *BMU* 的权重，使与其邻近的神经元向着输入向量调整。从而 *BMU* 的量会随着时间和距离而降低。因此拥有权值 $W_v(s)$ 的神经元 v 的更新公式为[8]

$$W_v(s+1)=W_v(s)+\Theta(u,v,s)\alpha(s)\left(D(t)-W_v(s)\right)$$

其中

s 为步长指数

t 为训练样本的指数

$D(t)$ 为输入向量

u 为 $D(t)$ 的 *BMU* 指数

$\alpha(s)$ 为一个单调递减的学习系数

$\Theta(u,v,s)$ 为在步长为 s 下给出神经元 u 和神经元 v 之间距离的邻近函数

t 可以系统地从 $(0,1,2,\cdots,T-1)$ 中多次重复选取（T 为训练样本的大小）。也可以随机的从数据集中取出（*Bootstrap* 抽样），或采用其他一些抽样方法（如 *Jackknifing*），在此选择系统地训练所有样本。

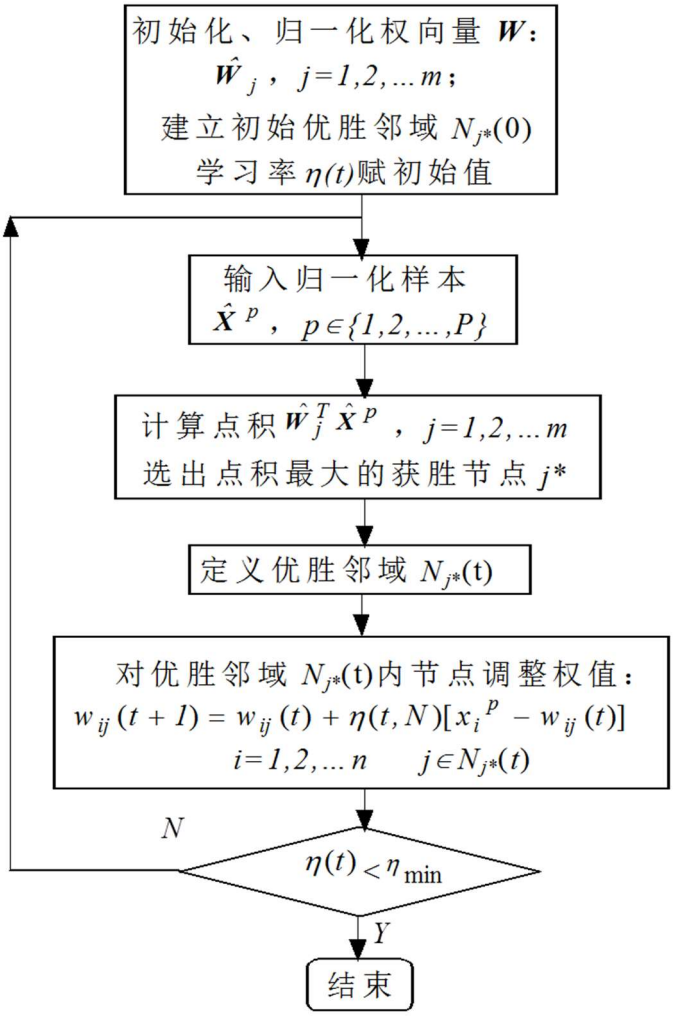

图 5:训练数据流程图

（4）模型的实现

通过 *MATLAB* 自带的神经网络工具箱代码进行编程，对交易量和响应时间两个数据进行训练。训练过程如下

1、标准化。采用一般的标准差法进行标准化。

2、距离函数的确定。由于考虑到输入向量为二维向量，且数据量较大，所以采用欧式距离。

3、神经元数量的选择（竞争层的大小）。神经元的数量会极大的影响到分类的效果。过多的神经元会导致分类结果过细，需要人工进行二次分类，甚至会出现“死节点”，即在训练过程中，某个节点从未获胜过且远离其他获胜节点，因此它们的权值从未得到过更新；而过少的神经元将会导致分类失败。经过多次尝试、筛选后发现 10*10 的竞争层网络效果最好，能顺利明确的将神经元分类。

4、选取初始权值的基本原则是尽量使权值的初始位置与输入样本的分布区域充分重合，避免出现大量的初始“死节点”，在此利用从训练集中随机抽取 $m=100$ 个输入样本作为初始权值。

5、学习率的选择。待分类的数据量较大这一明显特点，学习率过大会影响精度，容易产生过拟合；而学习率小则可能导致算法效率低下，花费资源过长。湖北工业大学的刘幺和等[9]指出，采用一种学习率随迭代次数下降的动态

学习率可以达到很好的优化效果。在训练开始时，学习率可以选取较大的值，之后以较快的速度下降，这样有利于快速捕捉到输入向量的大致结构；之后学习率在较小的值上缓降至 0 值，这样精细地调整权值使之符合输入空间的样本分布结构。这种动态学习率的具体函数为

$$\eta = Ae^{-\lambda n}$$

其中

A 为常数。在本模型中 $A=3$，$\lambda=0.0005$

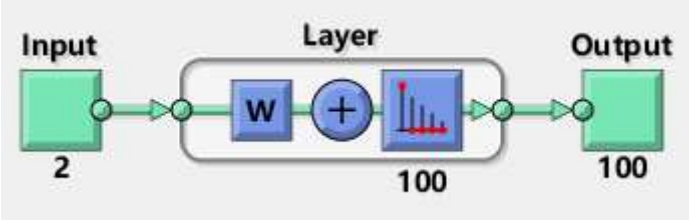

图 6：SOM 训练过程拓扑图

MATLAB 训练结果如下

经过三次训练后，*SOM* 将 131040 个样本数据分类到了 100 个神经元之中，神经元之间的距离反映了其离群的程度。

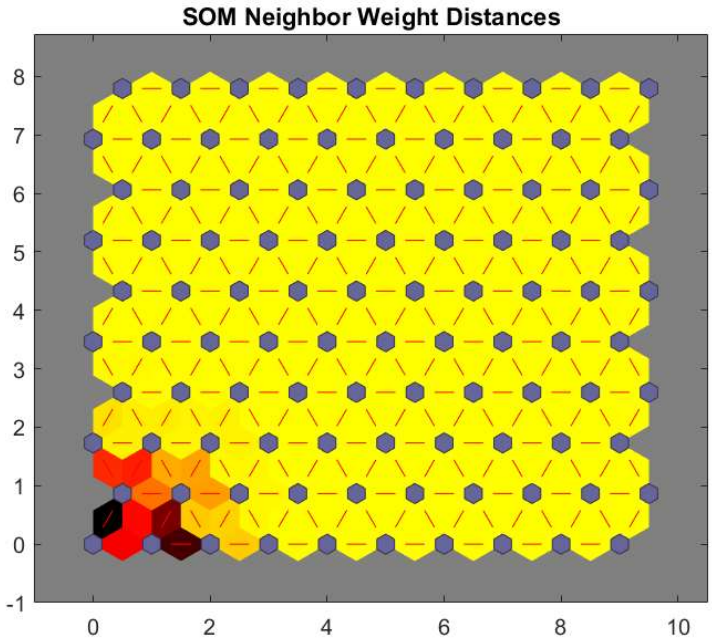

图 7：神经元距离图

图 7 中小六边形为神经元，红线代表神经元之间的距离，线的颜色越深表示相邻的神经元之间距离越远，由图 7 可知第 1、2、11、12 号神经元有明显的离群特征。考察其数据可知 11、12 号神经元对应第二类异常，共 12 组；1、2 号神经元对应第四类异常，共 19 组。部分结果如下

表 2：不同神经元对应的异常类型

日期	时间	成功率	响应时间	神经元
0209	217	0.76	6,735	12
0209	219	0.87	7,691	12
0209	220	0.82	6,684	12
0209	227	0.81	7,168	12
0209	228	0.72	6,311	12

0209	229	0.82	6,552	12
0309	304	0.60	16,896	11
0323	48	0.18	46,256	1
0323	49	0.31	39,376	2
0323	50	0.00	50,624	1
0323	51	0.06	53,543	1
0323	52	0.14	49,018	1
0323	53	0.13	49,593	1
0323	54	0.00	57,211	1
0323	55	0.06	53,889	1
0323	56	0.17	47,397	1
0323	57	0.10	51,697	1
0323	58	0.00	56,758	1
0323	59	0.19	45,991	1
0323	100	0.26	44,476	1
0323	101	0.79	28,820	2
…	…	…	…	…

由表 2 可见分类效果良好，因此得到历史数据中的故障情况如下

同时经过对结果的分析，可知第二类错误其异常值可能存在不连续分布。猜测可能是因为第二类异常是由于某一分行侧网络节点故障所致，与系统后端的异常相比，分行发生交易请求不连续（特别在交易量本来就低的凌晨时段），所以出现了成功率较低的异常点断续分布的情况。

分行侧网络出现故障

1 月 30 日 06:36～06:38

数据中心后端处理系统应用进程异常

2 月 09 日 02:17～02:29

3 月 23 日 00:47～01:04

4 月 16 日 06:00～06:03

5.2.3 后端 CPU 载荷模型的建立

（1）模型分析

由题目所给出的条件和信息可知，一笔交易沿着 ATM 前端经过分行侧网络节点到后端服务器，经过处理后再原路返回。由前面的假设，在系统其它功能正常的情况下，单笔交易的响应时间只受中心服务器处理速度的影响，即每一笔交易响应时间的差异是由中心处理器处理速度不同引起的。在不考虑侧网络节点的故障的情况下，当后端中心服务器的负载过大的时候，会表现出单笔交易响应时间较长这一特征；CPU 载荷情况还会受到同一时刻并行处理事务数量的影响，在 CPU 载荷大的时刻，交易量会表现出相对增加这一特征。因此认为只有同时满足单笔交易响应时间长、单位时刻交易量大这两个特征的异常点，才能被判断为 CPU 过载。

引入 CPU 载荷指数 L 作为评估后端中心服务器交易处理情况的特征参数。为了避免某一时刻由于单笔交易响应时间和交易量中其中一个值过大而另一值正常，造成对 CPU 过载的误判，所以在定义 CPU 载荷指数时引入对数计算，来保证 L 值大时，交易量和响应时间均较大。同时为响应时间增加一个 1.4 倍的指数因

子使得 L 值对响应时间更加敏感。综上，定义 CPU 载荷指数 L 为

$$L=\log(N)\times\log(\tau)^{1.4}$$

其中

N 为每分钟的交易量

τ 为每笔交易的平均响应时间

与交易量一样，CPU 载荷指数 L 也是一个随时间呈现周期性变化的量。在此将第三类的异常定义为 CPU 载荷指数明显大于正常值，并且认为当 ATM 系统出现异常情况时，之后一段时间内均会连续出现异常；而当只有单个点异常时，不能判定为 ATM 系统出现异常。

（2）阈值的确定

对 CPU 载荷指数的数据进行探索，发现其正常值数量巨大，离群点较少，具有和重尾分布相似的分布特征。相比于正态分布，其部分统计参数容易受到离群点的影响（如标准差），因此可应用方差分析法来寻找其置信区间。方差分析法作为一种数理统计方法广泛应用于气象、水文、地震等行业数据的科学统计与分析，常被用来计算最新采集数据与均值的离散程度。

正常的 CPU 载荷的数据分布集中在一定范围内，而当数据变化的绝对值超过 N 倍标准差，则说明数据存在异常。在利用方差分析数据异常时，N 的取值通常可采用两种方法来实现

（1）单次 N 值比较法。通常情况下 N 值默认为 3，即数据变化超过 3 倍标准方差即认为该点数据不正常。当这种不正常的数据点个数超过用户设定的某个数值时即认为数据存在异常，其中 N 的取值和不正常数据点个数可由用户根据被测项类型与长期统计结果具体设定，通过该方法可检测出数据超出 3 倍均方差的数据异常。

（2）双次 N 值比较法。利用第 1 次 N 值比较去除干扰，即认为数据变化超过 N 次标准方差的数据点为干扰点，去掉干扰点后进行第 2 次 N 值比较，通常取 N 为 2，即去掉干扰后，数据变化超过 2 倍方差的数据个数超过用户设定的某个数值时即认为数据存在异常。

在分析 CPU 载荷的数据时，发现 CPU 载荷存在极大偏离值，因此选用双次 N 值比较法更为合理。根据双次 N 值比较法的要求，首先进行第一次 N 值比较处理，在去除极端异常值后求出 CPU 载荷的平均值和方差作为其特征参数。

经过两次 N 值比较处理后的结果如下

表 3：两次 N 值比较处理后的结果

	第一次 N 值比较	第二次 N 值比较
μ	49515	49160
σ	61050	39193
异常点数量	28	117

由第三类异常的定义，找出偶然出现的单个时间点异常后，将这类点归为正常点。最终找出来了62组异常的数据，并根据这些数据的连续性将其划分为三个出现连续异常的时间段

3月01日 22:04～22:13

4月05日 07:17～07:23

4月14日 07:17～07:23

5.3 问题二模型的建立与求解

5.3.1 基于 *ARIMA* 的交易量时序模型

（1）模型分析与预处理

使用 *MATLAB* 绘制部分天数日交易量随时间变化的分布图。通过对图像的观察，发现不论是工作日还是非工作日，交易量随时间的变化趋势及发生变化的时间段大致相同，满足时间序列的特征，因此使用时序模型对交易量进行分析。

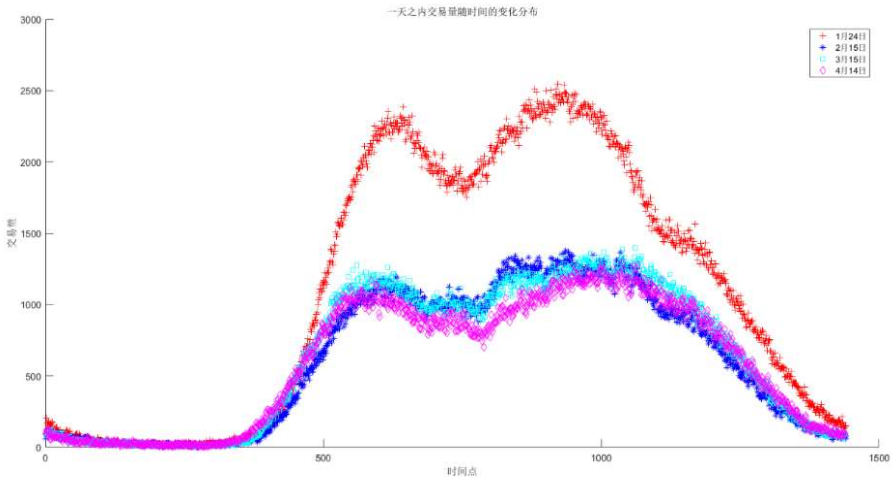

图 9：日交易量随时间的变化分布

由图 9 可看出，虽然大体上交易量随的时间变化具有直观的趋势性，但就单个数据而言，模型的噪声较大，不利于后续分析，也不能对某一分钟作出精确预测，因此首先需要对数据进行降噪处理。

由模型一中筛选出的异常点可知，一个异常所持续的时间一般为三至五分钟。为了去除噪声同时尽可能不丢失原有数据特征，选取三分钟作为时间窗口，以三分钟内交易量的平均值作为其对应交易量，结果如下

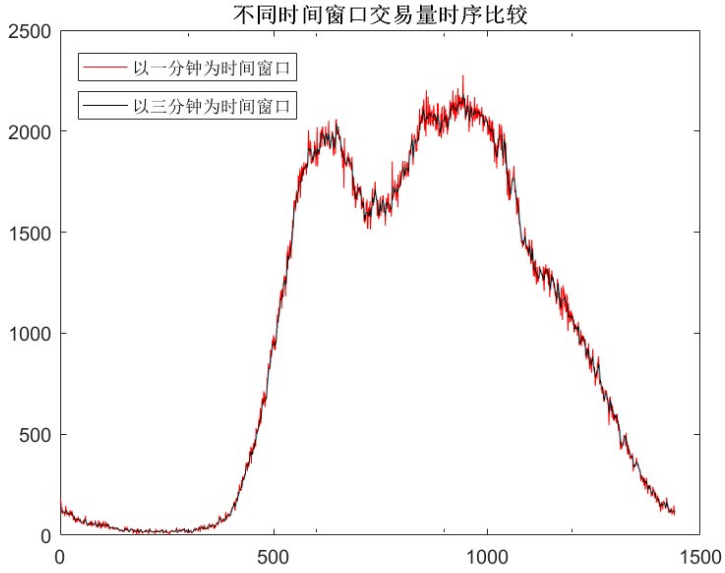

图 10：不同时间窗口的交易量时序分布比较

由图 10 可验证取三分钟作为时间窗口时，可较好的降低交易量变化过程中的产生噪声，使得后续建模过程更加精确。

同时，由于时序模型的主要目的是依据历史的健康数据，对未来交易量的正常值进行预测，为异常的判断提供实时依据。因此在后续的建模过程中已去除模型一中筛选出来的异常数据，并采取相邻数据取平均的方法补齐，之后利用相应的健康数据进行拟合。

（2）模型简介

时间序列法是一种定量预测方法，在统计学中作为一种常用的预测手段被广泛应用。时间序列分析则是根据系统观测得到的时间序列数据，通过拟合和参数估计来建立数学模型的理论和方法。对于平稳时间序列，可使用通用 *ARMA* 模型（自回归滑动平均模型）及其特殊情况的自回归模型、滑动平均模型或组合 *ARMA* 模型等来进行拟合。当观测值多于 50 个时一般都采用 *ARMA* 模型。对于非平稳时间序列则要先将观测到的时间序列进行差分运算，化为平稳时间序列，再用适当模型去拟合这个差分序列。

（3）平稳性检验（*ADF* 检验）

通过对交易量随时间变化趋势的分析，发现当差分次数 $d=0$ 时，数据具有明显的周期性，不满足 *ARMA* 模型对数据平稳性的要求，因此判断交易量的时间序列为非平稳时间序列。

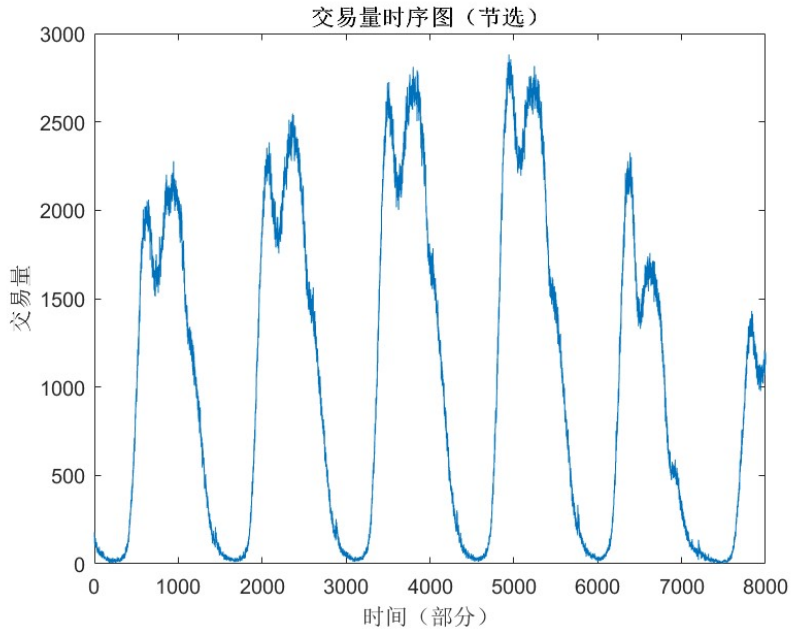

图 11：差分次数 $d=0$ 的交易量时序图

此时使用 *ARMA* 模型的推广模型 *ARIMA* 模型来对交易量进行分析。

ARIMA 模型即差分自回归滑动平均模型，是时间序列预测分析方法之一。不同于 *ARMA* 模型的只能用于检验平稳时间序列，*ARIMA* 模型在对原始时间序列进行差分后，将非平稳的时间序列组合成平稳的序列进而实现对非平稳时间序列的分析。在 $ARIMA(p,d,q)$ 中，*AR* 为“自回归”，p 为自回归项数；*MA* 为“滑动平均”，q 为滑动平均项数，d 为使之成为平稳序列所做的差分次数（阶数）。其数学表达式为：

$$\left(1-\sum_{i=1}^{p}\phi_i L^i\right)(1-L)^d X_t=\left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_i$$

其中

L 是滞后算子

$d\in Z, d>0$

因为平稳性检验其实也是探求 *ARIMA* 模型中 d 值的过程，所以首先进行一阶差分，即令 $d=1$，绘制出此时的交易量时序图。

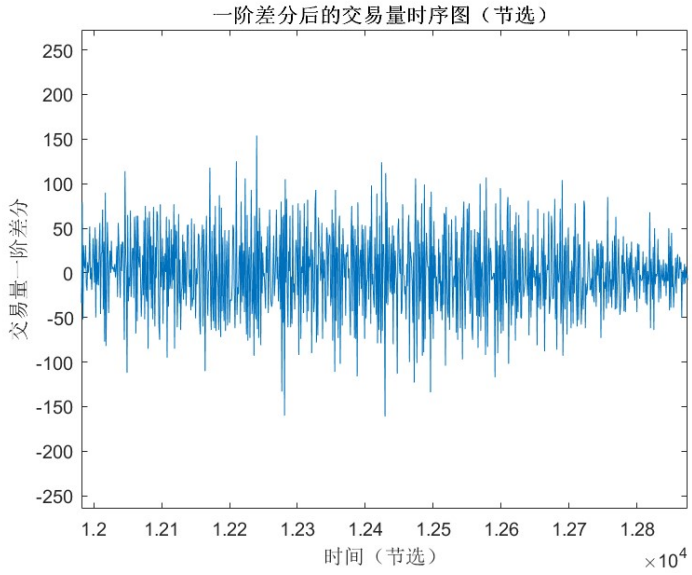

图 12：差分次数 $d=1$ 的交易量时序图

观察图 12 发现此时没有明显的趋势性，为了更加严格地检验一阶差分后数据的平稳性，对差分后的数据进行 *ADF* 检验。*ADF* 检验是平稳性检验中最为常用的一种方法，它由美国的两位统计学家 *D.A.Dickey* 和 *W.A. Fuller* 于 20 世纪 70 年代提出，用以验证其自相关性系数是否等于 1[10]。

在采用 *MATLAB* 对交易量时序的一阶差分进行 *ADF* 检验后，得到序列的 p 值小于 10^{-5}，证明时间序列是平稳的。同时计算序列的自相关系数和偏相关系数，得到的结果如下：

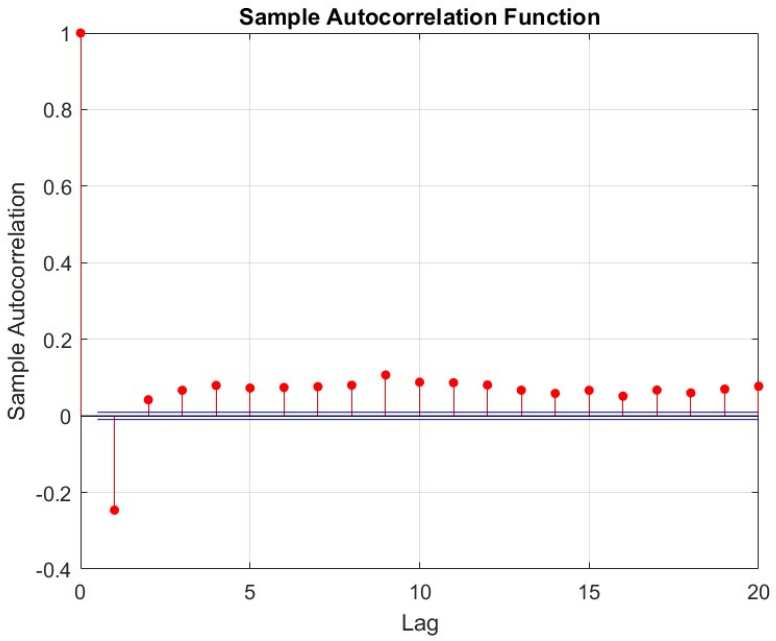

图 13：差分后序列的自相关性图

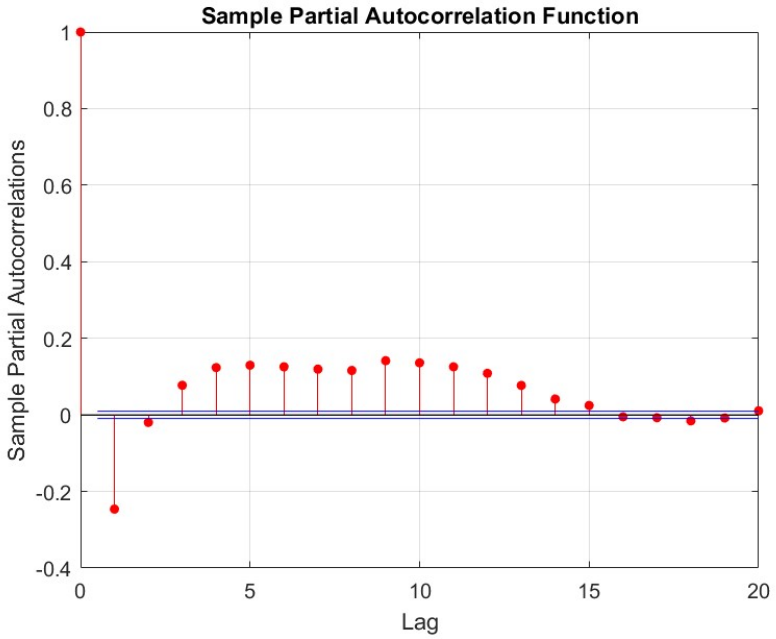

图 14：差分后序列的偏相关性图

由图 13 和图 14, 序列的自相关系数具有拖尾性；偏相关系数在 lag=15 时截断，在 lag=3 的时候也进入了置信区间。因此考虑采用 $ARIMA(15,1,0)$ 或者 $ARIMA(3,1,0)$ 模型。

（4）*AIC* 准则定阶

在验证了使用 *ARIMA* 模型的合理性后，需要对模型进行定阶，这里利用 *AIC* 准则。*AIC* 准则又称 *Akaike* 信息准则，是由日本统计学家 *Akaike* 于 1974 年提出的。*AIC* 准则起源于 *Kullback—Leibler* 信息量, 是信息论与统计学的重要研究成果，具有重要的意义[11]。

根据 *AIC* 准则，若设 x 为随机变量，则 x 的概率密度为 $f(x)$（其中含有 k 个未知参数），此时有

$$f(x)=g(x|\beta^0)$$

其中

$\beta^0=(\beta_1^0,\beta_2^0,\cdots,\beta_k^0)^T$ 为未知参数向量

$f(x)$ 属于分布族 $g(x|\beta)$

$K-L$ 是一种用来刻画 $f(x)$ 与 $g(x|\beta)$ 的接近程度的信息量，通过求出 $K-L$ 的最小值寻得最接近于 $f(x)$ 的参数概率密度 $g(x|\beta)$，其定义为

$$I(f(\cdot),g(\cdot|\beta))=\int f(x)\ln\frac{f(x)}{g(x|\beta)}dx$$

当给定容量为 n 的样本时，设样本观测值 $x=(x_1,x_2,\cdots,x_n)$，模型参数 $\beta=(\beta_1,\beta_2,\cdots,\beta_k)^T$，$\ln(L(\beta))$ 为其对数似然函数，则 AIC 信息准则满足

$$AIC(k)=-2\ln\left(L\left({}_k\hat{\beta}^{(m)}\right)\right)+2k=\min$$

其中

${}_k\hat{\beta}^{(m)}$ 是模型参数 $\beta=(\beta_1,\beta_2,\cdots,\beta_k)^T$ 的最大似然估计

k 为未知参数的个数，$k=p+q+1$

计算得到 $ARIMA(15,1,0)$ 模型的 AIC 值为 408031.5；$ARIMA(3,1,0)$ 模型的 AIC 值为 414603.9。选取 $ARIMA(15,1,0)$ 模型更加合理。

（5）系数求解

采用最小二乘法估计参数。设 X_t 是 $ARMA(p,q)$ 序列，X_t 的一个观测样本为 $X_1,X_2,\cdots,X_n$。现取 $k=t-1$，即由 $\{X_{t-1},X_{t-2},\cdots,X_1\}$ 的线性组合来预报 X_t，X_t 的估计量为

$$\sum_{j=1}^{t-1}\varphi_{t-1,j}X_{t-j},\quad j=2,3,\cdots,n$$

其系数满足如下方程

$$\begin{bmatrix}1 & \rho_1 & \cdots & \rho_{t-2}\\ \rho_1 & 1 & \cdots & \rho_{t-3}\\ \vdots & \vdots & & \vdots\\ \rho_{t-2} & \rho_{t-3} & \cdots & 1\end{bmatrix}\begin{bmatrix}\varphi_{t-1,1}\\ \varphi_{t-1,2}\\ \vdots\\ \varphi_{t-1,t-1}\end{bmatrix}=\begin{bmatrix}\rho_1\\ \rho_2\\ \vdots\\ \rho_{t-1}\end{bmatrix},\quad t=2,3,\cdots,n$$

此时预报的残差平方和为

$$S(\varphi,\theta)=\sum_{j=2}^{n}\left(X_t-\sum_{j=1}^{t-1}\varphi_{t-1,j}X_{t-j}\right)^2$$

在 X_t 的平稳可逆域中寻求 $\hat{\varphi}_L,\hat{\theta}_L$，使得 $S(\hat{\varphi}_L,\hat{\theta}_L)=\min$，则满足上式的 $\hat{\varphi}_L,\hat{\theta}_L$ 称为 φ,θ 的（无条件）最小二乘估计（ULS 估计）。

利用 $MATLAB$ 进行上述过程，可得到序列 $ARIMA(15,1,0)$。

（6）$ARIMA$ 模型的检验

为了进一步检验 $ARIMA(15,1,0)$ 模型是否符合统计规律，利用 χ^2 检验判断数据经过时间序列处理后产生的是否为白噪声。

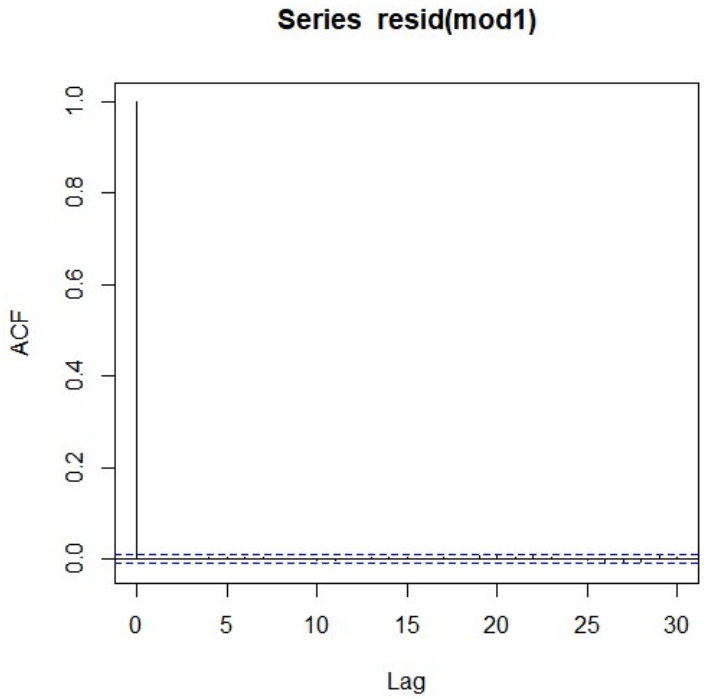

图 15：经过时序处理后的数据拟合

可见模型拟合效果良好，通过模型后的数据为白噪声。下面严格按照理论判断

若拟合模型的残差记为 $\hat{\varepsilon}_t$，它是 ε_t 的估计，记

$$\eta_k = \frac{\sum_{t=1}^{n-k} \hat{\varepsilon}_t \hat{\varepsilon}_{t+k}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}, k = 1,2,\cdots,m$$

$Ljung-Box$ 的 χ^2 检验统计量是

$$\chi^2 = n(n+2)\sum_{k=1}^{m} \frac{\eta_k^2}{n-k}$$

检验的假设是

$$H_0: \rho_k = 0, \text{当} k \le m \ ; H_1: \text{对某些} k \le m, \ \rho_k \ne 0$$

在 H_0 成立时，若 n 充分大，χ^2 近似于 $\chi^2(m-r)$ 分布，其中 r 是估计的模型参数个数。

在进行 χ^2 检验时，给定显著水平 α，查表得上 α 分位数为 $\chi_\alpha^2(m-r)$。则当 $\chi^2 > \chi_\alpha^2(m-r)$ 时，拒绝 H_0，即认为 ε_t 非白噪声，模型检验未通过；而当 $\chi^2 \le \chi_\alpha^2(m-r)$ 时，接受 H_0，即认为 ε_t 是白噪声，模型检验通过。

计算可得模型的 $Ljung-Box$ 统计量 $\chi^2 = 5.439 \le \chi_{0.95}^2(15)$，在 95%的置信度下通过检验，因此模型的拟合效果良好[12]。

（7）基于 *ARIMA* 模型的预报

已知时间序列为

$$\hat{X}_k(m)=\varphi_1\hat{X}_k(m-1)+\varphi_2\hat{X}_k(m-2)+\cdots+\varphi_p\hat{X}_k(m-p),\quad m>p$$

在已知前 p 个数据的情况下，时间序列的预报公式为

$$\hat{X}_{k+1}^{(q)}=\begin{bmatrix}-G_1 & 1 & 0 & \cdots & 0\\ -G_2 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ -G_{q-1} & 0 & 0 & \cdots & 1\\ -G_q+\varphi_q^* & \varphi_{q-1}^* & \varphi_{q-2}^* & \cdots & \varphi_1^*\end{bmatrix}\hat{X}_k^{(q)}+\begin{bmatrix}G_1\\ G_2\\ \vdots\\ G_{q-1}\\ G_q\end{bmatrix}X_{k+1}+\begin{bmatrix}0\\ 0\\ \vdots\\ 0\\ \sum_{j=q+1}^{p}\varphi_j^*X_{k+q+1-j}\end{bmatrix}$$

其中

式中常数项在 $p\le q$ 时为 0

$$\varphi_j^*=\begin{cases}\varphi_j, j=1,2,\cdots,p\\ 0, j>p\end{cases}$$

$\hat{X}_k(q)=\left(\hat{X}_k(1),\hat{X}_k(2),\cdots,\hat{X}_k(q)\right)^T$ 为预报向量

由此得到各个时刻交易量的估计值。部分结果如下

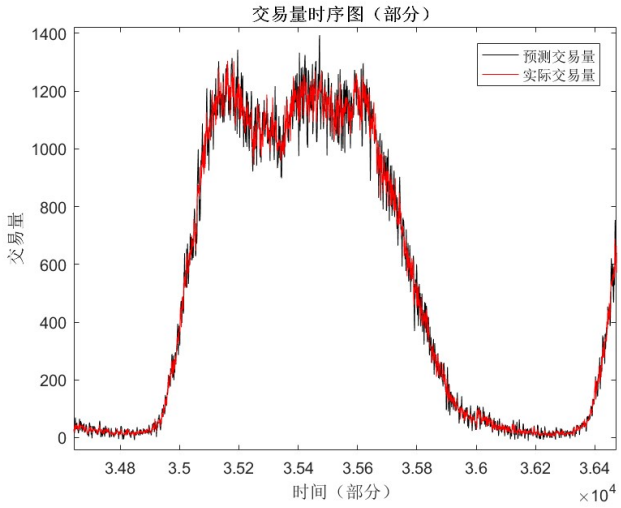

图 16：部分交易量估计值时序图

两曲线基本重合，可见预报的精度良好。

（8）系统状态的监测

根据已建立的 $ARIMA(15,1,0)$ 模型，利用历史数据对未来数据的正常值做出预测，为数据异常的监控提供依据。

对于第一类异常，交易量在某个时间区间内会出现明显的下降，此时预测的健康数据大于实际数据。两者之间的差值在一定程度上反映了数据的下降情况，但由于数据基数对第一类异常也有影响，因此定义实时的数据下降指标为

$$U=\frac{\hat{x}_t-x_t}{x_t}\times 100\%$$

其中

$\hat{x}_t$ 为时间序列预测的三分钟内的平均交易量

x_t 为三分钟内的平均交易量实际值

上式给出了交易量下降的量化指标，在系统出现故障时，这个值会明显偏大，因此利用第一问筛选出来的数据对 U 进行检测

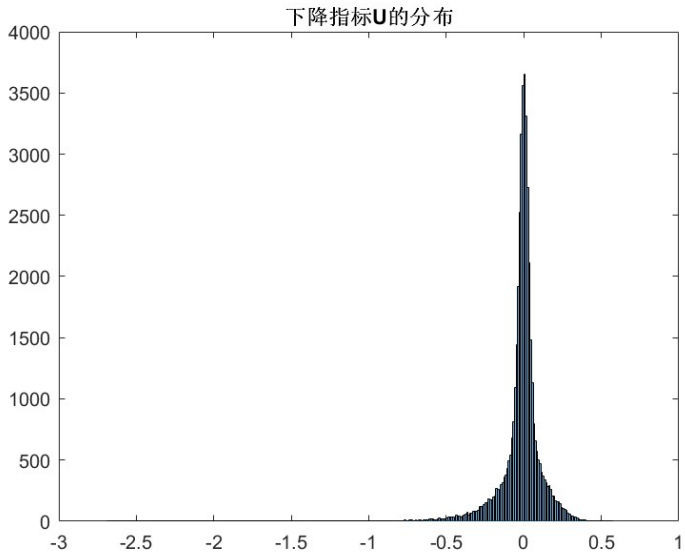

图 17：下降指标 U 的分布

由图 17 可知 U 是一个峰度较大的近似正态分布，大部分值集中在半径为 0.5 的 0 的邻域中。计算可知有 58 个值小于-1 的数据点，和第一类异常的数据点基本重合。有 99%的数据大于-0.5，定义 U 处于[-1,-0.5]时，系统为亚健康状态。第一类异常的报警、预警规则为

$$\begin{cases} U > -0.5 & \text{正常} \\ -1 < U < -0.5 & \text{预警} \\ U < -1 & \text{报警} \end{cases}$$

5.3.2 基于贝叶斯定理的异常预警及监测模型

（1）模型分析

相比于第一类异常，第二、三、四类异常的发生更加迅速、随机性更加大，而且一般在发生之前没有预兆，这样给预测带来了困难。但是二、三、四类异常其参数的特征均十分明显

表 3：各类异常的参数特征

类别	交易成功率指标	响应时间指标	交易量指标	CPU 载荷指标
正常	高	正常	正常	正常
分行侧网络出现故障	高	正常	正常	正常
数据中心后端处理系统异常	高	缓慢	偏高	偏高
数据中心后端处理系统进程异常	低	缓慢	正常	偏高

考察这些数据的相关性，计算其相关性矩阵

表 4：参数间的相关性

	CPU 载荷	交易量	响应时间	成功率
CPU 载荷	1.00	0.43	0.02	-0.11

交易量	0.43	1.00	-0.03	-0.08
相应时间	0.02	-0.03	1.00	-0.37
成功率	-0.11	-0.08	-0.37	1.00

可见数据之间的相关性都很低，数据之间比较独立，说明这些指标能够高效地对系统的状态信息进行表征。结合上述特征综合考虑，引入贝叶斯分类器模型。

（2）朴素贝叶斯分类器介绍

朴素贝叶斯分类器是一系列在假设各特征之间强（朴素）独立的情况下，以贝叶斯定理为基础的简单概率分类器。该分类器模型会从自有限集合中取出用特征值表示的类标签，之后分配给问题实例。

其中，朴素贝叶斯算法不是训练这种分类器的单一算法，而是一系列基于相同原理的算法：所有朴素贝叶斯分类器都假定样本每个特征与其他特征不相正常关。尽管这些特征相互依赖或者有些特征由其他特征决定，然而朴素贝叶斯分类器认为这些属性在判定类型时概率分布上是独立的[13]。

理论上，概率模型分类器是一个条件概率模型

$$p(C|F_1,....,F_n)$$

其中

独立的类别变量C有若干类别

条件依赖于若干特征变量$F_1,\ldots,F_N$

（3）模型分析

然而，当特征数量n较大或者每个特征能取大量值时，基于概率模型列出概率表变得不现实。于是修改模型为下式

$$p(C|F_1,....,F_n)=\frac{p(C)p(F_1,...,F_n|C)}{p(F_1,...,F_n)}$$

在实际应用中，只需要关心分式中的分子部分，因为分母并不依赖于C，而特征值F_i又是给定的，所以可以认为是分母一个常数。此时分子等价于联合分布模型

$$p(C,\ F_1,....,F_n)$$

使用链式法则

$$\begin{aligned}
&p(C,\ F_1,....,F_n)\\
&\propto p(C)p(F_1,....,F_n|C)\\
&\propto p(C)p(F_1|C)p(F_1,....,F_n|C,F_1)\\
&\propto p(C)p(F_1|C)p(F_2|C)p(F_1,....,F_n|C,F_1,F_2)\\
&.....
\end{aligned}$$

根据贝叶斯分类器原理中的条件独立假设，每一个F_i对其他的特征F_j，$j\neq i$时是条件独立的。此时有

$$p(F_i|C,F_j)=p(F_i|C)$$

联合分布模型表示为

$$p(C,\ F_1,....,F_n)=\frac{1}{Z}p(C)\prod_{i=1}^{n}p(F_i|C)$$

其中

Z 为证据因子，是一个只依赖于 $F_1,\ldots,F_N$ 的缩放因子，当特征和变量的值已知时为常数。

由于分解成所谓的类先验概率 $p(C)$ 和独立概率分布 $p(F_i|C)$，上述概率模型的可掌控性得到很大的提高。

从得出的先验概率中通过比较以及时序分析得到很多有用信息，通过构造特征参量以及求解，直接对系统的状态以及将来可能出现的状态做出预警或者报警。

（4）模型建立

在前面所述模型的基础上，从给出的数据中建模，并提取出来了发生异常的时候的数据点。利用这些数据点，结合构建的特征参数（CPU 载荷），构造朴素贝叶斯模型。则每一时刻系统的状态一共有以下四种情况：

表 5：系统的交易状态及其对应代号

正常	分行侧参数数据变更或者配置错误	数据中心后端处理系统异常	数据中心后端处理系统应用进程异常
R	E_1	E_2	E_3

求出每一种交易状态的先验概率：

$$p(S)=\frac{p(X\mid S)p(S)}{p(X)}=\frac{\prod_{k=1}^{n}p(x_k|S)p(S)}{\prod_{k=1}^{n}p(x_k)}$$

其中

S 为交易状态，$S\in\{R,E_1,E_2,E_3\}$

X_I为特征参数，$X_I\in\{$交易量，响应时间，成功率，CPU载荷$\}$

（5）模型求解

首先采用标准差法将数据标准化：

$$x'_{ij}=\frac{x_{ij}-\bar{x}_j}{s_j}$$

其中

$$\begin{cases}\bar{x}_j=\dfrac{1}{n}\sum_{i=1}^{n}x_{ij}\\ s_j=\sqrt{\dfrac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}\end{cases}$$

标准化之后的数据均值为零，标准差为 1，无量纲。

其次求出各个状态的独立概率 $P(s)$

$$P(S_i)=\frac{n_i}{N},i=1,2,3,4,5,6$$

结果如下

表 6：各交易状态的独立概率

交易状态 S_i	R	E_2	E_3	E_4
$P(S_i)$（百分数）	99.6478	0.2130	0.1103	0.0289

之后求解条件概率 $P(x_j|S_i)$，由于这些量的分布都是连续的，所以其概率密度为一连续函数，因此只需要求出其概率密度函数即可。由于样本数据足够大，所以可用样本的频率来求解原始随机变量的分布，此处由于各个数据的分布不符合现有已知的标准分布，在概率论中常采用核密度估计来拟合其密度函数[14]，属于非参数检验方法之一。其求出的概率密度仍然符合下列规则：

利用概率密度函数的两条性质

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$P(0 < x \le a) = \int_0^a f(x)dx$$

其中

$P(0 < x \le a) = F(0 < x \le a)$，即概率用频率近似

用 *MATLAB* 求解可得到各个参量的条件分布。对于所有点求出其对应各种状态的概率后，认为最大概率所对应的状态即是当前状态。

（6）系统全局指标的构建

1、健康度指标

利用贝叶斯定理以及现有数据（先验数据）可得出现在系统处于各个状态的先验概率 $p(S)$，$S \in \{R, E_1, E_2, E_3\}$。但单一的先验概率不能准确表征系统的健康程度，例如当其他概率相同时，对 $\{p(R)=80\%, p(E_1)=20\%\}$ 和 $\{p(R)=80\%, p(E_1)=80\%\}$，后者系统的健康度明显低于前者。因此构造表征系统健康度的特征参数，定义一分钟内系统健康度为

$$s = \frac{p(R)}{\sum_{\text{对}S\text{求和}} p(S_i)}, S \in \{R, E_1, E_2, E_3\}$$

2、健康度下降指标

在实际的系统中，系统的各个参数随时间动态的、连续的变化，由于数据采集精度的限制，只能以最小一分钟为单位，将数据进行离散的采集。前面构造贝叶斯模型的时候，利用离散的数据对其分布进行拟合以达到将概率连续化的目的。这里，利用前面已经使用过的动态时间窗口的方法，计算特征参量在一段时间内的变化，以监测系统的连续变化趋势。

在实践中，一个正常的系统，其系统正常的先验概率 $p(R)$ 应该随时间平稳，当系统进入异常状态时，其 $p(R)$ 会经历一段连续下降的过程，利用一段时间窗口内的 $p(R)$ 进行曲线拟合所得的斜率作为衡量，如果斜率随时间不断增大，那系统即将进入异常状态的可能性就很大。

根据前面的时间窗口的设置，计算每一分钟的系统健康度下降趋势，选取统计时间窗口 W=6，则定义一分钟内的健康度趋势为

$$\eta_i=\frac{\sum_{l=i-W}^{i}(s_l-\overline{s})(l-\overline{l})}{\sum_{l=i-W}^{i}(l-\overline{l})^2}$$

其中

$$\overline{s}=\frac{1}{W+1}\sum_{l=i-5}^{i}s_i$$

$$\overline{l}=\frac{1}{W+1}\sum_{l=i-5}^{i}l_i$$

η_i 为第 i 分钟的健康度下降趋势

s_l 为第 l 分钟的健康度

若健康度趋势为不断下降的，则认为有很大几率系统即将发生异常，计算 W 时间窗口内递减的分钟数为：

$$D(i)=\begin{cases}0,\eta_i>\eta_{i-1}\\1,\eta_i\le\eta_{i-1}\end{cases}$$

定义这 W 时间窗口内的健康度下降指标为

$$T=\frac{\sum_{n=i-W}^{i}D(n)}{W+1}$$

在后面的应用中，认为当健康度下降指标 $T>0.2$ 时，认为系统有很大倾向在向着异常发展，可以根据最大的异常先验概率对系统进行预警，在整体异常前提醒工作人员进行排查。

基于上述指标，结合前面的先验概率的计算，此时模型不仅能够对系统的异常进行报警，还可以对系统即将发生的异常进行预警，以保证 ATM 系统工作的正常。

综上，基于贝叶斯分类器的 ATM 系统预警及报警的模型流程图如下

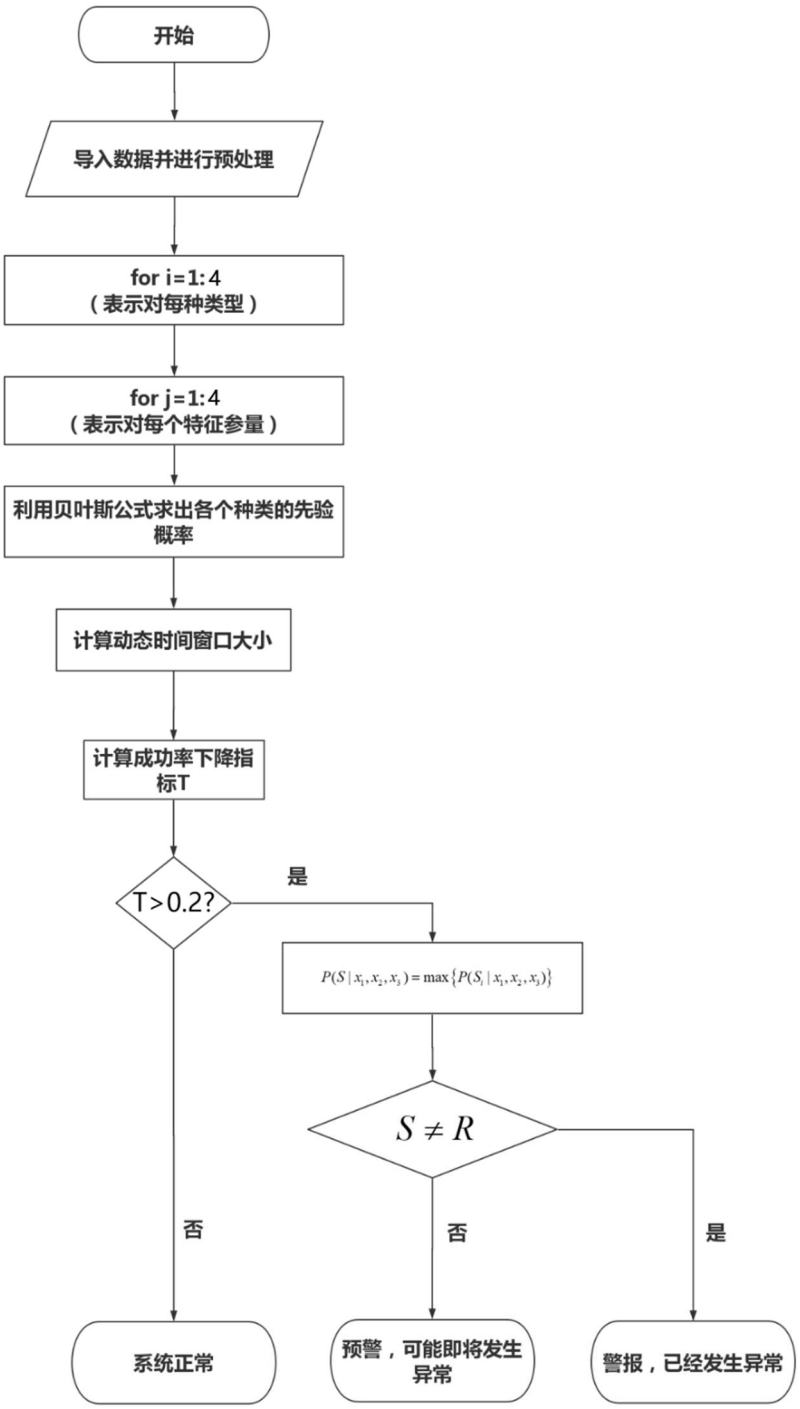

图 18：基于贝叶斯分类器的 ATM 系统预警及报警的模型流程图

（7）模型检验

随机从各个类别的交易状态内抽取测试组数据 15 组，利用上述的贝叶斯分类器进行分类，测试数据是否能准确报错。

表 7：各类状态实际报错次数与贝叶斯分类后预测报错次数

预测值 实际值	R	E_1	E_2	E_3
R	15	0	0	0
E_1	1	14	1	0

E_2	0	0	11	4
E_3	0	0	3	11

由表 7 知，实际检验过程中分类器的效果良好。

由问题一可知，在 3 月 23 日 00:49～01:06 出现了一次较大的系统异常，为不失一般性，从 00:45 开始对系统进行监测直到 01:08 分结束。结果如下：

表 8：分类器对 3 月 23 日 00：45~01：08 监测情况

时间	$p(R)$	$p(E_1)$	$p(E_2)$	$p(E_3)$	T
0:45	0.52	0.00	0.00	0.00	0.00
0:46	0.36	0.00	0.00	0.00	0.00
0:47	0.50	0.00	0.00	0.00	0.00
0:48	0.04	0.00	0.00	0.08	0.17
0:49	0.00	0.00	0.00	44.19	0.33
0:50	0.00	0.00	0.00	126.01	0.50
0:51	0.00	0.00	0.00	151.56	0.50
0:52	0.00	0.00	0.00	92.04	0.50
0:53	0.00	0.00	0.00	200.54	0.50
0:54	0.00	0.00	0.00	190.32	0.33
0:55	0.00	0.00	0.00	183.17	0.33
0:56	0.00	0.00	0.00	182.70	0.33
0:57	0.00	0.00	0.00	200.44	0.33
0:58	0.00	0.00	0.00	142.77	0.33
0:59	0.00	0.00	0.00	129.86	0.50
1:00	0.00	0.00	0.00	119.97	0.67
1:01	0.00	0.00	0.00	194.81	0.67
1:02	0.00	0.00	0.00	195.27	0.50
1:03	0.00	0.00	0.00	283.42	0.50
1:04	0.00	0.00	0.00	149.19	0.50
1:05	0.00	0.15	0.00	7.86	0.33
1:06	0.01	0.00	0.00	0.00	0.17
1:07	0.12	0.00	0.00	0.00	0.00
1:08	1.66	0.00	0.00	0.00	0.17

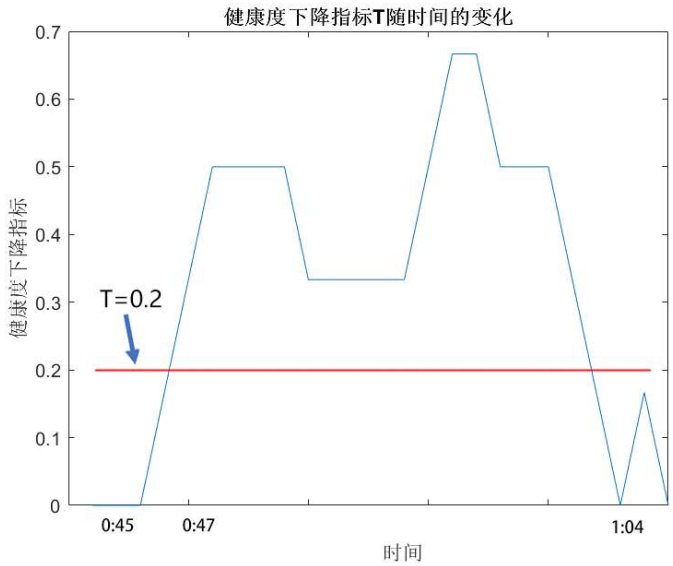

图 19：健康度下降指标 T 随时间的变化

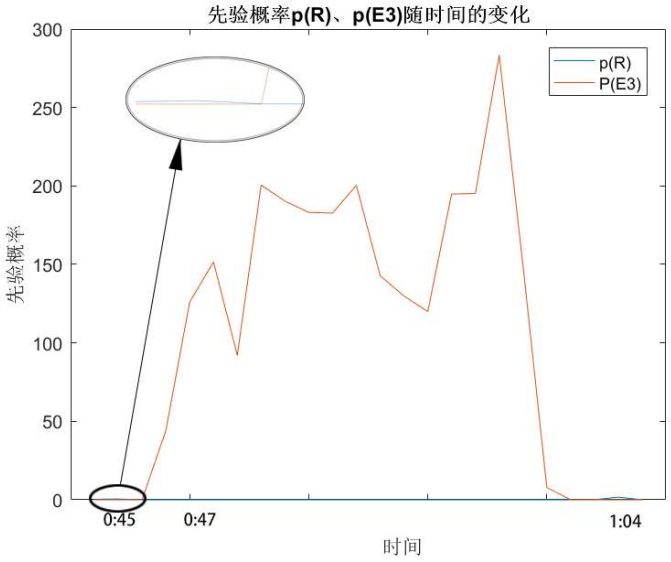

图 20：先验概率 $p(R)$ 和 $p(E)$ 随时间的变化

由第一题结论可知，此时系统应该是发生了第三种交易状态异常，结合数据和图 20 看出，开始发生异常时，$p(R)$ 也就是系统正常的先验概率迅速下降，而第三种异常的先验概率迅速上升。在异常开始的的第二分钟 $(00:48)$，$p(R)<p(E_3)$，可见变化十分的灵敏。同时，受到数据变化的影响，异常的中间时段，$p(R)$ 保持在较低的水平，而其他的几种异常的先验概率也有不同程度的上升，但是仍小于 $p(E_3)$。

在开始异常的第一分钟 $(00:47)$，系统迅速给出了预警，在第二分钟 $(00:48)$，系统的预警变为警报，且正确地指出了第三种异常。可见模型反应迅速，效果直观明显。

5.3.3. 异常数据时间窗口的确定

在离线的异常检测过程中基于已有的分类模型可以对数据点逐一分类，以达

到目的。但是这样做往往会忽略前后数据之间的关联，某些系统一段连续的时间之内的异常可能会被判定为很多次异常。设计系统时我们采用了双次计数的方法，将比较连续的异常值进行时间窗口的划定，从而可以直接生成报告，知道异常数据的时间跨度。

5.4 问题三模型的建立与求解

5.4.1 基于更多的交易参数

（1）交易的种类

每一笔种类交易可能会涉及到不同工作量，比如跨行转账和查询余额。而前者需要调动的信息及操作更多，相应的会更加消耗时间和资源。

根据这些参数，丰富 ATM 系统的拓扑图，能够为故障的发生以及成因提供更多分析的依据。

如基于交易种类的数据，可以更加精确地对前面所述的后台 CPU 载荷进行计算：

$$L'=\gamma * f(N)\times g(\tau)$$

其中

N 为每分钟的交易量

τ 为每笔交易的平均响应时间

γ为修正系数，针对每一笔交易，根据其工作量（交易类型）不同

（2）其他分行的数据

如果能得到其他分行的数据，可以减少由于其他原因产生的虚报后端异常。基于前端后端系统相对独立的假设，若如果多个分行的数据同时异常，则有很大几率是后端处理系统发生了故障；若只有一个分行的数据发生故障，则更有可能是前端或者分行测网络有故障。

（3）用户的操作细节

虽然在分析问题前，假设 ATM 系统数据与持卡人的操作水平无关，但事实上每个人对银行系统的熟悉程度、对密码的记忆力、输入速度都有区别。如果能提供操作者的操作信息，在建模过程中去除这些因素的干扰，系统的精度将会得到提升。

5.4.2 基于更多的数据总量

（1）更加精确的时序模型

如果能够拥有数量更多的数据，则可以建立一个更加精准的时间序列模型。从现有的数据来讲，ATM 交易的笔数不仅仅只是工作日和非工作日有区别，一周之内的各个工作日中，ATM 交易笔数也有稳定的差别。星期六和星期天的分布也有差别，甚至这些分布随着月份的变化也会相应的变化，例如一月份和二月份比

其 ATM 交易量日平均值更高。随着数据的增加，可以将各个时间序列做进一步的细分，从而可以提高模型的精度，减少误报。

同时，ATM 交易量还受一些固定节日的影响。从现有的数据来看，一月份由于经历了春节，在除夕之前的日子可能受工资结算或者年底资金回笼的影响，交易量与其他日子相比，显著增加，春节之后，ATM 交易量下降。相类似的，在其他的节日，人们的消费或者资金容易受到节日的影响。在获取更多数据后，可以对各个节日建立不同的序列模型，对可能的异常进行预测，从而应对节日里交易量突增的情况。

基于时间跨度更大的数据，我们可以对数据进行季节性分析，建立梳系数 $ARIMA$ 模型，使得结果更加准确。

（2）表现更好的贝叶斯模型

贝叶斯分类器模型中，训练数据的分布直接决定了先验概率的计算结果。如果训练数据过少，在对一些没有出现过的异常进行分类时可能产生严重的错误。基于更大的数据库，能够丰富贝叶斯分类器的训练集，从而提高监测系统应对各种异常的灵敏度与准确率。

六、灵敏度分析

系统在实际的应用过程中，数据将会更加复杂多变，能否保证在遇到异常时迅速、准确的预警对整个 ATM 系统的安全十分重要，接下来主要针对预警系统进行灵敏度分析。

（1）时间窗口的改变

在定义健康度下降指标中，将健康度下降指标定义为前六分钟内健康度下降的比例，这样得到的健康度下降指标剔除了随机产生的误差，如果窗口值过小，会影响报警的效果。为此分别设定 2-8 分钟的时间窗口对系统进行测试，测试数据为 3 月 23 日 00:00～01:40（包含一段时间的第四类异常），结果如下：

表 9：不同窗口时长的报警情况（3 月 23 日 00：00~01：40）

窗口时长	2	3	4	5	6	7	8
报警次数	15.00	19.00	22	25	23	25	28
准确率	遗漏报警	100.00%	90.91%	80.00%	86.96%	80.00%	71.43%
预警提前时间	2.00	2.00	2	2	1	1	1
报警结束滞后时间	3.00	4.00	6	6	6	7	8

可见时间窗口越长，报警结束滞后也会越长，相反预警提前时间会减少，实际应用过程中显然前一个指标更加重要，同样还发现窗口越长，报警的准确率会降低，但是当时间窗口小于 3，将会有异常点被遗漏，这是不允许的。

综合上述分析可知，当系统选择 $W=3$ 的时间窗口时，效果最好。模型对 W 灵敏度较高，当改变其他因素时，应重新计算选择 W。

（2）下降指标临界值的改变

在预警系统中另外一个重要的参数是 T 报错的阈值，在前述模型中我们采用

了 0.2 作为阈值。通过改变这个阈值，对系统进行测试，可得结果如下

表 10：不同阈值的报警预警次数情况

T	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
报警预警次数	19	19	19	11	11	11	3	3	3

由表 10，当T变高时，系统将遗漏较多报警，验证了 0.2 作为阈值的合理性，模型对T灵敏度不高。

七、模型的评价与推广

7.1 模型的评价

（1）模型优势

1、对每个特征参数根据其特点采取不同的模型进行分析，使得模型对参数信息的捕捉更加准确。

2、与其他常用的异常预警报警机制相比，朴素贝叶斯分类器在处理大量数据时更加迅速、高效。

3、模型中的预警报警机制除了可以区分出前、后端异常，还能够具体判断出是题目中所给四种异常中的哪一种。

（2）模型缺点

1、朴素贝叶斯分类模型对于离群点过于敏感，无法排除掉偶然产生的异常点。

7.2 模型的推广

此模型可以应用于监测互联网信息传输问题。例如通过每秒钟传输的字节数、传输的错误率、传输节点数和终端数，提取特征参数来判断网络传输是否处于异常状态，并对产生的故障分类报警，对可能的异常提前预警。帮助运行商及用户及时发现问；同理，通过分析调制速率、能量利用率、噪声率等，可以检测数据信号的编码和传输是否正常。

同时模型也可以应用于研究经济方面的股票的波浪变化问题以及一些期货公司所遇到的财务困境问题，通过分析经济指标提取出相应的特征参数，对相关历史数据进行深入挖掘后，对可能发生的经济问题进行预报。

参考文献

[1] Denning D E. An intrusion-detection model[J]. IEEE Transactions on software engineering, 1987 (2): 222-232.

[2] Gruau F. Neural Network Synthesis using Cellular Encoding and the Genetic Algorithm[J]. 1994.

[3] Grossberg S. Nonlinear neural networks: Principles, mechanisms, and architectures[J]. Neural networks, 1988, 1(1): 17-61.

[4] Teng H S, Chen K, Lu S C. Adaptive real-time anomaly detection using inductively generated sequential patterns[C]//Research in Security and Privacy, 1990. Proceedings., 1990 IEEE Computer Society Symposium on. IEEE, 1990: 278-284.

[5] 张良均. MATLAB 数据分析与挖掘实战[M]. 机械工业出版社, 2015.

[6] 张学功. 金融时间序列中加性异常值的鉴别与校正[J]. 价值工程, 2009, 28(2): 4-7.

[7] [8] Kohonen T. Self-organized formation of topologically correct feature maps[J]. Biological cybernetics, 1982, 43(1): 59-69.

[8] Kohonen T, Honkela T. Kohonen network[J]. Scholarpedia, 2007, 2(1): 1568.

[9] 刘幺和, 陈睿, 彭伟, 等. 一种 BP 神经网络学习率的优化设计[J]. 湖北工业大学学报, 2007, 22(3): 1-3.

[10] Dickey D A, Fuller W A. Distribution of the estimators for autoregressive time series with a unit root[J]. Journal of the American statistical association, 1979, 74(366a): 427-431.

[11] Akaike H. A new look at the statistical model identification[J]. IEEE transactions on automatic control, 1974, 19(6): 716-723.

[12] Box G E P, Jenkins G M, Reinsel G C, et al. Time series analysis: forecasting and control[M]. John Wiley & Sons, 2015.

[13] 吴立增, 朱永利, 苑津莎. 基于贝叶斯网络分类器的变压器综合故障诊断方法[J]. 电工技术学报, 2005, 20(4): 45-51.

[14] 李存华, 孙志挥, 陈耿, 等. 核密度估计及其在聚类算法构造中的应用[J]. 计算机研究与发展, 2004, 41(10): 1712-1719.