华北电力大学

毕 业 设 计(论文)

题　　目　　基于数据挖掘的 KQI,KPI 关联

院　　系　　电子与通信工程系

专业班级　　通信 1702 班

学生姓名　　熊梓豪

指导教师　　张珂

二〇二一年六月

基于数据挖掘的KQI,KPI关联

摘要

为实现通信设备质量与用户感知的关联，即KPI与KQI之间的映射，以作为运营商更准确、针对性地优化服务设备的参考基准。本文基于统计和机器学习算法，构建监控、评价通信运行质量和服务状态的综合模型，对KPI、KQI数据进行挖掘，找到样本分布规律和指标间关联规则程度，实现了本课题预期的研究目标。本文主要工作如下：

1) 针对单维异常检测问题，本文基于统计方法，首先利用指数、Gamma、对数正态等经典概率分布，通过极大似然法来拟合KPI、KQI直方图包络，然后基于卡方检验、偏度峰度检测评估拟合效果判决是否接受，最后，根据分布函数、二分法、箱型图等方法确认阈值。同时对单极、二极、特殊数据，也针对性设计了合适的检测方法。

2) 针对关联规则挖掘问题，本文基于Apriori算法，计算出KPI、KQI指标内部和两者之间的关联规则和程度，讨论了导致结果的实际原因。

3) 针对多维异常检测问题，本文利用主成分分析、K-Means聚类等多种机器学习方法挖掘KPI、KQI分布特征，通过将多指标向量特征降维等方法映射到单一维度，并配合单维异常检测模型，实现多指标阈值判定的目的。

4) 针对时间序列预测问题，本文基于LSTM算法，以KPI上行吞吐量为对象，训练LSTM网络收敛到理想状态，然后，我们对比该模型预测结果和原始时间序列曲线的异同，分析评价了预测模型的性能。

5) 最后，本文将上述各数据挖掘模型整理、统一，构建一套实时监控、关联评价KPI、KQI的综合模型。

关键词：KPI；KQI；数据挖掘；异常检测；Apriori关联算法

KQI, KPI ASSOCIATION BASED ON DATA MINING

Abstract

In order to realize the relationship between the quality of communication equipment and user service satisfaction, mean the mapping between KPI and KQI, which can be used as a reference for ISP to optimize service equipment more pertinently. Based on statistics and ML algorithm, this thesis constructs a comprehensive model to mines KPI and KQI data, finds out the distribution rule of samples and the degree of association rules between indicators, and achieves the expected research goal of this topic. The main work of this thesis is as follows:

1) Aiming at the problem of 1D anomaly detection, based on the statistical method, this thesis first uses the classical probability distributions such as exponential, gamma and lognormal to fit the histogram envelope of KPI and KQI by the maximum likelihood estimation. Then evaluate the fitting effect based on Chi square test and D'Agostino's K-squared test to decide whether to accept it. Finally, distribution function, dichotomy and box diagram are used to calculation the threshold. Meanwhile, for unipolar, bipolar and special data, the appropriate method is also designed.

2) Aiming at the problem of mining association rules, this thesis designs a model based on Apriori, calculates and analyzes the degree of association between KPI and KQI.

3) Aiming at the problem of multidimensional anomaly detection, this thesis uses PCA, K-Means clustering and other ML methods to mine KPI and KQI samples. By mapping multi-index vector features to 1D and cooperating with the single dimension anomaly detection model, the purpose of multi-index threshold determination is realized.

4) Aiming at the problem of time series prediction, based on LSTM algorithm, taking KPI uplink throughput as the object, we train LSTM to converge to the ideal state. Then, we compare the prediction results of the model with the original time serie, and evaluate the prediction model.

5) Finally, this thesis unifies the above data mining models to build a comprehensive model of real-time monitoring, association evaluation KPI and KQI.

Keywords: KPI; KQI; Data Mining; Anomaly Detection; Apriori

目　　录

摘要.....I
Abstract.....II
1 绪论.....1
1.1 课题背景.....1
1.2 国内外研究现状.....1
1.2.1 KPI、KQI 数据挖掘研究概况.....1
1.2.2 阈值检测研究现状.....1
1.3 本文主要研究内容.....3
1.4 论文结构.....3
2 单维数据异常检测.....4
2.1 相关工作.....4
2.2 概率分布拟合方法.....4
2.2.1 指数分布、极大似然估计.....7
2.2.2 卡方分布拟合检验.....8
2.2.3 Gamma 分布.....10
2.2.4 对数正态分布.....13
2.2.5 偏度-峰度检测.....14
2.3 异常检测方法.....16
2.3.1 分布函数阈值确定.....16
2.3.2 箱型图.....17
2.3.3 特殊类型数据.....19
2.4 单维异常检测综合模型.....20
2.5 本章小结.....21
3 Apriori 关联规则.....22
3.1 Apriori 关联规则算法.....22
3.1.1 基本概念.....22
3.1.2 算法流程.....22
3.2 数据挖掘结果.....24
3.3 本章小结.....27
4 多维向量异常检测.....28
4.1 相关工作.....28
4.2 算法模型及数据挖掘结果.....28
4.2.1 主成分分析+ Mahalanobis 距离.....28
4.2.2 聚类方法.....29

4.3 模型总结 32
5. 数据预测 33
5.1 相关工作 33
5.2 LSTM 长短期记忆模型 33
5.3 数据挖掘结果 35
5.3.1 原始样本时间序列 35
5.3.2 预测时间序列 36
5.4 本章小结 36
6 KPI、KQI 综合关联评价模型 37
6.1 综合模型流程 37
6.2 本章小结 38
7. 总结与展望 39
7.1 总结 39
7.2 不足与展望 39
参考文献 40
附录 42
致谢 43

1 绪论

1.1 课题背景

随着 5G 技术在中国的全面建设，如何合理准确地对 5G 通信服务质量和用户对服务的满意程度进行评价，以作为进一步升级和改造通信系统设备的参考基准，已经成为一个重要的关键性问题。因此，关键性能指标 KPI，和关键业务指标 KQI，便是电信管理论坛 TMF 和各大通信供应商一起共同提出的针对通信服务质量的评价体系和标准。

KPI 关键性能指标，指的是面向网络设备的评价方法，其主要对六个领域进行重点关注：(1)接入能力、(2)保持能力、(3)容量、(4)移动能力、(5)利用率和(6)完整性。而如何实现对这六个领域的评价，是通过对该领域下具体的设备性能参数的测量来完成的，这些细分的子指标包括，RRC 连接建立/重建/失败次数、小区用户面上/下行丢包率、同/异频切换成功率、上/下行流量、平均/最大用户数等。其次，关键业务 KQI 指标，则是面向用户感知的业务层面的评价方法，对于不同的服务类型诸如视频类、网页类、即时通讯类和游戏类等，其进行评价的对象也各有差异，以视频类业务为例，评价指标有播放成功率、下载速率(kbps)、平均等待时长(ms)、播放中断率、用户数等。

1.2 国内外研究现状

1.2.1 KPI、KQI 数据挖掘研究概况

赵刚[1]以彩铃系统为背景，开发了 K2K 算法来实现 KPI 对 KQI 指标的映射和关联，他通过层次式树形分析算法，设置指标评价对象的因素论域和评价等级集，来计算每个 KPI 指标的模糊综合评价数值，并基于评价数值和构建的指标阈值门限集合，通过线性分段法计算得到 KQI 得分集合，最后通过 Bolta 算法向上逐层回溯，更新每个树节点的得分评价数值。

倪萍[2]则基于流数据挖掘算法 EARA，通过改进电信管理论坛发布的 KPI 层次结构，让不同量纲的评价对象能够相对平滑地聚合到同一个目标指标，并利用层次分析算法模型，对各层次的相对权重一致性检验来实现对权重的计算。同时，倪萍在此工作基础上进一步开发了可视化的压缩模型 VPC，通过压缩事件来帮助管理人员更容易的从大规模的关联规则集中发现有特征有价值的信息模式，帮助使用者更容易理解数据挖掘结果。

1.2.2 阈值检测研究现状

主流的无监督异常检测模型有以下几种：

(1) 概率统计方法

Markus 和 Andreas 在 2012 年[3]，提出一种基于直方图的孤立点检测模型 HBOS，来实现针对网络安全无监督模式下入侵数据检测的快速算法，该模型的优势在于其时间复杂度较低且运行效率远高于多元方法，其中固定组距模式时间复杂度为 $O(n)$而动态组距模

式则为 $O(n \cdot \log(n))$，作者测试 HBOS 模型比聚类方法快 2 倍也比紧邻方法快 7 倍，适合运用在大规模数据集上。模型首先将单维特征数据绘制直方图、排序并归一化处理，且对分布未知的样本数据采用动态宽度模式，然后 HBOS 模型对每个样例计算其箱格的相对高度并作为判决异常数据的参考基准。最后，作者测试分析 HBOS 在全局检测下性能表现较优异但在局部异常检测模式下则存在欠缺。

(2) 主成分分析方法

Shyu M.L.，Chen S.C.，Sarinnapakorn K.和 Chang L.在 2003 年提出了基于主成分分析的异常检测算法模型[4]，论文中作者以计算机安全和入侵防御为研究背景，并期望在不提出任何分布假设的条件下，即无监督模式下，能够检测出隐藏在正常数据中的作为入侵攻击的异常点值。因此，作者设计了主成分分析器 PCC 模型，将 PCA 算法和离群点检测算法相结合，来检测孤立异常点。在该模型中，PCA 算法并非设计用来为异常检测问题服务的，其目标是在最大程度保留样本信息的条件下，降低样本特征数量也即维数，于是作者使用 Mahalanobis 距离作为离群点检测来架起 PCA 和异常检测之间的桥梁，其中，Mahalanobis 距离不同于欧氏距离，其公式中协方差矩阵的计算能够帮助消除异常值带来的协变性和相关性对 PCA 算法的不良影响，以提高 PCC 模型的稳健性，而稳健性的加入也使得 PCC 模型更适合处理无监督异常检测问题。同时，论文实验结果显示，作者设计的 PCC 模型查全率、准确率、误警率分布达到 98.94%、97.89%、0.92%，结果优于近邻法、LOF 算法、基于 Canberra 离群点检测等异常检测方法。其次，关健、刘大昕[5]也同样利用主成分分析作为模型基础，然后利用多层感知机和最小均方误差原则学习样本行为特征，最后模型通过输出均方误差值，来分析当前行为与正常情况的差异并作为判决参考，如果超出阈值则判为异常。

(3) 聚类方法

Portnoy，Eleazar 和 Stolfo[6]，在 2001 年针对网络入侵的无监督检测问题，提出用聚类方法作为异常检测的模型基础。在该模型中，作者使用层次聚类中的单连通 Single-Linkage 算法的一个简单异构变体作为其聚类模型方法，该算法在近似线性的时间复杂度下具有较好优势。然后，作者通过设置簇团宽度 Width 和样本点距离簇心的欧氏距离，来构建训练集并挖掘所有特征簇团，同时基于设置的最大簇团比例 N 和簇团成员的数量将其划分为正常簇和异常簇。其次，当簇团从训练集中被挖掘后该模型即可开始异常检测工作，通过计算输入样本点距离各个簇心的欧氏距离来找到距离其最近的簇团，如果最近簇团为异常簇则输入样本点判决为异常数据，从而实现异常检测的目的。最后，作者选取筛选后的 KDD-Cup1999 数据集作为其模型测试的样本对象和训练集合，以准确率和误检率为性能评价指标，分别给出了在不同簇团宽度 Width 和最大簇团比例 N 的参数设置下，该聚类模型的准确率和误检率的测试数值。同时，杨斌[7]在 2008 年基于 K-Means 聚类方法建立异常检测模型，通过选取聚类器 DB 值最小的结果并利用簇间欧氏距离的方法来挖掘正常簇和入侵攻击簇，来达到判决异常数据的目的。

1.3 本文主要研究内容

本次设计主要针对通信业务中，面向网络设备的关键性能指标 KPI 和面向用户感知的关键质量指标 KQI 之间的数据关联问题提出解决方案。研究内容如下：

1) 基于异常检测算法，实现对 KPI、KQI 指标数据中异常点指标的捕捉和检测。

2) 基于 Apriori 关联规则等数据挖掘算法，实现对 KPI 和以视频类业务为代表的 KQI 之间的数据关联分析。

3) 利用多种机器学习算法，对 KPI、KQI 进行分析和预测，测试并对比分析各算法之间的差异、性能。

4) 构建一套 KPI、KQI 关联分析系统，能够监控和评价通信业务的运行质量和服务状态。

5) 实现以 C++语言为基础，以 Apriori 关联规则等机器学习算法为代表的，通用式算法程序模板库。

1.4 论文结构

第 1 章，提纲挈领地介绍了论文的研究背景和当前该课题的研究概况，简明扼要地交代了研究的目的。

第 2 章，针对单维数据异常检测问题，介绍了基于统计方法设计的单维异常检测综合模型，同时递进式地描述了经典分布的参数估计、卡方拟合检验、峰值-偏度检测、箱型图等模型算法，最后给出了 KPI、KQI 共 29 个指标的阈值结果。

第 3 章，针对关联规则的数据挖掘问题，介绍了 Apriori 关联挖掘算法模型，计算出了 KPI、KQI 指标内部和两者之间的关联程度，并对结果进行了分析。

第 4 章，利用多种机器学习算法对 KPI、KQI 的数据挖掘，分别阐述了主成分分析+Mahalanobis 距离、K-Means 聚类两种模型，分析了对 KPI、KQI 数据处理的结果，实现了多指标异常检测问题，最后比较了优劣特点。

第 5 章，针对时间序列预测问题，介绍了长短期记忆模型，搭建了 LSTM 神经网络，实现了对 KPI、KQI 指标的时间序列预测。

第 6 章，综合上述设计的所有模型，构建一套 KPI、KQI 关联分析系统，能够实时地监控和评价通信业务的运行质量和服务状态。

2 单维数据异常检测

2.1 相关工作

本章节的目标，是处理单维数据异常检测问题。即给出一列数据，我们能够通过分析数据自身规律确定一个阈值，并将其作为检测出所有高于或低于该阈值的异常数据的基准。

针对该问题，本章节拟采用统计学方法进行建模。首先，本文从已有的 KPI、KQI 样本数据的基础统计特征开始分析，利用指数分布、Gamma 分布、对数正态分布等多种经典随机变量分布，基于统计学中分布拟合方法的极大似然估计算法，对上述分布的典型参数进行估值，实现给定概率分布函数对样本频数直方图的拟合。其次，利用统计学中的假设检验方法，如卡方拟合检验、峰度-偏度检验等，对 KPI、KQI 共 29 个样本指标在上述各类概率分布函数拟合的情况，进行了评价和判决，只有在满足前提置信度的条件下才对拟合予以接受。同时，本章节也设计了箱型图等单维异常检测模型，对单极化、多极化等特殊数据，也给出了针对性判决方法。

最后，本章节综合上述所有算法，总结并设计完成了单维异常检测综合模型，实现了对 KPI、KQI 所有 29 个指标的异常检测并计算给出了每个指标各自的判决门限。同时，本章节的综合模型具有普适性，能够适应并处理未来新的指标。综上所述，本章节顺利地实现了对单维数据异常检测的问题的处理，为下一章节的关联规则挖掘模型打好了基础。

2.2 概率分布拟合方法

首先，本文对已有的 KPI、KQI 样本数据，进行基础的统计分析，列出其各指标的均值、方差和标准差，如表 2-1、2-2 所示。从表中我们可以发现，诸如(2) RRC 连接建立完成次数、(4)E-RAB 建立成功总次数、(6)eNodeB 发起 E-RAB 释放次数、(16)下行 PUSCH PRB 利用率、(17) 下行 PRB 利用率、(18) 下行 PDSCH PRB 利用率 2015、(19)小区内的平均用户数等指标，其样本均值与标准差数据基本相同，即其指标的样本方差是样本均值的平方。通过查阅概率分布表，我们可以知道，以指数分布等为代表的经典概率分布可以满足这个规律。同时，本文绘制出 KPI、KQI 各指标的频数分布直方图，以进一步更清晰地研究指标样本的分布规律。如图 2-1、2-2 所示。

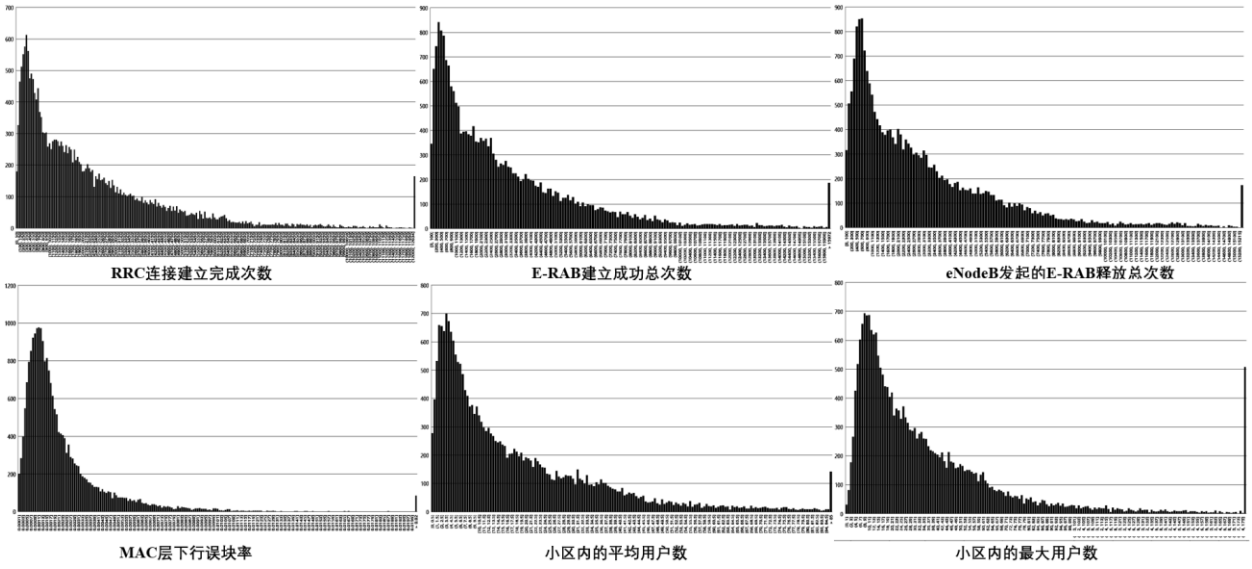
表 2-1　KPI 样本均值、方差、标准差基本参数一览表

	指标名称	均值	方差	标准差
1	RRC 连接建立失败次数	1.474	22.906	4.786
2	RRC 连接建立完成次数	2245.840	5426614	2329.510
3	RRC 重建成功次数	28.290	2243.544	47.366
4	E-RAB 建立成功总次数	3155.134	10708241	3272.345
5	E-RAB 异常释放次数	1.177	8.648	2.941
6	eNodeB 发起 E-RAB 释放次数	3058.053	9316441	3052.285

7	小区用户面上行丢包率	0.000036	0	0.000301
8	MAC 层上行误块率	0.0195	0.00218	0.0467
9	MAC 层下行误块率	0.00238	0.000009	0.003045
10	E-RAB 建立成功率(%)	99.912	4.084	2.021
11	同频切换成功率(%)	97.978	185.627	13.625
12	异频切换成功率(%)	96.725	307.267	17.529
13	ESRVCC 切换成功率	0.335	0.222	0.471
14	eNodeB 内切换成功率(%)	0.996	0.00365	0.0604
15	eNodeB 间切换成功率(%)	0.993	0.00434	0.0659
16	下行 PUSCH PRB 利用率	40.341	1630.858	40.384
17	下行 PRB 利用率(%)	17.107	279.586	16.721
18	下行 PDSCH PRB 利用率 2015	34.219	1118.381	33.442
19	小区内平均用户数	17.777	313.243	17.699
20	小区内最大用户数	38.362	1874.799	43.299
21	上行流量(GB)	0.077	0.00841	0.0917
22	下行流量(GB)	0.792	0.789	0.888
23	上行总吞吐量	66×10^{6}	620×10^{15}	789×10^{6}
24	下行总吞吐量	681×10^{6}	58196×10^{15}	7629×10^{6}

表 2-2　KQI 样本均值、方差、标准差基本参数一览表

	指标名称	均值	方差	标准差
1	视频播放成功率(%)	99.492	22.289	4.721
2	下载速率(kbps)	3688.626	26000189.057	5099.038
3	播放平均等待时长(ms)	1020.399	3032596.256	1741.435
4	播放中断率(%)	23.447	26976.518	164.245
5	视频用户数(个)	3.923	22.331	4.726



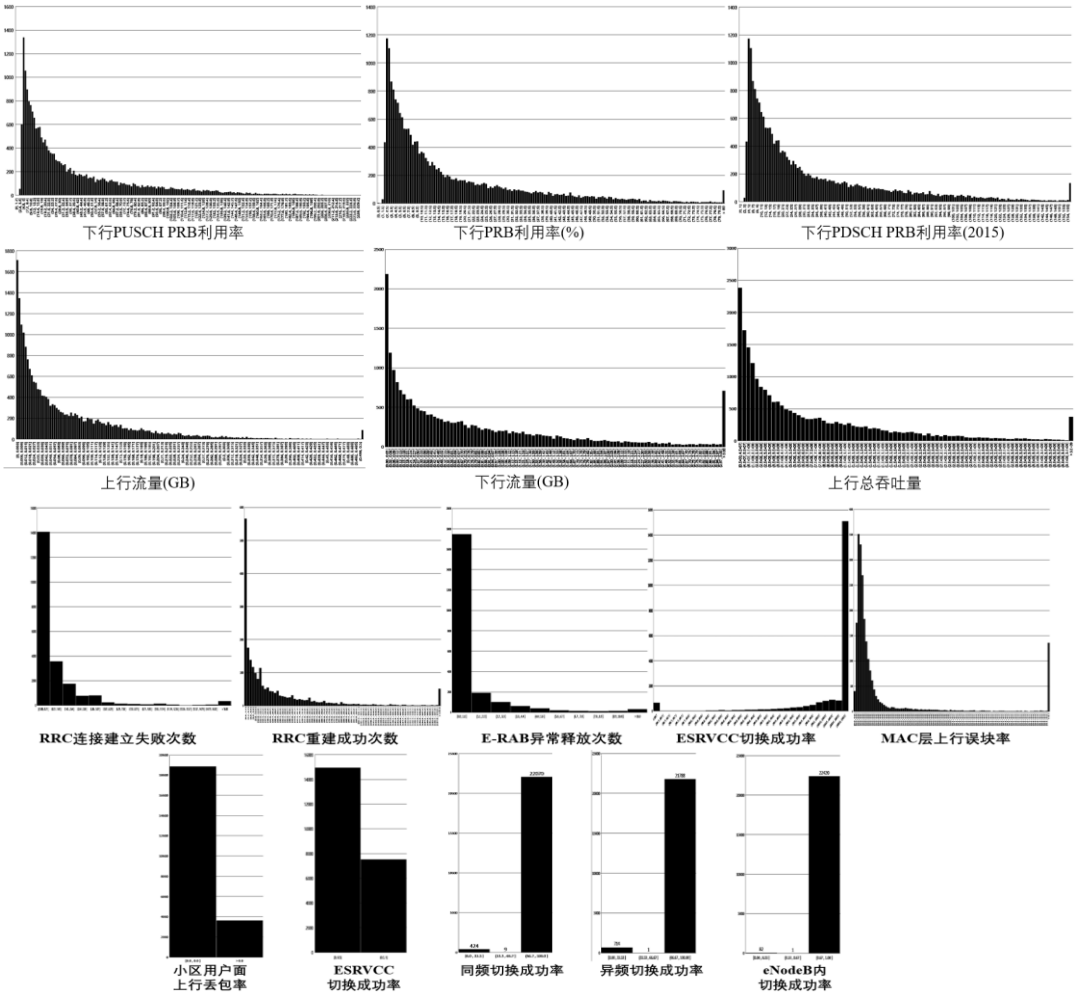


图 2-1 KPI 指标频数分布直方图

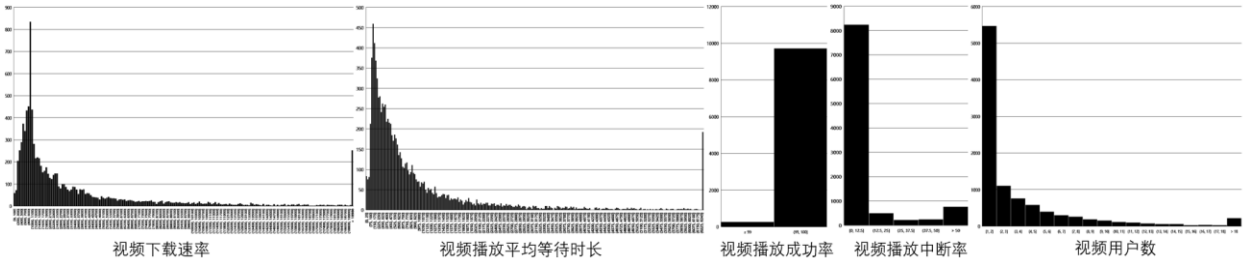


图 2-2 KQI 指标频数分布直方图

上述各指标的评述直方图显示，绝大多数指标样本的分布都相对均匀，峰值都靠近于零点且数据恒正然后沿着 X 轴正方向呈现逐渐下降的趋势直至零，同时频数直方图的包络曲线相对平滑，剧烈的数据起伏几乎不存在，说明 KPI、KQI 指标样本数据的规律性较强，更容易进行分析和解析。其次，部分诸如成功率的指标，分布则比较集中，频数分布的箱数较少，数据基本都聚集在 0%和 100%的极点附近。

2.2.1 指数分布、极大似然估计

指数分布是概率密度曲线服从负指数函数的一种连续性随机变量分布。其随机变量恒大于零且按 x 单调递减，指数分布的均值为θ，方差为θ^2，即均值等于标准差。指数分布的概率密度函数公式，如式 2-1 所示。

$$f(x)=\frac{1}{\theta}e^{-\frac{x}{\theta}}\quad(x>0) \tag{2-1}$$

基于本文已有的样本数据，为了计算并拟合假设的分布的参数θ，我们可以采用极大似然估计方法。极大似然估计的思想是在采样样本满足独立同分布的条件下，使得目标估计的参数值，能够极大化似然函数，即使得似然函数对未知参数的导数为零 (似然函数，指 x 已知而参数θ未知的条件概率$P(x|\theta)$)。如公式 2-2 所示。

$$L(\theta)=\prod_{\mathrm{i=n}}^{1}\frac{1}{\theta}e^{-\frac{x_i}{\theta}}=\frac{1}{\theta^n}e^{-\frac{\sum x_i}{\theta}} \tag{2-2}$$

$$\Rightarrow \ln L=-n\ln\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i \tag{2-3}$$

令似然函数极大化，有

$$\hat{\theta}=maxarg\,L(\theta) \tag{2-4}$$

$$\Rightarrow \frac{\partial L}{\partial\theta}|_{\theta=\hat{\theta}}=0 \tag{2-5}$$

解得

$$\hat{\theta}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x} \tag{2-6}$$

从上述结果我们可以得知，指数函数的样本参数估计值等于样本均值。于是，我们通过计算的均值推导每个指标数据的指数概率密度函数方程，并绘制出了 KPI、KQI 中可能满足指数分布的指标的拟合图像，如图 2-3 所示。

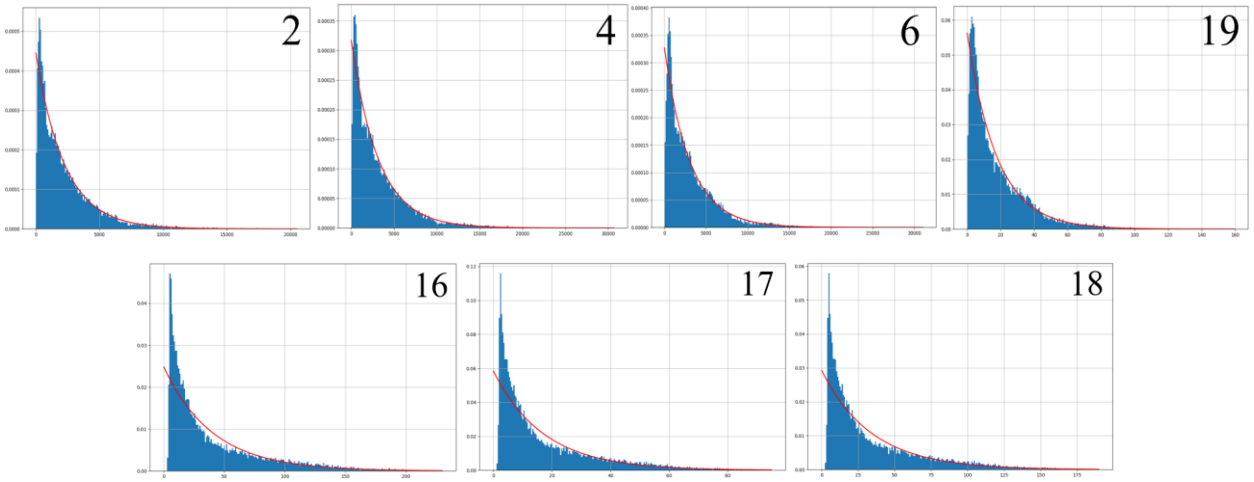


图 2-3 指数分布拟合图

从上述的随机变量分布拟合图像，我们可以看出，KPI 样本 2、4、6、19 四个指标的拟合程度良好，直方图的顶部即 X 趋近于零的一小段数据稍有突出，而 X 紧随其后的小部分数据由低于曲线形成稍浅的低谷，其中高出和低于的两部分频数分布的总面积相同。我们可以分析认为，在趋近于 0 的一小段数据中，样本分布的面积和估计的指数分布相同，但小部分数据在数据采集时被判断的更接近 0，致使峰值处出现一小段高出和低于的问题。与此同时，2、4、6、19 四个指标的中尾部的绝大部分样本和指数分布曲线保持一致、基本重合，证明我们估计的指数分布能够很好的拟合这四个指标的随机变量分布特征。而 KPI 样本 16、17、18 这三个指标和拟合的指数函数曲线相比较，峰值处过于突出，而样本分布呈现曲率程度要高于指数函数能拟合的曲率。而指标数据的尾部部分与估计的指数曲线贴合程度良好。

2.2.2 卡方分布拟合检验

为了检验和评价上述数据分布的拟合结果，本文采用卡方拟合检验方法。首先，我们做出假设，

H_0: 总体 X 满足指数分布，且参数值等于样本均值$\theta = \hat{\mu}$.

其次，我们以Δ为组距，将样本数据按频数分为互不相交，左开右闭的n组，并按实际情况设置上溢箱和下溢箱$\{A_1, A_2, \dots, A_n\}$。实际上，我们将样本分布按组距分为了 N+2 个箱子，每个区间的频数即是箱子里的小球，然后我们从随机的一个箱子里取出一个球，那么取出这个球的若是 K 号箱子的概率应当是$p_k = f_k/n_k$。同时，我们按照前提估计的分布，按相同的组距和组数，分为 N 的箱子，那么其对应的概率p_k应当近似同样本数据的箱子一致，而估计分布箱子中的概率则可由分布函数给出，即

$$p_i = F(i+1) - F(i) \tag{2-7}$$

对于上溢箱和下溢箱，我们同样有，

$$p_{overflow} = 1 - F(i_{overflow}) \tag{2-8}$$

$$p_{underflow} = F(i_{underflow}) \tag{2-9}$$

于是，我们可以将样本每个箱子中的概率f_i/n，同估计分布箱子的概率p_i之间作比较，即类似于方差方法，将二者之差取平方和作为评价基准，从而判断提出的假设是否正确。 同时，概率论中存在定理：若样本数量足够多，当$Ci = n/p_i$时，则当H_0为真的统计量$\sum_{i=0}^{n} f_i^2/np_i - n$，近似服从卡方分布$X^2(k-1)$。检验统计量公式如 2-10 所示。

$$\sum_{i=0}^{n} C_i\left(\frac{f_i}{n} - p_i\right)^2 = \sum_{i=0}^{n} \frac{n}{p_i}\left(\frac{f_i}{n} - p_i\right)^2 = \sum_{i=0}^{n} \frac{f_i^2}{np_i} - n = X^2 \tag{2-10}$$

综上所述，卡方分布拟合检验模型的算法流程图，如图 2-4 所示。

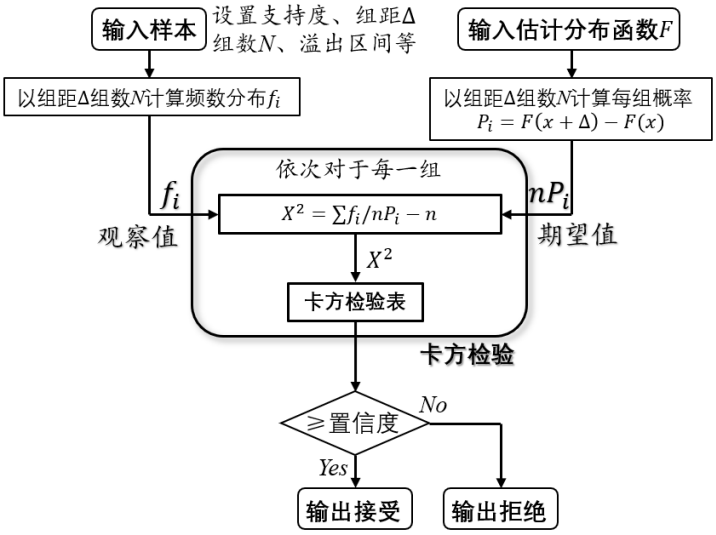


图 2-4 卡方分布拟合检验算法流程图

本文以 KPI 指标(2)RRC 连接建立完成次数为例，将数据导入卡方模型，计算得到的 X^2 结果为 29.503717，满足我们设定的置信区间，故接受其为指数分布的假设。其中，RRC 连接建立完成次数指标的 X^2 分布拟合检验表，如表 2-3 所示。

表 2-3 KPI 指标(2)RRC 连接建立完成次数 X^2 分布拟合检验表

	f_i	np	p
[0,530]	5029	4730.421	0.210
(530,1060]	3700	3736.026	0.166
(1060,1590]	2805	2950.665	0.131
(1590,2120]	2319	2330.397	0.104
(2120,2650]	1809	1840.517	0.082
(2650,3180]	1451	1453.617	0.065
(3180,3710]	1104	1148.048	0.051
(3710,4240]	901	906.713	0.040
(4240,4770]	727	716.110	0.032
(4770,5300]	569	565.575	0.025
(5300,∞]	2089	2124.911	0.094

从上表中，我们可以分析出，样本直方图区间频数和假设的分布函数区间概率二者的数据基本一致，例如在(530, ∞]的各个区间段中，两者频数数值相差极小，即样本的直方图曲线基本贴合指数分布概率密度曲线，而在第一区间[0,530]内，样本数据的频数值要稍大于假设分布的概率值，说明在这个区间内样本直方图曲线要略高于所拟合的指数分布曲线，这与我们在上面第 3.2.1 章节分析的结论相一致，说明 X^2 分布拟合检验及其数据表格，能够很好的反映样本数据和我们所假设的分布之间在各区间段的相关程度，并对所做出的假设进行判定和评价。

表 2-4　KPI 指标 X^2 分布拟合检验表

	指标	X^2 值	上溢门限	组数
2	RRC 建立完成次数	29.504	5300	10
4	E-RAB 建立成功次数	56.706	7000	10
6	E-RAB 释放次数	51.455	7000	10
16	下行 PUSCH PRB 利用率	486.843	141	10
17	下行 PRB 利用率	409.824	66	10
18	下行 PDSCH PRB 利用率 2	407.976	131	10
19	小区内平均用户数	145.502	60	10

表 2-5　KPI 指标(18)下行 PDSCH-PRB 利用率(2015) X^2 分布拟合检验表

	f_i	np	p
[0,13]	7743	7157.483	0.318
(13,26]	5119	4880.917	0.217
(26,39]	2641	3328.454	0.148
(39,52]	1876	2269.779	0.101
(52,66]	1408	1547.835	0.069
(66,79]	1058	1055.518	0.047
(79,92]	839	719.792	0.032
(92,105]	635	490.849	0.022
(105,118]	454	334.726	0.015
(118,131]	289	228.260	0.010
(131,∞]	441	489.386	0.022

对于 KPI 指标 18 的下行 PDSCH-PRB 利用率，X^2 拟合检验表显示，在靠近零的初始区间(0, 26]内拟合曲线的要低于样本频数直方图，而在中部过渡区间 (26,66]内指数分布拟合曲线则相比样本直方图数值要稍微更高一些，而在尾部区间又回到拟合曲线稍低一些的状态。即指数分布拟合曲线并不能很好地反映样本直方图的变化曲率，在首部无法达到样本的数值高度，曲率上样本直方图先比较拟合曲线更陡峭，从拟合图上也可以看出，在中间过渡区间拟合曲线和样本直方图之间存在明显的偏差缝隙。故 X^2 分布拟合检验模型得到检验值为 407.976，数值过大以至于不满足模型的要求，因此我们拒绝 KPI 第 18 号指标下行 PDSCH-PRB 利用率服从指数分布的假设。

最后，卡方拟合检验模型具有通用性，即其检验的使用范围并不局限于指数分布，该算法能够很好的对任意给定的分布函数和样本数据分布之间的拟合程度，进行假设检验并给出评价分析结果。

2.2.3 Gamma 分布

Gamma 分布，是指一种双参数的统计学上重要的连续概率分布。指数分布和卡方分布都是 Gamma 分布的子集，当参数 α 取 1 时，Gamma 分布自动退化为指数分布，当 $\alpha = n/2$ 且 $\beta = 1/2$ 时，Gamma 分布则退化为卡方分布。Gamma 分布的概率密度函数和分布函数

如公式 2-11 所示。其中，Gamma 函数Γ和不完全 Gamma 函数γ，如 2-13、2-14 所示。

$$f(x)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \tag{2-11}$$

$$F(x)=\frac{1}{\Gamma(\alpha)}\gamma(\alpha,\beta x) \tag{2-12}$$

$$\Gamma(x)=\int_0^{\infty}t^{x-1}e^{-t}dt \tag{2-13}$$

$$\gamma(x,s)=\int_0^{s}t^{x-1}e^{-t}dt \tag{2-14}$$

相比于上一章节的指数分布，Gamma 分布更具有普适性。但是，Gamma 分布的统计特征并没有指数分布那么明显，指数分布可以很方便地从均值近似等于标准差这一特征快速发掘，同时 Gamma 分布的数值计算也相对指数分布复杂地多，需要通过计算Γ函数来计算 Gamma 分布拟合曲线，且在卡方拟合检验时，其分布函数 F(x)的计算量也非常繁琐，需要计算不完全 Gamma 函数$\gamma(\alpha,\beta x)$。因此，指数函数可以看作 Gamma 分布的一种快速算法的特例，而 Gamma 函数则是指数函数更加普适性的衍生，二者在阈值检测方法上不可或缺，且相互补充。Gamma 分布的平均值、方差表达式，如公式 2-15、2-16 所示。

$$\mu=\alpha/\beta \tag{2-15}$$

$$\sigma^2=\alpha/\beta^2 \tag{2-16}$$

运用同指数分布章节一样的最大似然参数估计方法，我们可以得出 Gamma 分布的两个关键参数α、β的估计表达式，如式 2-17、2-18 所示。基于如上过程，本文绘制出 Gamma 分布对 KPI、KQI 各指标样本数据的概率密度曲线拟合图，如图 2-5、2-6 所示。

$$\alpha=\bar{X}^2/V[X] \tag{2-17}$$

$$\beta=\bar{X}/V[X] \tag{2-18}$$

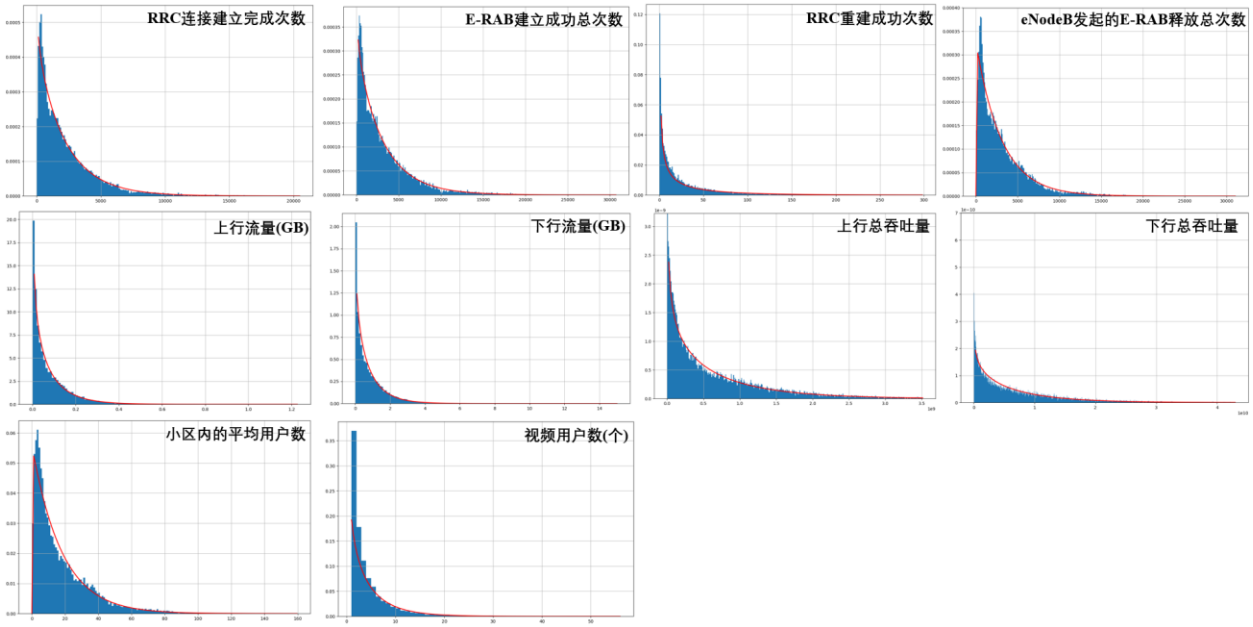


图 2-5 KPI、KQI 指标 Gamma 分布拟合图 (拟合程度较好)

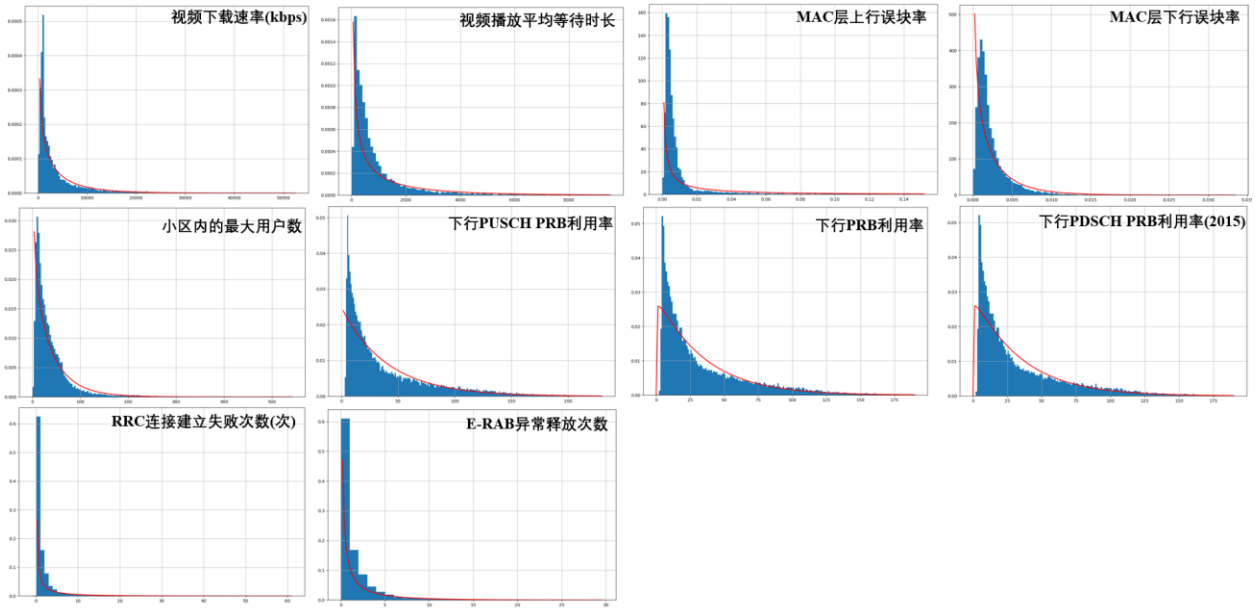


图 2-6 KPI、KQI 指标 Gamma 分布拟合图 (拟合程度一般)

通过上面各指标样本 Gamma 分布拟合图，我们将 KPI、KQI 指标基于其各自 Gamma 分布拟合程度的优劣分为两组分别进行讨论，对于 KPI、KQI 中大部分诸如次数、流量/吞吐量、用户数等类型的指标，其分布拟合曲线能够同频数直方图在趋势上基本保持一致，大尺度上能够近乎完美地贴合直方图曲线。而对于部分诸如利用率、误块率以及 KQI 中下载速率、播放等待时长等指标，虽然从整体上分布拟合曲线能大体反映频数直方图的趋势，但拟合曲线同样本原始直方图相比又存在相对明显的偏差，拟合曲线导数过大过小等问题，导致 Gamma 方法并不能完美的反映指标分布的趋势。例如 PRB 利用率，Gamma 分布的拟合曲线在该指标的中部过渡段区间，拟合的斜率小于样本原始直方图曲线，导致在左侧靠近零的区间拟合曲线过低，而在中间区间拟合曲线又过于高的偏差。还有一些如 RRC 失败次数、E-RAB 异常释放次数等指标，虽然 Gamma 分布能够拟合其直方图趋势，但由于其自身的离散性和分布较狭窄的特征，并不适合用 Gamma 方法来进行拟合反映其分布。同样的，我们基于 3.2.2 提出的卡方拟合假设检验方法，来对 Gamma 分布的拟合结果进行评价和判决。Gamma 分布卡方检验表，如表 2-6 所示。

表 2-6　KPI、KQI 指标 X^2 分布拟合检验表

	指标	X^2 值	上溢门限	组数
2	RRC 连接完成次数	16.915	5300	10
3	RRC 重建成功次数	371.79	20	10
4	E-RAB 建立成功次数	10.454	7700	10
6	E-RAB 释放次数	50.935	7400	10
19	小区内平均用户数	152.952	53	10
21	上行流量	42.097	0.68	10
22	下行流量	137.436	3.68	10
23	上行吞吐量	78.34	4.7E8	10
24	下行吞吐量	154.259	2.89E0	10

Q2	下载速率	124.832	18500	10
Q5	视频用户数	173.456	23	10

卡方分布检验表显示，Gamma 分布对 KPI 第(2)、(4)指标的拟合程度非常好，卡方检验值略超过 10，满足自由度为 11 时(包括上溢区间)，显著性水平 α 为 0.25 和 0.10 的判决条件。而其他指标的卡方检验值显示拟合程度一般，除了 RRC 重建成功次数外检验值都在 50 或 100 左右，但因为本章节阈值检测的目的，关注的重点在样本分布的尾部即存在异常的区域，而曲线往往在峰值和中部过渡区间存在拟合欠佳，尾部拟合程度较好。因此，我们可以基于实际情况对卡方检验值一般但分布曲线拟合尚可的指标考虑接受。

2.2.4 对数正态分布

对数正态分布是一种随机变量恒大于零，且样本数据的对数满足正态函数的分布。对数正态分布的概率密度函数，如公式 2-19 所示。于是，我们可以将样本数据进行对数变换后，再对其分布进行分析。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}\quad(x>0) \tag{2-19}$$

我们采用与上述分析指数分布时一样的参数估计方法——极大似然估计，有

$$L(\mu,\sigma)=\prod_{\mathrm{i=n}}^{1}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}=(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}}e^{-\frac{\sum_1^n(x_i-\mu)^2}{2\sigma^2}} \tag{2-20}$$

$$\ln L=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{2-21}$$

令

$$\begin{cases}\dfrac{\partial L}{\partial\mu}=0\\[2ex]\dfrac{\partial L}{\partial\sigma^2}=0\end{cases} \tag{2-22}$$

解得

$$\begin{cases}\hat{\mu}=\bar{x}\\[2ex]\widehat{\sigma^2}=\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\end{cases} \tag{2-23}$$

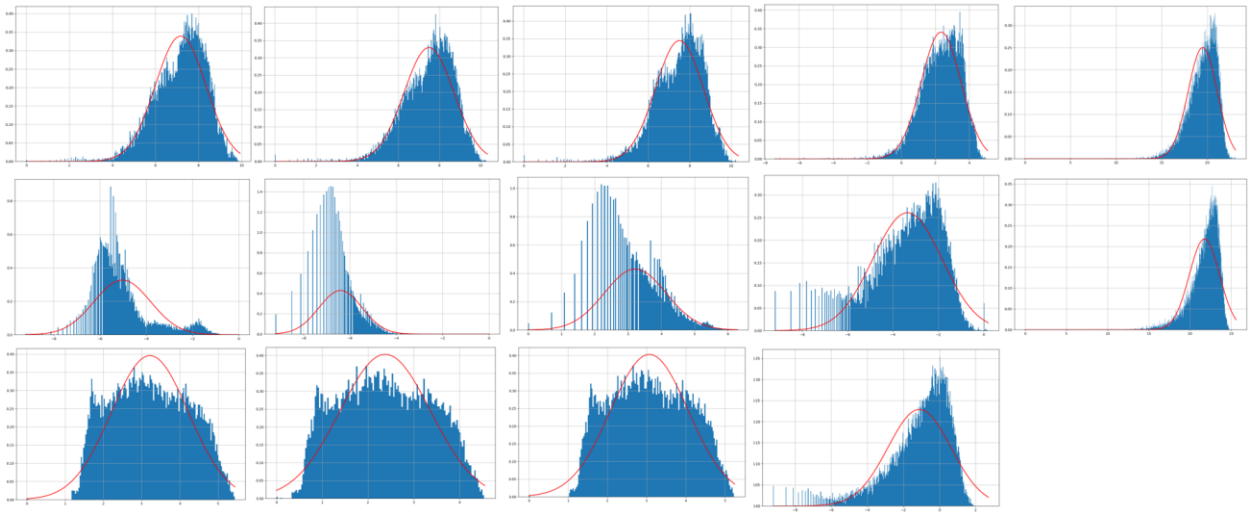


图 2-7 部分 KPI 指标极大似然估计下的正态曲线拟合概览图

2.2.5 偏度-峰度检测

对于正态分布的假设检验和评价方法，可以使用卡方拟合检测，但统计上有更适合的专门针对正态分布的检测模型——偏度-峰度检测。偏度，是指样本概率密度分布相对于其均值的不对称程度，从数值的正负上看，可分为左偏和右偏，在统计学上偏度特征可以通过样本标准化变量的三阶矩计算得出，如公式 2-24 所示。峰度，指的是样本概率密度分布在均值附近的陡峭程度，同样的，该特征可以通过样本标准化变量的四阶矩计算得出，如公式 2-25 所示。

偏度公式：

$$E[\left(\frac{X-E(X)}{\sqrt{D(X)}}\right)^3]=\frac{E\left[\left(X-E(X)\right)^3\right]}{D(X)^{3/2}} \tag{2-24}$$

峰度公式：

$$E[\left(\frac{X-E(X)}{\sqrt{D(X)}}\right)^4]=\frac{E\left[\left(X-E(X)\right)^4\right]}{D(X)^2} \tag{2-25}$$

而针对正态分布而言，概率密度分布应当具有无偏性，即偏度为零，均值、中位数、方差为同一个数。其次，正态分布函数的峰度恒为 3。基于上述特征，倘如样本服从正态分布，则应当有偏度等于 0，且峰度等于 3。综上，本算法首先计算样本的中心矩 D(X)、标准化变量的三阶矩和四阶矩 $E[(X-E(X))^3]$, $E[(X-E(X))^4]$。然后，计算得到样本分布的偏度、峰度数值。最后，将得到的结果代入假设检验，根据是否满足前提设定的置信度进行假设检验分析，从而实现对正态分布和对数正态分布拟合函数的检测和评价。基于该模型，我们计算并给出 KPI、KQI 各指标样本的偏度-峰度数值结果，如表 2-7、2-8 所示。

表 2-7　KQI 指标偏度-峰度检测表

	指标名称	偏度	峰度	偏度置信度	峰度置信度
2	下载速率	0.147	3.035	6.002	0.723
3	播放平均等待时长	0.109	3.505	4.467	10.332

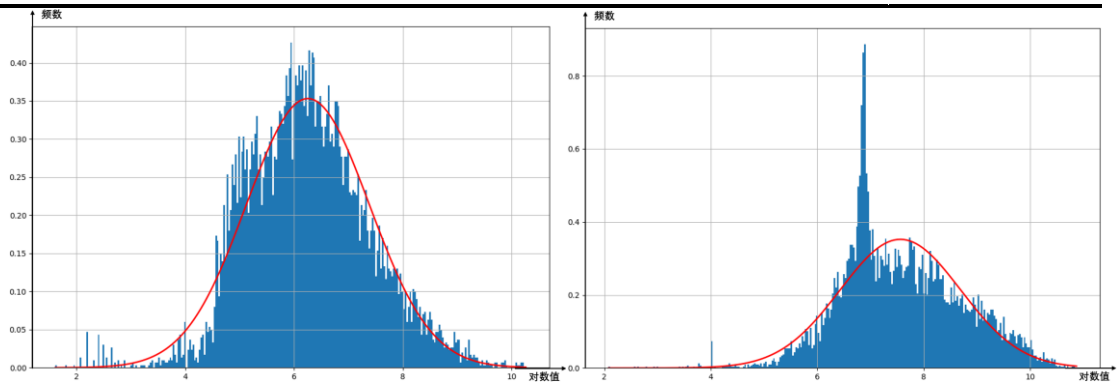


图 2-8. KQI(左)视频播放平均等待时长(右)下载速率的正态曲线拟合

此处，我们绘制出 KQI 指标(2)下载速率和(3)视频播放平均等待时长的对数值的频数直方图及其正态分布拟合曲线，如图 2-8 所示，连同上面偏度-峰度数值表一起来分析。

首先，对于 KQI 第 3 指标视频播放平均等待时长，其峰度略大于 0，即显示该指标有轻微右偏倾向，可知其平均数 6.251、中位数 6.206、众数 4.912，满足一般右偏情况下均值>中位数>方差的通常特征，且该指标与正态拟合曲线基本贴合，偏度不会太大，故所计算的偏度数据与直方图相符，计算无误。同时，该指标峰度相较于理想的数值 3 略大 0.5，即其峰度比标准正态分布稍显陡峭，直方图中显示，因样本数据噪音影响，在频数曲线在小尺度与正态拟合曲线相比有所起伏，其峰值区间同样受噪音影响略高于拟合曲线并更陡峭一些。从模型所得峰度-偏度置信值相对较小，且该指标对数的频数直方图在整体上和拟合的正态曲线趋势基本一致，因此我们接受该指标服从对数正态分布的假设。

其次，对于 KQI 第 2 指标视频播放平均等待时长分析，虽然该指标对数值的峰度和偏度都相对目标值差距较小，但从其频数直方图同拟合的正态曲线对比来看，其概率密度分布在(6.5, 7)的小范围区间内有一明显突起，突起的幅度相对较大数值超过 0.8，且其样本噪点的小尺度起伏相较于拟合曲线，偏差也相对较大。因样本数据分布不规则和小范围突起的原因，使得检测模型误以为其峰度-偏度满足正态分布，但其频数直方图图像并不满足，于是我们拒绝该指标服从正态分布的假设。

表 2-8　KPI 指标偏度-峰度检测表

	指标	偏度	峰度	偏度置信度	峰度置信度
2	RRC 建立完成次数	-0.669	3.748	40.946	22.913
4	E-RAB 建立成功次数	-0.801	4.514	49.091	46.376
6	E-RAB 释放次数	-0.886	5.136	54.272	65.425
8	MAC 层上行误块率	1.283	4.612	78.580	49.372

9	MAC 层下行误块率	0.985	9.607	60.327	202.395
10	E-RAB 建立成功率	-49.884	2492.297	3055.398	76249.349
11	同频切换成功率	-7.072	51.036	433.155	1471.388
12	异频切换成功率	-5.342	29.542	327.200	813.005
14	eNodeB 内切换成功率	-42.393	3202.215	2596.578	97994.761
16	下行 PUSCH PRB 利用率	0.071	1.994	4.341	30.798
17	下行 PRB 利用率	0.059	1.984	3.617	31.122
18	下行 PDSCH PRB 利用率 2	0.057	1.989	3.509	30.952
19	小区内平均用户数	-0.805	4.930	49.336	59.140
20	小区内最大用户数	0.048	2.900	2.911	3.043
21	上行流量	-0.794	3.545	48.613	16.703
22	下行流量	-1.298	5.107	79.511	64.545
23	上行吞吐量	-1.451	10.947	88.889	243.421
24	下行吞吐量	-1.970	13.370	120.637	317.639

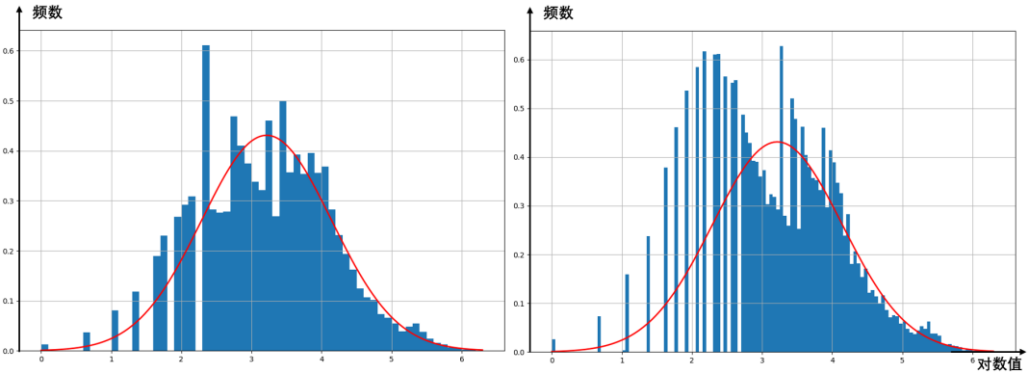


图 2-9. KPI 小区内的最大用户数正态曲线拟合(左 0.1 组距, 右 0.03 组距)(均值: 3.2164, 方差: 0.8569)

同上述 KQI 分析一样，我们对 KPI 指标(20) 小区内的最大用户数的对数值的偏度-峰度检测数据和频数直方图及其正态分布拟合曲线一同分析。该指标偏度略大于 0 显示轻微右偏倾向，峰度略小于目标值 3 则其波峰部分的陡峭程度相比标准正态稍显平缓，从频数直方图上，在横坐标大于 4 时样本数据概率密度基本和正态拟合曲线趋势保持一致，而在小于 4 的区间内，因为对数运算和人数指标本身的离散性特征，而造成数据分布在(0,4)的区间里产生分离使得图像看起来不均匀连续，当频数直方图组距较小时(例 $\Delta = 0.1$)，正态拟合曲线能够基本贴合概率密度分布反映其变化趋势，当组距较大时(例 $\Delta = 0.03$)上述影响便凸显出来。不过，本文阈值检测判定的是大于门限的区间的数据，即图像右边大于 4 的区间不受上述影响。因此，我们可以接受 KPI 小区内的最大用户数指标满足服从对数分布的假设。

2.3 异常检测方法

2.3.1 分布函数阈值确定

对于已知分布函数的指标，在给定非异常值百分比门限 K 的条件下，我们可以基于具体的分布函数方程直接确定阈值的表达式。例如，指数分布的阈值 $X_{threshold}$，可以推算得出，

$$X_{threshold} = -\theta \ln(1-K) \tag{2-26}$$

对数正态分布，我们可先将数据取对数，转为正态分布，然后选择教科书上经典的几个值作为阈值参考，如公式 2-27 所示。也可以从正态分布的分布函数来计算，分布函数如公式 2-28 所示，其中 *erf* 是指误差函数，是 Gamma 分布一章节中叙述过的不完全 Gamma 函数$\gamma(x,s)$的一类特殊子集。因此，对 Gamma 分布的阈值推算，和正态分布情景是相类似的，如公式 2-29 所示，于是我们二者放在一起进行讨论，

$$\begin{aligned} P\{\mu - \ \sigma < X < \mu + \ \sigma\} &= 1 - 2 \cdot \Phi(-1) = 68.26\% \\ P\{\mu - 2\sigma < X < \mu + 2\sigma\} &= 1 - 2 \cdot \Phi(-2) = 95.44\% \\ P\{\mu - 3\sigma < X < \mu + 3\sigma\} &= 1 - 2 \cdot \Phi(-3) = 99.74\% \end{aligned} \tag{2-27}$$

$$F(x) = \frac{1}{2}(1 + erf(\frac{x-\mu}{\sigma\sqrt{2}})) \tag{2-28}$$

$$F(x) = \frac{1}{\Gamma(\alpha)}\gamma(\alpha, \beta x) \tag{2-29}$$

如果我们直接从上述分布函数方程来推导阈值门限是困难的，以为我们需要知道不完全 Gamma 函数$\gamma(x,s)$和误差函数 erf 的反函数，而其反函数并非简单的初等函数。因此，我们可以利用数值计算中的二分法，通过迭代收敛来确定其具体阈值数据。具体程序如 Algorithm 1 所示。我们以 KPI 第 2 号指标 RRC 连接建立完成次数指标作为测试样本，设定 90%为非异常比例参数，基于指数分布阈值公式 2-26 算得的门限数值为 5171.238，而利用二分法和 Gamma 分布拟合计算得到的阈值门限为 5264.992，二者数值相差较小，说明二分算法能够正确地给出估计分布函数下的指标样本数据的阈值门限结果。

Algorithm 1 二分法 C++程序

```
template<typename F>
double BisectionMethod(double st, double ed, F&& f) {
    double mid = (st + ed) / 2;
    if (f(st) * f(ed) > 0 || f(st) * f(mid) == 0 || ed - st < 1E-9) return mid;
    return f(mid) * f(ed) > 0 ?
        BisectionMethod(st, mid, f) :
        BisectionMethod(mid, ed, f);
}
```

2.3.2 箱型图

箱型图是一类基于样本变量的分位数性质来进行统计的分析的模型，该统计方法具有标记疑似异常值的能力。箱型图主要标记数据的中位数、上四分位数和下四分位数，作为“箱子”的中间线、顶部和底部。然后，计算上下四分位数的差值，作为参考量Δ。同时，箱型图取高于上四分位数1.5Δ的数值为上阈值，而低于下四分位数1.5Δ的数值为下阈值。从而实现了对异常值的检测和标记。该模型的异常检测方法具有通用性，对是否知道样本数据具体的分布函数并没有要求，可以适应各种类型和不容易找到样本分布函数的数据，并给出能够接受并进一步研究的异常阈值。基于箱型图模型算法，我们计算得到 KPI、KQI 各指标的中位数和上下阈值，如表 2-9、2-10 所示。

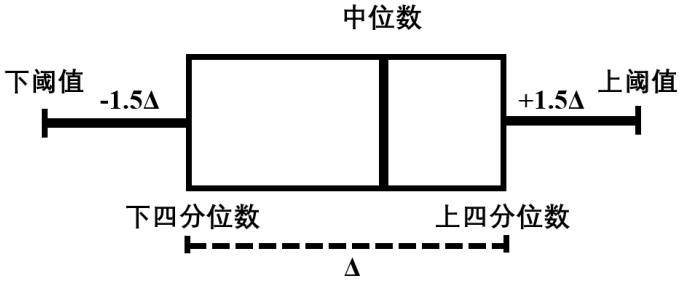


图 2-10 箱型图原理示意图

表 2-9 KPI 箱型图数据表

指标序号	中位数	下四分位数	上四分位数	下阈值	上阈值
1	0	0	1	-1.5	2.5
2	1534	597.5	3082	-3129.25	6808.75
3	11	2	34	-46	82
4	2135	818	4361	-4496.5	9675.5
5	0	0	1	-1.5	2.5
6	2140	848	4254	-4261	9363
7	0	0	0	0	0
8	0.0049	0.003	0.0092	-0.0063	0.0185
9	0.0016	0.9	0.0028	-0.00195	0.00565
10	100	99.969	100	99.922	100.047
11	100	99.856	100	99.641	100.216
12	100	100	100	100	100
13	0	0	1	-1.5	2.5
14	1	1	1	1	1
15	0.9998	0.998	1	0.994	1.0036
16	24.196	10.950	56.898	-57.972	125.821
17	10.481	4.806	24.303	-24.439	53.548
18	20.966	9.616	48.617	-48.886	107.119
19	11.685	4.866	25.275	-25.748	55.888
20	25	13	48	-39.5	100.5
21	0.044	0.0136	0.1107	-0.132	0.256
22	0.484	0.139	1.152	-1.382	2.672
23	378×10^6	116×10^6	951×10^6	-1135×10^6	2202×10^6
24	4159×10^6	1191×10^6	9897×10^6	-12×10^9	23×10^9

表 2-10 KQI 箱型图数据表

指标序号	中位数	下四分位数	上四分位数	下阈值	上阈值
1	100	100	100	100	100
2	1681	904	4146	-3959	9009
3	496	235	1053	-992	2280
4	0	0	0	0	0
5	2	1	5	-5	11

2.3.3 特殊类型数据

对于以成功率、失败率为代表 KPI、KQI 指标数据，往往呈现单极化和双极化的特征。首先，对于单极化特征，我们以 KPI 指标(7)小区用户面上行丢包率为例，如图 2-11 所示，可以看到 83.83%的数据都为零，而其它非零样本数据绝大部分也只与零偏差在万分之一附近，该指标最大值仅仅为 0.0304。于是，对于诸如此类的单极化数据，我们可以将其集中的那一个数值极点(如零)视作正常数据，而偏离的样本点看作为异常数据，从而确定判决异常的区间和阈值。而对于双极化数据，即在 0 和 1 两个极端点附近大概率分布，而在 0 和 1 中间的过渡区域分布较少。我们以 KPI 指标(13) ESRVCC 切换成功率为例进行讨论，如图 2-11 所示，可以从直方图中看出，绝大多数数据均分布在 0 和 1 两个极值点上，百分率而在中间过渡带的数据非常稀少。于是，对于双极化数据，可以以 1/2 为基准阈值，将数据二值化为 0 和 1，也可以将箱型图结果作为阈值门限，两种方法择其优，同时，我们选取代表“否定”的指标作为异常数据值(如，成功率应为 0%，失败率应为 100%)。

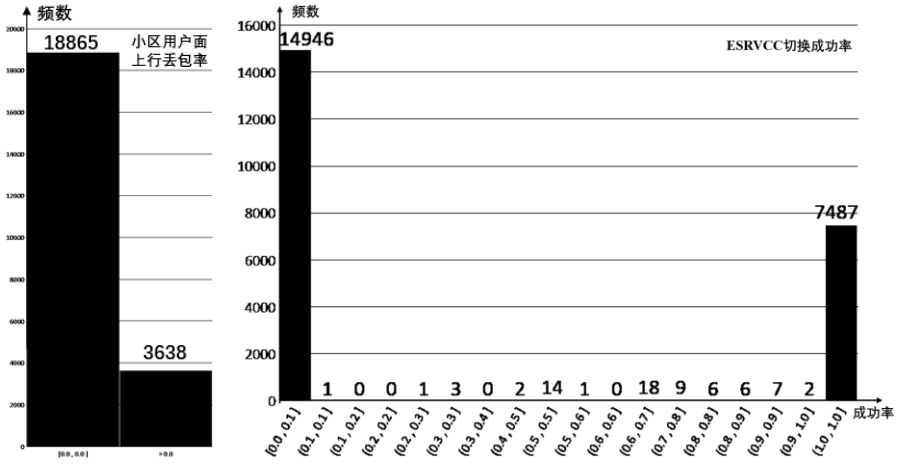


图 2-11 单极化和双极化指标样例
(左)小区用户面上行丢包率(右) ESRVCC 切换成功率

此外，还有部分的特殊数据，如 KQI 指标(2) 视频下载速率，对于形似指数函数左密右疏的样本数据分布，异常区间往往右侧稀疏的数据段，而对于下载速率，可以看到其在(800，900] 达到很高的峰值，说明普遍的下载速率是在该区间内，右侧稀疏段则指的是速率较快样本点，从指标意义来看异常数据应当在左侧靠近零的区间内，而这部分区间无法使用上述分析的拟合的指数分布和箱型图来判定(箱型图阈值数据 < 0)，需要特殊情况特殊分析，可以根据其当前已知的频数直方图数据，以给定的比率划定取值门限，将其升序排列，9999 个样本点第 999 号样本点为 539，即该指标阈值数据。同时，对于部分尚未找到其具体分布函数的指标，我们也可以使用百分比比例划定方法，确定其判决门限。

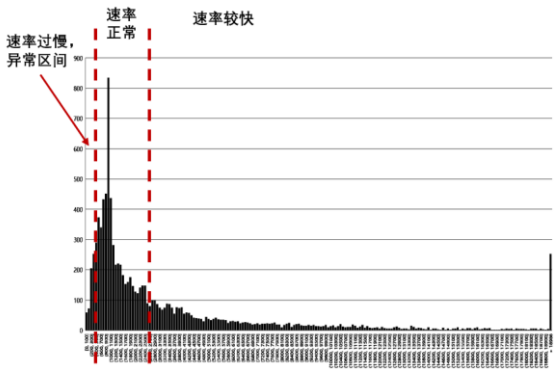


图 2-12 特殊数据指标样例

2.4 单维异常检测综合模型

综上所述，我们将上述所有算法模型整理并统一，设计出单维数据异常阈值检测综合模型体系。在综合模型中，对于输入的任意的 KPI、KQI 单维度指标样本数据，模型首先考虑使用统计学上经典的随机变量分布对样本的频数直方图包络进行拟合，如果样本分布具有均值近似等于标准差等专有性特征，模型优先使用与其特征匹配的分布函数去拟合。

其次，在随机变量分布函数的拟合处理过程中，模型使用极大似然估计方法，来通过输入的 KPI、KQI 样本数据实现对分布函数特征参数数值的估计，完成拟合工作。然后，模型进一步采样卡方拟合检验方法，来评价分布函数对样本直方图包络拟合的优劣程度，并判决我们是否接受样本数据服从该随机变量分布函数的假设。同时，针对正态分布、对数正态分布，本模型则使用更专门的峰度-偏度假设检验方法来进行评估。如果卡方拟合检验的判决结构为接受，那么我们就可以通过分布函数或二分法等算法，去计算在给定比率下该指标的异常判决的阈值门限结果，实现模型的最终目的。

如果卡方拟合检验的判决为拒绝，现有的分布函数均不能对该 KPI、KQI 单维指标很好地拟合，那么我们可以直接交给采用箱型图模型，通过上下四分位、中位数确定该指标的阈值，或根据百分比比率划定阈值。同时，对于单极化、多极化数据，则通过二值化等方法判定，而对于特殊数据，则要人为地具体情况来具体分析。单维数据异常阈值检测综合模型流程图，如图 2-13 所示。最后，我们依次对 KPI、KQI 各指标单维数据，构建了适合其数据自身分布的异常检测的判定方法，计算并给出了每个指标判定异常的阈值，如表 2-11、2-12 所示。将阈值结果同之前箱型图数据表相比，数量级相同且数据近似，进一步交叉验证了本章节单维度异常检测模型方法的正确可靠且性能良好。

表 2-11 KQI 样本异常检测算法及其阈值

指标序号	异常检测方法	判断阈值
1	单极化	< 100
2	特殊分析	< 539
3	对数正态分布	> 2219.961
4	单极化	> 0
5	Gamma 分布	> 9.883

表 2-12　KPI 样本异常检测算法及其阈值

指标序号	异常检测方法	阈值门限	指标序号	异常检测方法	阈值门限
1	箱型图	> 2.5	13	二值化	< 1/2
2	指数分布	> 5171.238	14	单极化	< 1
3	Gamma 分布	> 81.449	15	单极化	< 0.994
4	指数分布	> 7264.965	16	百分比比例	> 100.931
5	箱型图	> 2.5	17	百分比比例	> 41.882
6	指数分布	> 7041.427	18	百分比比例	>83.765
7	单极化	> 0	19	指数分布	> 40.933
8	箱型图	> 0.0185	20	对数正态分布	> 81.709
9	箱型图	> 0.00565	21	Gamma 分布	> 0.193
10	单极化	< 99.922	22	Gamma 分布	> 1.929
11	单极化	< 99.641	23	Gamma 分布	> 1.658E9
12	单极化	< 100	24	Gamma 分布	>1.657E10

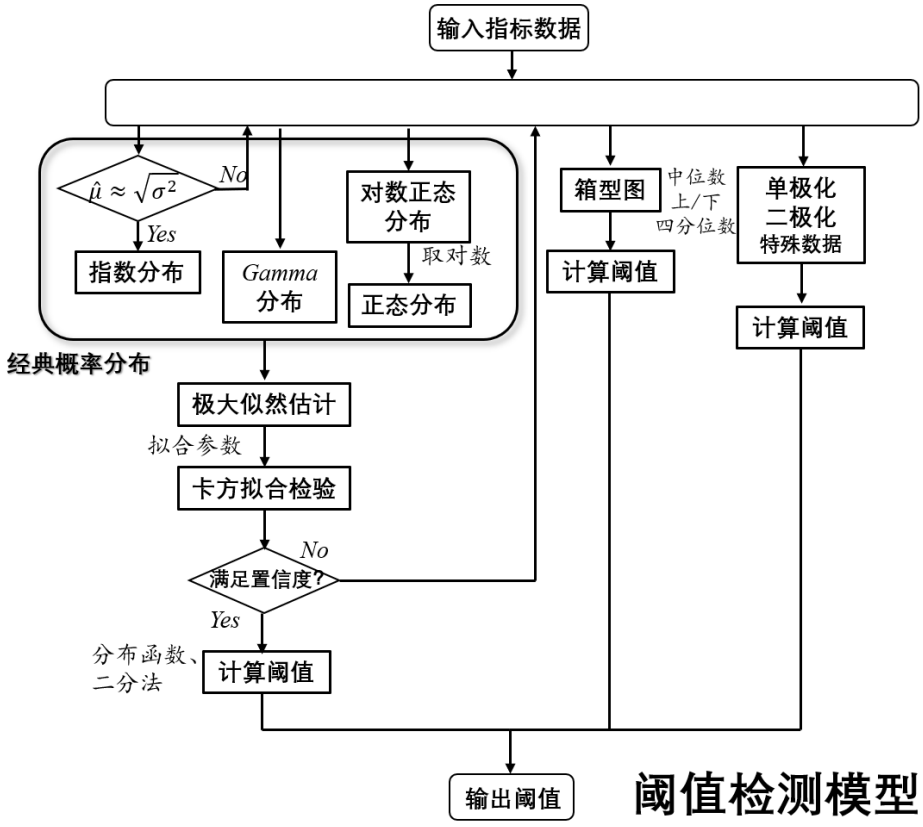


图 2-13 单维数据异常检测综合模型流程图

2.5 本章小结

本章基于统计学方法提出一种单维数据异常检测模型，构建模型并确定了 KPI、KQI 共 29 个指标的异常数据阈值，同时基于该结果对原样本数据进行判决，计算得到了筛选后异常数据的表格，并运用到我们的下一个关键算法——Apriori 关联规则上。

3 Apriori 关联规则

3.1 Apriori 关联规则算法

3.1.1 基本概念

关联规则，指的是蕴含式$A \to B$，其中 A 是先导项，B 是后继项，反映了 A 与 B 之间的依存与关联特征。**支持度**$P(AB \dots C)$，是指$AB \dots N$两个及两个以上的元素项同时存在的集合，在随机出现的集合中的概率。我们可以用已知的同时存在目标元素$AB \dots N$的集合占总集合数的比率进行估算。如式 3-1 所示。

$$P(AB \dots C) \cong \frac{N_{AB \dots C}}{N_{all}} \tag{3-1}$$

置信度$P(A \to B)$，是指关联规则$A \to B$可被信任的程度，其数值等同于当 A 元素项存在时 B 项也存在的条件概率$P(B|A)$ ，我们可以使用条件概率公式，完成对置信度数值的估算，如式 3-2 所示。进一步，我们可以推广至任意数量的多元素项之间的关联规则置信度测算，如式 3-3 所示。

$$P(A \to B) = P(B|A) = \frac{P(AB)}{P(A)} \tag{3-2}$$

$$P(A \dots B \to C \dots D) = P(C \dots D|A \dots B) = \frac{P(A \dots BC \dots D)}{P(A \dots B)} \tag{3-3}$$

3.1.2 算法流程

(*input*) 输入数据：需要分析的指标项集合，设定的支持度、置信度。

首先，本算法对数据进行预处理。我们将 *KPI*、*KQI* 各指标经过第二章异常检测后的数据，以时间、基站地点为基准，进行对齐合并成一张表格，表格中的每一行数据即是一个项集，在该集合中保证其元素在时间和空间上是相同的，而项集中的元素即是异常数据所对应的指标。其次，我们将该表格中每一行存在异常的数据所对应的指标，按其顺序为编码提取出来，如指标 *RRC* 连接建立失败次数在序列的第一位编码为(1)。于是，我们便得到了一张可以进行下一步分析的指标项集合。而，对于所需输入的目标支持度、置信度阈值，我们可以按照经验选择 0.03，0.1，因为部分指标异常样本占总表格的比例相对较低，如果目标支持度设置太高算法就会将这些指标忽略，以至于无法计算他们的关联规则数据，所以支持度、置信度选择应当适合实际需求，同时，在进一步的实验中，可以按实际需求进行修正。

(1) 依次生成 *K* 个元素的频繁项集。

本算法从二元素的项开始，按元素个数逐次增加，依次生成满足条件 *K* 个元素的项的集合。而生成项集方法，我们可以采用$F_{k-1} \times F_{k-1}$ 方法。在该方法中，我们以上一次生成的 *K-1* 个元素项的集合表为原表，将其前 *K-2* 个元素为参考基准(表已经按顺序排列好)，

当两个集合前 *K-2* 个元素完全相同时，且第 *K-1* 个元素不相同时 (因为集合表应当满足各个集合两两不同，所以该条件会自动满足)，我们将这两个集合进行合并，所生成的含有 K 个元素的集合放入新表。如此步骤循环遍历，便实现了对项集表的生成。同时，对与初始一个元素的项集表F_1，我们可以之间遍历输入表，将其所有出现过的元素项，拷贝至F_1表中作为单元素集合即可。

(2) 计算每个生成的频繁项集的支持度 *P*(*AB…C*)，保存满足目标支持度阈值的集合。

由基本概念所述，支持度 *P(AB)* 可以用同时存在目标元素$AB\ldots N$的集合占总集合数的比率进行估算。于是我们对第一步生成的 *K* 元素项集表的每一个集合逐个操作，对输入的指标项集合表依次遍历进行比较，如果相对比的两个集合，前者(生成表)是后者(输入表)的子集时，我们将生成表集合的频数记录加一。直至所有生成表集合都计算完为止结束。然后，我们对所有记录的频数，除以输入表项集总数，从而实现支持度的计算。最后，我们将计算得到的结果，对比输入的支持度阈值的，滤除所有不满足门限的集合，并将满足的数据保存至频繁项集总表中，以进行下一步的分析。

(3) 关联规则蕴含式生成

基于上述整理的频繁项集，本算法以每一项的元素个数 K 为顺序，从二元素项开始，依次生成关联规则蕴含式。具体操作为，对于当前 K 元素的项集(记为 C_k)，首先统计有哪些元素出现过，整理成一张单元素项的表，作为蕴含式右侧的后继项(记为 Y_{k1})，然后，对 C_k 的每一项遍历，如果 Y_{k1} 的某个后继项真包含于 C_k 的第 i 项，则将 C_k[i]减去 Y_{k1} 的哪一项并记录在先导项表 X_{k1} 中，由此计算出 C_k 关于后继项表 Y_{k1} 的先导项表 X_{k1}，而后继项表 Y_{k1} 和先导项表 X_{k1} 的每一项相对应的形成了我们所需的关联规则蕴含式。基于该流程，我们利用与步骤(1)相同的项集生成方法，将 Y_{k1} 生成 $Y_{k2}\ldots Y_{kj}$ 直至 j＝k-1 为止，并相应的生成与之对应的先导项表 $X_{k2}\ldots X_{kj}$，如此，算法就实现了生成所有关联规则蕴含式的目的。

(4) 计算置信度 *P(A...C→B…D)*

如上所述，置信度的计算基于条件概率公式，我们用处理的集合的本身的支持度，除以规则左边 X 集合的支持度，即得到当前规则的置信度，然后我们将结果对比置信度门限如果满足，则将该规则保存至规则总表，如小于门限，则将该规则删去。

(*output*) 最后我们保存的满足目标置信度的关联规则表，即是最后输出的结果。

综上所述，Apriori 关联算法流程图，如图 3-1 所示。

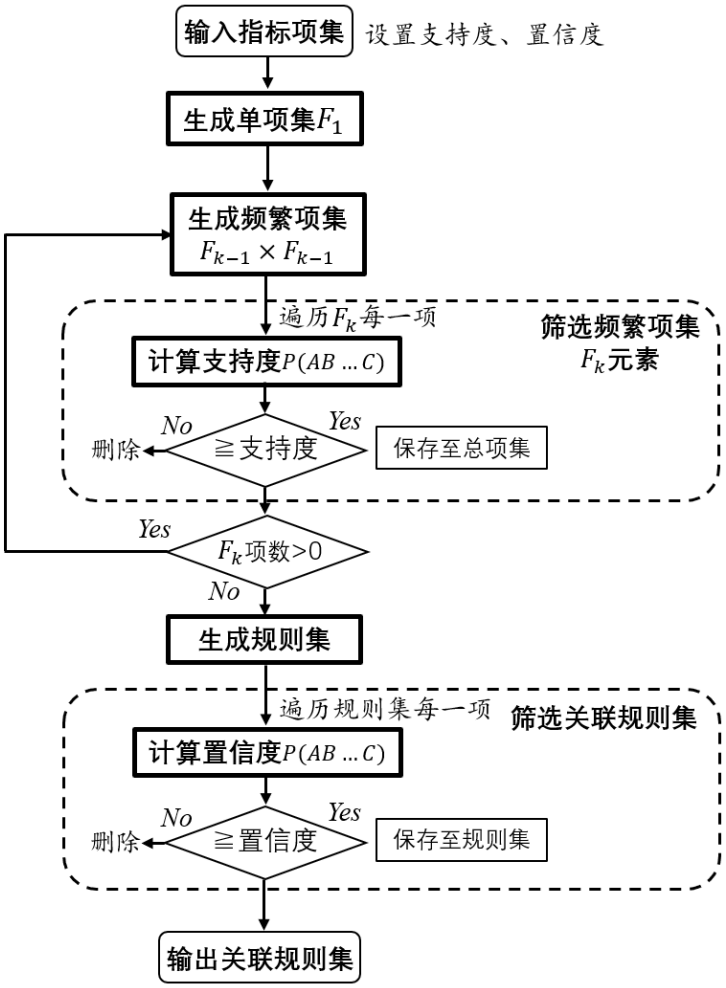


图 3-1 Apriori 关联算法流程图

3.2 数据挖掘结果

我们基于上述搭建的Apriori关联模型，将第二章单维度异常检测模型确定的阈值门限，对 KPI、KQI 共 29 个特征指标进行判决，并将每一行中判定异常的样本以其所在指标的序号加以保存，然后我们将每一行存储的异常指标的序号作为一个项集整理成表，输入至本章节的 Apriori 关联模型中，并生成各指标集合之间的关联规则，计算出每一个满足条件的关联规则的置信度输出，整理并绘制成表格，其中，KPI、KQI 单对单指标间关联程度热力图，如图 3-2 所示。

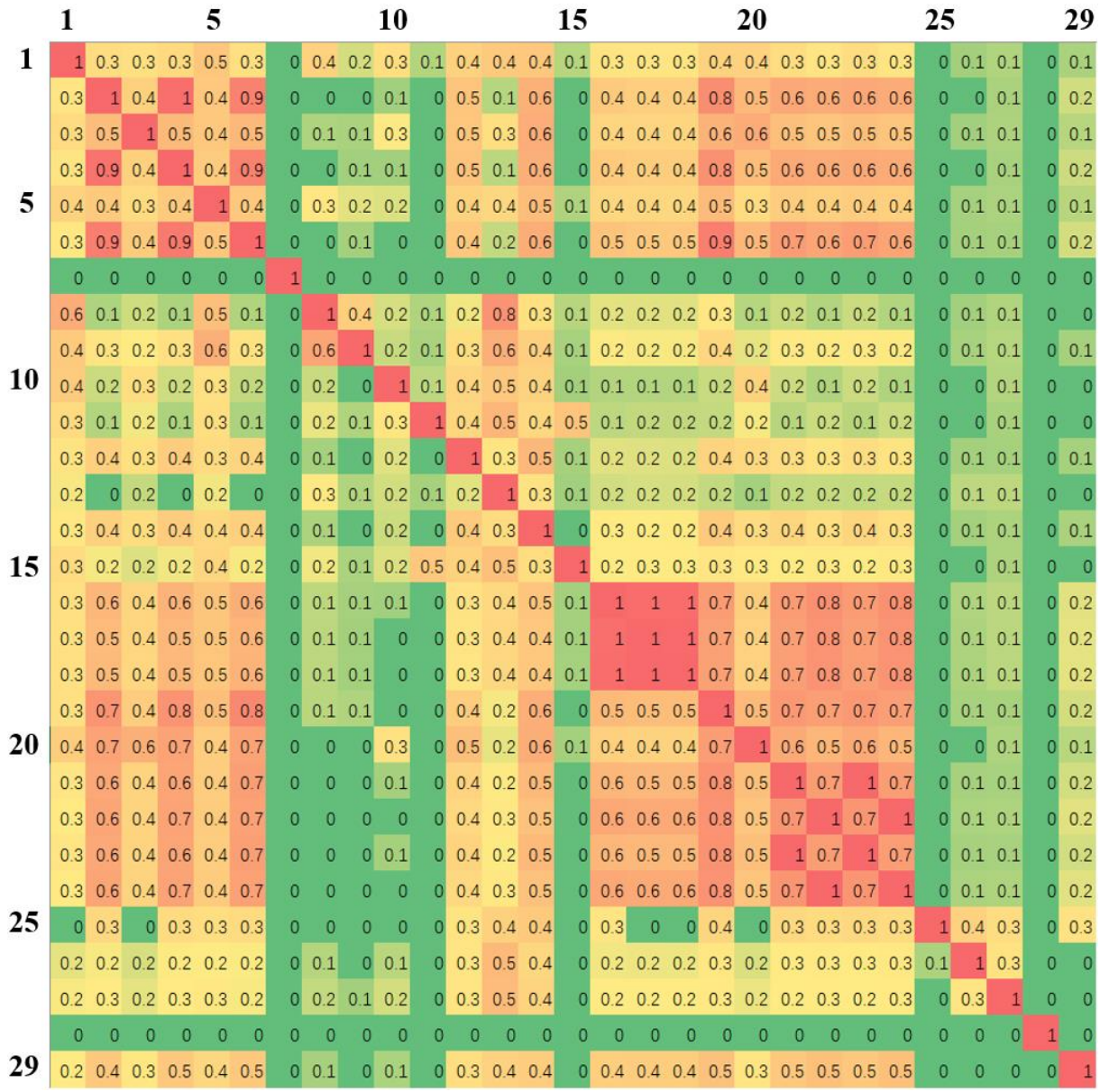


图 3-2 单对单指标间关联程度热力图

对上面关联程度热力图进行分析，横坐标是 KPI、KQI 共 29 个指标，正方形矩阵的每一个数据点(A, B)，指的是从 A 到 B 的关联置信度数值 P(A->B)。该矩阵的对角线只均为 1，这是显然的，因为自己本身的条件概率为 1，$P(A \rightarrow A) = P(A|A) = P(A)/P(A) = 1$。其次，热力图矩阵有数个明显特征的局部子区块。例如，KPI 第 16、17、18 指标，名称是下行 PUSCH.PRB 利用率、下行 PRB 利用率、下行 PUSCH.PRB 利用率(2015)，从指标的实际物理意义来看，这三个指标应当属于同一种物理本质，只是测量的规则和标准不同而导致其样本数值之间有所差异，但在热力图中显示这三个指标之间的关联程度均为 100%，清晰地指出了 KPI 第 16、17、18 指标物理本质的同一性。而对于 KPI 第 21、22、23、24 号指标，名称为上/下行流量、上/下行总吞吐量。我们可以从热力图上看到，上行流量和上行总吞吐量之间，与下行流量和下行总吞吐量之间的关联程度为 1，这是显然的，因为流量和总吞吐量指的都是数据量在给定时间内的通过量，只是其测量计算的公式稍有差异，

一个在通信专业中常用，一个在计算机领域中常用。而在上/下行之间的关联程度均为 70%，说明上/下行之间的异常数据的发生有很强的关联性。

同时，对于热力图中关联置信度数值较大，颜色较红热的数据点，例如(16,17,18)=>(21,22,23,24)，即 PRB 利用率=>上/下行流量的关联度达到了 0.7 至 0.8，而反向关联度在 0.5 至 0.6，PRB 利用率反映的是无限通信资源的利用情况，为系统是否需要扩容及优化给出参考，于是，流量/吞吐量与 PRB 利用率之间存在关联是正确的。并且，平均用户量相比最大用户量，对通信系统各指标的关联程度要更高一些。说明用户人数的平均数比峰值，对通信系统产生的影响更显著，需要关注的价值更大，符合实际的生活经验。同时，平均用户量指标，对 RRC 连接完成/异常次数、E-RAB 建立成功/异常/释放次数的关联程度也较显著，因为用户人数的增加通常导致各种次数指标的上涨，并使得通信系统产生了更大的数据流量。

针对 KPI 与 KQI 指标之间的关联性分析，播放成功率指标与 ESRVCC/eBodeB 间切换成功率关联程度相对更大，可能是由于在通信基站之间的切换更容易导致视频播放的不顺利乃至失败。其次，下载速率指标的热力图与播放成功率相似，数值差距不大，说明播放成功率和下载速率指标的物理意义和本质，应当是相近的。而用户数指标则与上下行流量、PRB 利用率、平均用户数、RRC/E-RAB 的连接/释放数等 KPI 指标关联相对紧密，同时播放用户数的行列的热力图规律与 KPI 中平均用户数的热力规律基本相似，因为这是同一性指标，而 Apriori 关联模型结果的热力相似性，也交叉证明了我们的模型和算法程序正确性良好，能够挖掘并反映 KPI 和 KQI 数据之间关联特征。 同时，KQI=>KPI 的关联程度数值，要大于 KPI=> KQI 数值，热力图颜色也要更红热一些， 数学上有 P(KQI->KPI) = P(KPI | KQI) = P(KPI，KQI) / P(KQI) 大于 P(KPI->KQI) = P(KPI，KQI) / P(KPI)，即 P(KPI)- > P(KQI) KPI 指标异常的概率要大于 KQI 异常的概率，我们可以合理推测，可能因为 KPI 设备指标的异常偏离并不一定会直接影响视频播放的用户体验，视频类业务对设备异常的反应相对迟缓，容错弹性区间较大，而 KPI 设备指标更容易检测到通信系统的异常特征等。

Apriori 关联模型不仅能计算生成单指标对单指标之间关联规则的置信度，同时也能计算多指标集合之间的关联规则的置信数据，因为多指标关联规则数量过大，29 类指标 9999 个样本数据点往往能生成数十万条符合条件的关联规则，论文里无法容纳，于是，我们在此处给出少部分多指标集合关联程度数据，如表 3-1 所示。

表 3-1 *Apriori* 多指标集合关联程度数据挖掘部分结果表

规则 *X*	规则 *Y*	置信度	规则 *X*	规则 *Y*	置信度
5 17	15 16	0.998024	19 28	5 18	0.911647
5 16	15 17	0.998024	1 28	3 5	0.897698
3 17	15 16	0.997940	3 28	1 5	0.888608
3 16	15 17	0.997940	5 26	1 18	0.831933
1 17	15 16	0.997872	1 25	5 18	0.831858
1 16	15 17	0.997872	15 23	18 21	0.831640
1 16	3 15	0.997872	15 21	3 5	0.678520
1 17	3 15	0.997872	17 23	3 18	0.658050
17 18	15 16	0.995898	18 26	20 21	0.657795
16 18	15 17	0.995898	18 26	20 23	0.657795
16 25	15 17	0.995305	21 28	3 18	0.657343
1 23	3 21	0.995011	23 28	3 18	0.657343
1 21	3 23	0.995011	16 28	19 22	0.435967

复合型多指标集之间的关联规则的分析是相对困难的，因为其存在繁琐且相互耦合的特征，从数据结果表来看，部分指标规则的关联程度较大接近于 1，即代表先导 X、后继 Y 项几乎都同时出现异常，说明当先导集合元素同时出现异常状况时，容易导致后继集合中的指标同样发生异常，表明了元素指标之间存在的联合耦合性的关联特征信息，可能是先导元素的联合作用的才能引起后继集合异常，也可能是先导元素同时异常反映通信设备出问题的可能性更大，从而也增大了后继集合受到影响而产生异常的概率大小。

3.3 本章小结

本章节针对关联规则挖掘问题基于 Apriori 算法构建了关联规则模型，计算并挖掘出了出 KPI、KQI 指标内部和两者之间的关联规则和程度，同时也讨论了导致结果的实际原因。

4 多维向量异常检测

4.1 相关工作

本章节的目的是使用多种机器学习方法，对 KPI、KQI 各个指标的样本数据的规律进行挖掘，并实现多指标高维向量数据的异常检测。我们将 KPI 一行 24 个指标、KQI 一行 5 个指标看成一个多维向量，并以其为处理对象进行建模。基于上述目的，本章节使用了主成分分析算法 + Mahalanobis 距离公式、K-Means 聚类算法等多种机器学习模型，依次对我们的 KPI、KQI 指标样本数据进行数据挖掘，并对计算结果的有效性进行分析，同时通过将多指标高维向量特征降维等方法，映射到单维数据，并配合第二章的单维异常检测模型实现了多指标阈值门限判定的目的。

4.2 算法模型及数据挖掘结果

4.2.1 主成分分析+ *Mahalanobis* 距离

主成分分析 PCA，是一种特征降维方法，其目的是将多特征的原始高维数据，通过寻找一个目标的超平面，使得原数据在超平面投影后的低维数据，能够满足(1) 投影点分布的方差是所有可能的超平面集合中数值最小的；(2) 投影点距离超平面的距离的均值也是所有可能的超平面中最小的，这两条性质。即主成分分析是计算一种超平面，在保证最小方差和最小重构两条原则的条件下，使得原始多特征高维数据投影到低维，实现特征数量的缩减，并保留具有主要作用的特征。主成分分析的算法流程有，首先输入样本数据集和输出维度 N，并将样本数据中心化$\sum x_i = 0$，然后计算样本点的协方差矩阵，并进一步对协方差矩阵进行特征值、特征向量分解，然后取最大的 N 个特征值所对应的特征向量，组合成投影矩阵 W，最后，将每一个原始样本数据点通过与投影矩阵相乘，计算其在超平面上投影后的降维坐标值，即 PCA 模型的输出数据。

Mahalanobis 距离，目的是为了去除各维度之间存在关联耦合性和量纲差异性，避免影响其距离数据的一种计算方法。该方法的实现原理是通过计算协方差矩阵来实现对维度差异性的消除。如公式 4-1 所示，其中 S 即协方差矩阵，当 S 等于单位矩阵时则退化为欧拉距离。

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1}(x-\mu)} \tag{4-1}$$

针对本章节多维度异常检测的目的，首先我们利用主成分分析算法将 KPI 共 24 个特征指标即 24 个维度降维至 2 个维度。然后，计算降维处理后的所有样本数据点的均值中心点(质心)，并利用 Mahalanobis 距离测算所有样本点到该质心的距离，而这些距离即组成一列单维数据，因此，最后我们使用第二章建立的单维度异常检测模型即可实现对多指标 KPI、KQI 数据的门限值确定和异常点判定。基于所述算法流程，我们将 KPI、KQI 样本导入，计算并绘制出了 KPI、KQI 指标在 PCA 降维处理后样本点的二维处理、三维处理散点图、Mahalanobis 距离分布直方图，如图 4-1 所示。

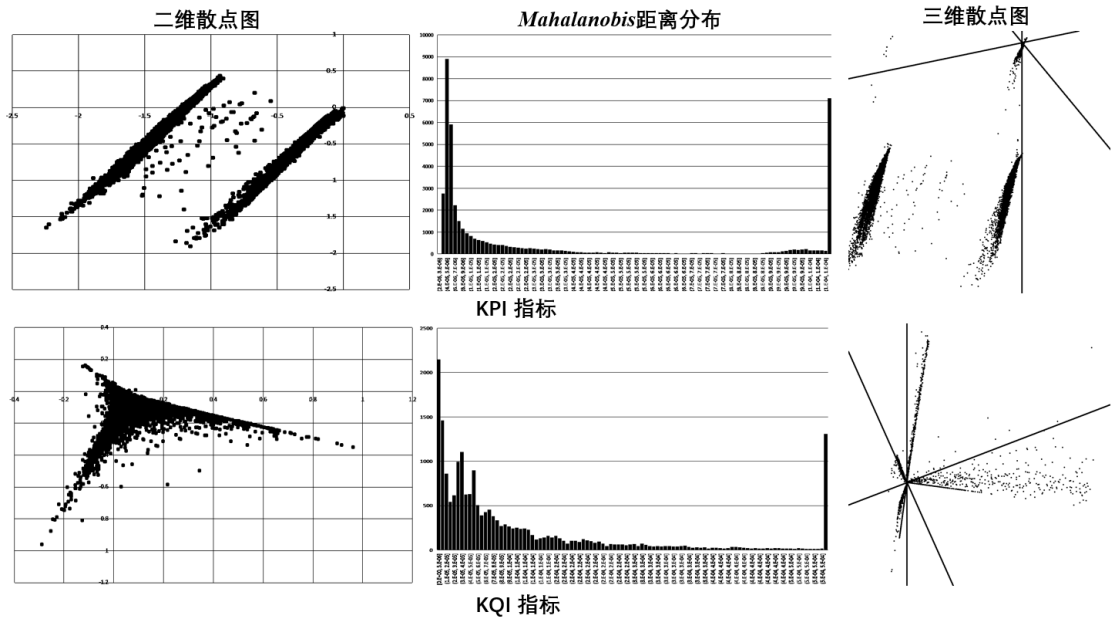


图 4-1 PCA 降维后样本点二维三维散点图、Mahalanobis 距离分布直方图

从主成分分析对指标样本数据的降维后的结果来看，即上图的二维散点图，KPI 指标的二元特征数据分布形如两条平行 1 的直线段，且有少量样本点分布在这两条线段之间，而从 PCA 三维处理的散点图中来看，样本点主要分布在两条平行线段，和一条靠近零值点的较短线段上。同时，我们可以从 PCA 降维至二元指标和三元指标的两幅散点图之间发现，二维散点平面的确是能够最大化三维散点图像上的样本点之间间距的目标平面，突出了三维图像中数据主要分布的两根平行线段的特征，而短线段则在二维投影后被融入了其中一条长线段，说明我们 PCA 主成分分析算法模型的效果和正确性是相对较好的。而对于 KQI 降维结果而言，其样本数据二维散点图呈现辐射状的三叉型形，且一条辐射臂相对较短，其中主要数据都分布在中心盘和三叉星的辐射臂上，而少量数据沿着辐射臂继续延伸或分散在主臂之间。同时，我们进一步从经 PCA 三维处理的散点图中可以看到，样本点沿四个主要辐射方向分布，其中三条辐射轴的数据相对集中，都聚集在轴线上，而第四条辐射轴的数据分布则相对更离散一些。

不过，KPI、KQI 指标二元降维的 Mahalanobis 距离直方图都相对不均匀，存在局部点的剧烈起伏，且数据沿 x 轴分布相对绵长，甚至在尾部有上升起伏的区间，因此并不适合作为阈值判决对象。

4.2.2 聚类方法

聚类算法的目的是，将原始样本数据，基于其自身在超空间的分布规律，自动地无监督地划分为目标的 K 个类别，并使得所有归类后的数据，能够保证其距离自己所在类的簇心(即所有属于该类的数据的质心或均值中心)，相比其他类的簇心都要更近。K-Means 算法是所有聚类算法中效率最高，且效果相对较好的模型。K-Means 算法模型的大致流

程，首先基于输入的样本数据集和设置的簇数 K，随机选择 K 个样本点作为初始簇心，然后开始循环迭代。在一个迭代周期内，对每个样本数据点计算其相对于每个簇心的距离，然后将该样本点纳入距离他距离最小的簇心所代表的集合中，当每一个样本点都计算并分类完成后，对于每一个新生成的簇心集合，重新计算其集合簇心，并将新的簇心数据结果取代旧的存入内存里。如果一个迭代周期里，所有旧有的簇心数据均未改变，即所有簇心都已经收敛到一个稳定的数值，则退出迭代循环，其簇心及其集合点，即是输出数据。如果存在簇心改变，则回到迭代开始，继续循环。综上所述，K-Means 算法的流程图，如图 4-2 所示。

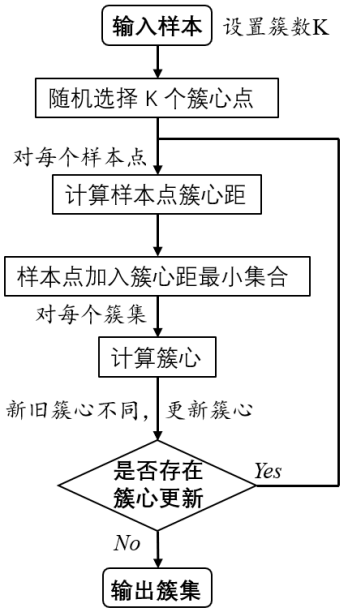


图 4-2 K-Means 算法流程图

针对本章节多维异常检测的目的，我们首先对 KPI、KQI 指标样本进行归一化使得数据均分布在 0 到 1 之间，然后利用 K-Means 聚类模型将 KPI、KQI 数据点聚类为设置的 K 个簇，计算每个样本点距离其自身所在簇的簇心的距离，作为判决该样本点是否异常的参考。同上述主成分分析方法一样，所有样本数据的距离组成一列单指标数据，用第二章单维异常检测模型即可解决。基于如上算法，我们将 KPI、KQI 指标数据代入，计算并绘制出不同 K 取值下 K-Means 聚类模型样本点到簇心距离分布直方图，如图 4-3 所示。

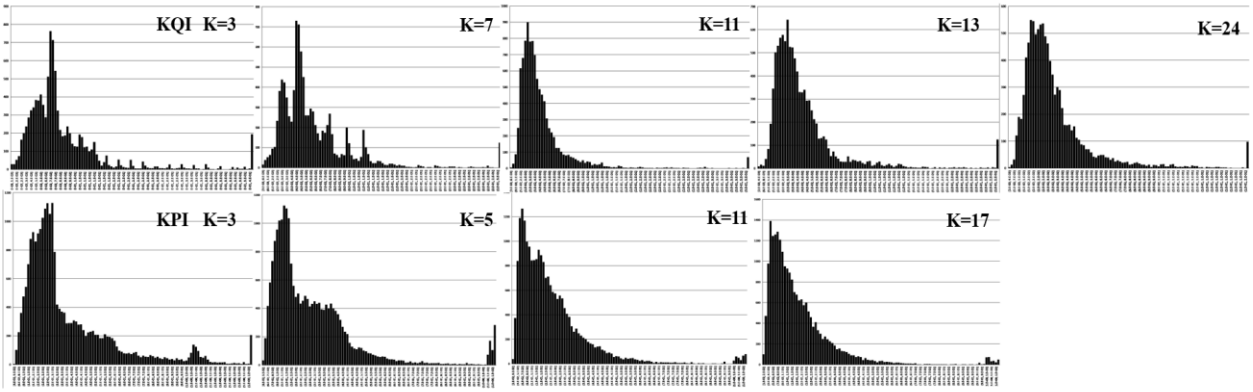


图 4-3 K-Means 聚类模型样本点到簇心距离分布直方图 (上:KQI 指标，下:KPI 指标)

从样本点到簇心距离分布直方图中可以分析，当 K 取值较低时，如 KQI 指标图像的(K=3，K=7)，分布呈明显锯齿状，且部分区间数值剧烈起伏，说明样本点的距离分布十分不均匀，我们可以推测，可能由于 K 值设置过小簇心数过少，使得部分抱团样本点的均值中心距离模型分配给他的簇心存在一定距离，而这些抱团样本点的距离数据就集体在直方图分布中呈现颗粒的锯齿状的。而诸如 KPI 指标的(K=3，K=5)，在逐渐过渡段区间内呈现阶梯状起伏，且在峰值到中间段的过程中出现较大落差，使得距离分布同样不均匀。当我们逐渐增大簇心个数的 K 的取值时，此前 KQI 指标出现的锯齿状颗粒逐渐被抚平，样本点到簇心的距离分布更均匀，曲线起伏更平滑，如 KQI 指标的(K=11，K=13)。而上述 KPI 指标出现的阶梯状和剧烈落差，随着 K 值的增加也逐渐消失，曲线相对平滑。通过上述分析可知，簇心数 K 值的适当增加，能够让 KPI、KQI 指标样本点分配到更适合的类别，从而使得所有样本点到其簇心的距离平滑分布，更能合理地反映和拟合 KPI、KQI 指标样本自身在高维空间内分布的特征。

同时，我们也将 K-Means 聚类得到的簇心点的坐标绘制出热图表格，如图 4-4 所示。簇心坐标数据热力图中显示，K 值较小时确定的簇心点，基本都被后续 K 值增加的簇心坐标结果所继承，即 K 值的增加倾向于增添更多簇心的同时尽量保留原有的簇心点距。而当 K 值继续增大到一定程度后，计算得到的部分簇心坐标之间基本相同，只在一到两个指标维度之间存在显著差异，说明此时簇心的个数已经充足甚至存在过分富余的可能。因此，通过对簇心坐标数据的分析能够帮助我们选择适当的簇心个数 K 的取值。

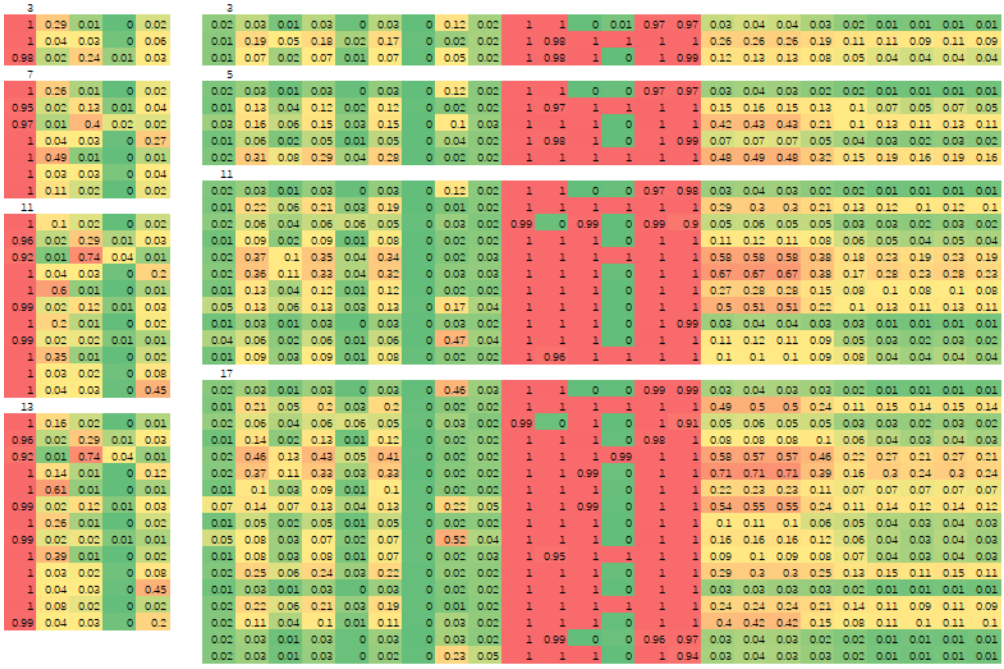


图 4-4 不同 K 值下的簇心坐标数据热力图 (左:KQI 指标，右:KPI 指标)

针对 K-Means 聚类后的样本点簇心距离分布进行阈值分析，对于 KPI 指标我们以簇心数 K=11 的结果为处理对象，可知数据的均值为 0.188、标准差 0.159、方差 0.025、Gamma 分布卡方拟合检测值为 1114.935，该数据的对数的偏度为 0.130、峰度为 2.741、偏度检测量 7.961、峰度检测量 7.912。我们进一步给出频数拟合图以帮助分析，如图 4-5 所示，该数据 Gamma 分布拟合的卡方数值过高，从直方图中同样可以看出在靠近零点峰值 Gamma 拟合曲线过低，而在中间及尾部区间拟合曲线又过高，因此卡方值显示偏差过大是合理的，该数据不适合用 Gamma 分布拟合方法。其次，该数据的对数的偏度-峰度接近于正态分布的 0-3，且偏度-峰度检验值也相对较低，从直方图来看正态曲线对对数值的拟合仍然存在缺陷，均值左侧曲线略小于频数，右侧曲线和直方图包络基本贴合，但在两层的尾部正态曲线均高于直方图。因此，对数正态拟合方法并不算太好，应当拒绝，也可根据工程实际考虑接受或拒绝。其次，对 KQI 指标我们选择 K=11 的聚类结果分析，分布拟合图 4-5 上显示，Gamma 分布和对数正态分布的拟合方法和直方图包络之间偏差太大，均不能反映其数据本身的分布，因此拒绝其拟合结果。因此，我们可以通过使用箱型图或百分比划定等方法确认 KPI、KQI 指标聚类数据的阈值门限。

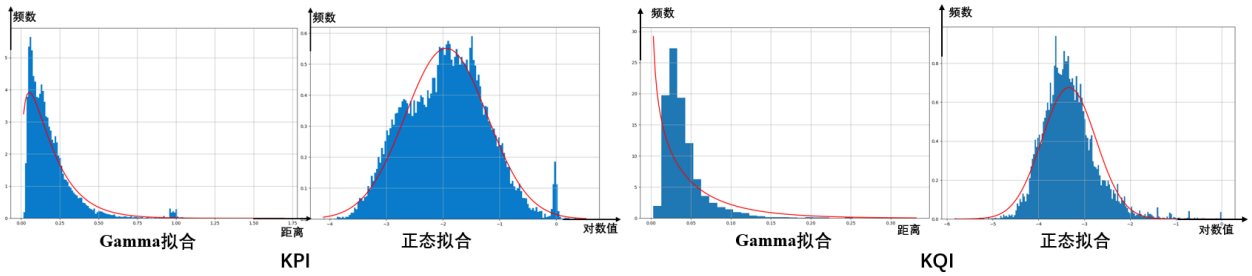


图 4-5 样本点到簇心距离分布频数拟合图 (左:KQI，右:KPI)

4.3 模型总结

主成分分析方法的优点是，可以通过特征降维使得多指标高维度的原始样本数据，按照要求地投影到二维平面和三维立体空间，让原本耦合复杂的数字能够可视化地展现其分布的具体几何结构，帮助我们更清晰的掌握 KPI、KQI 指标数据分布的规律和本质。但通过 Mahalanobis 公式计算的样本点到均值中心的距离分布效果并不好，直方图分布的不均匀性使得异常检测的结论较差。

K-Means 聚类模型的优点，在于能反映 KPI、KQI 指标数据在高维空间聚集的情况，通过将样本根据自己的分布特征分配至合适的类别簇团，并得到各个簇团的簇心坐标，帮助我们清晰地分析数据分布的特征情况。同时，样本点到簇心的距离的分布直方图，相比与比主成分分析的结果更加均匀和平滑，异常数据点的阈值门限也能更容易找到。

两种数据挖掘算法可以相互补充，相辅相成，通过在其模型的优势领域使用，帮助我们更全面完善地探索 KPI、KQI 各个指标样本数据的分布规律和本质。

5. 数据预测

5.1 相关工作

本章节的目的是处理时间序列样本数据的预测问题，分析并挖掘出数据隐含在时间的规律信息。本章节以 LSTM 长短期记忆算法作为基础建立模型。以 BDXUS0963 徐水五站基站 HLHD-3 的 KPI 指标中上行总吞吐量，作为检测该模型性能的测试时间序列数据。

本章节，首先对 LSTM 算法的正向、反向传播的流程进行了描述，并在 LSTM 算法的基础上构建了服务于本文 KPI、KQI 数据的时间序列预测模型。然后，本章节对 KPI 指标上行总吞吐量样本曲线的特征和周期性进行了描述和分析，将样本数据划分为训练数据集和测试数据集，分别代入 LSTM 模型进行迭代训练，和对每一迭代周期的网络效果进行测试评价，以使得网络参数自我学习更新并收敛到预测状态。最后，我们对 LSTM 预测模型的预测结果和原始样本时间序列曲线进行对比，评价并分析了网络预测的正确性和性能效果。

5.2 *LSTM* 长短期记忆模型

长短期记忆网络 *LSTM*，是一种特异化的循环神经网络。该模型随运行迭代次数(即时间)计算的同时，也将上一时刻模型的状态，作为当前时刻模型的输入之一。因此，在该机制的作用下，*LSTM* 模型能够保持一种长期及短期的记忆性。同时，*LSTM* 模型的内部由 *G*、遗忘门 *F*、输入门 *I*、输出门 *O* 四个并联的神经网络层作为核心，*FOI* 的激活函数均为 *Sigmoid*，而 G 的激活函数则为 *tanh*。

对于 *LSTM* 的正向传播过程，首先，*LSTM* 模型在当前时刻的输入有x_t, S_{t-1}, H_{t-1}，其中x_t和H_{t-1}以列为堆积方向进行合并，作为 4 个神经网络 *FGIO* 的输入，然后将遗忘门 *F* 的输出与上一次状态值S_{t-1}相乘，同时加上 *GI* 两个神经网络的输出的积，作为新的状态值S_t。其次，当前时刻的状态值在 *tanh* 激活函数处理后乘上输出门 *O* 的输出，作为当前时刻模型的输出y_t和H_t，最后 *LSTM* 模型输出y_t, S_t, H_t并作为下一时刻模型的输入。*LSTM* 正向传播流程示意图，如图 5-1 所示。

同时，对于 *LSTM* 的反向传播过程，首先，模型输入当前的正确标签数据 y_{0t}，然后利用设定的误差函数 *err()*，对 LSTM 之前的输出 yt 与正确答案 y0t 计算误差值，连同模型输入(t+1)时刻反向传播进行修正的ΔS_{t+1}, ΔH_{t+1}数据，作为 *LSTM* 模型反向传递的输入。其次，基于ΔS_{t+1}, ΔH_{t+1}模型先计算 S 的修正参数ΔS_n，基于Δy, ΔH_{t+1}可以算得神经网络 O 的修正参数。然后，将ΔS_n反向传播至遗忘和输入神经物理层 FGI，得到其各自的修正参数，并基于结果得到 H 的修正值ΔH_t。最后，LSTM 模型完成当前时刻的反向传播修正，并输出ΔS_t, ΔH_t继续向(*t-1*)时刻进行传播作为其输入数据。LSTM 反向传播流程示意图，如图 5-2 所示。

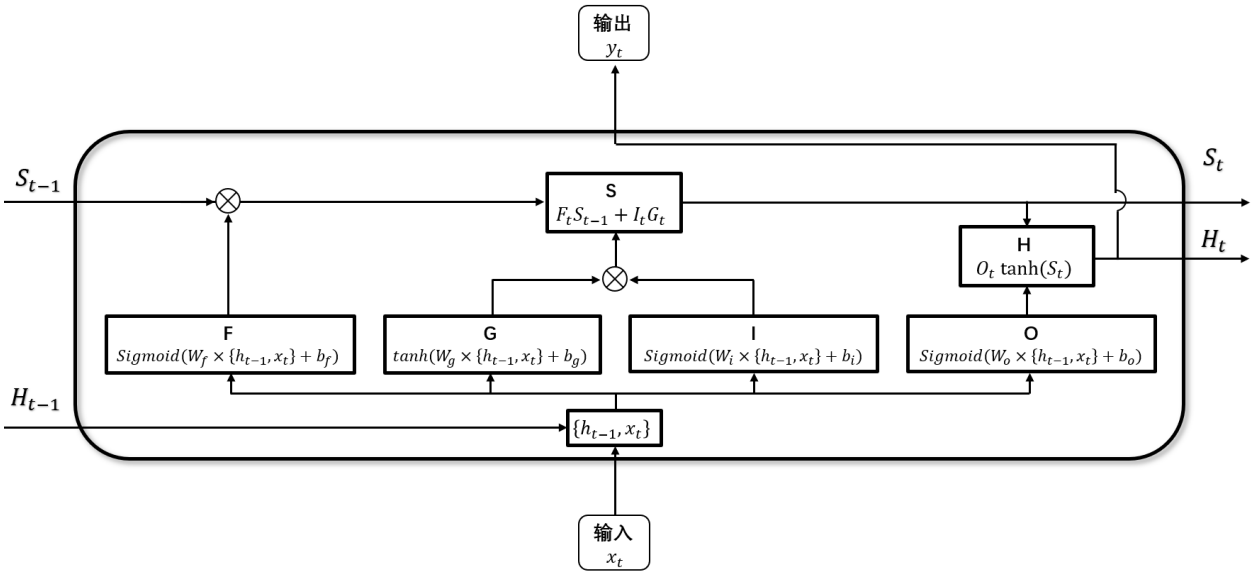


图 5-1 *LSTM* 正向传播流程示意图

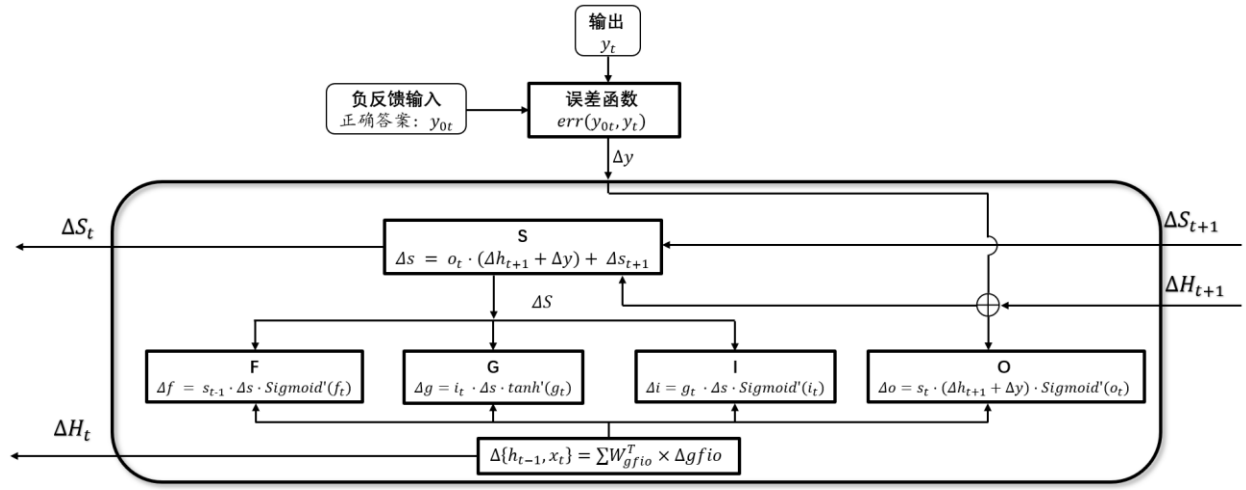


图 5-2 LSTM 反向传播流程示意图

基于 *LSTM* 的时间序列预测模型，首先，我们假设输入的 *KPI*、*KQI* 时间序列长度为 *x_Len*，而 *LSTM* 的细胞个数，即 *LSTM* 的输出序列长度为 *y_lstm_Len*，同时我们所需的模型最终序列长度为 *y_Len*。因为 *y_Len* 与 *y_lstm_Len* 并不相等，例如，模型假设 LSTM 细胞个数为 64，则输出序列长度为 64，而本模型实际所需输出序列长度为 1，故并不匹配，所以我们在 LSTM 单元的输出端，在增添一个全连接层神经网络，作为 LSTM 输出的处理，全连接层的输入输出层长度为{*y_lstm_Len, y_Len*}。流程示意图，如图 5-3 所示。

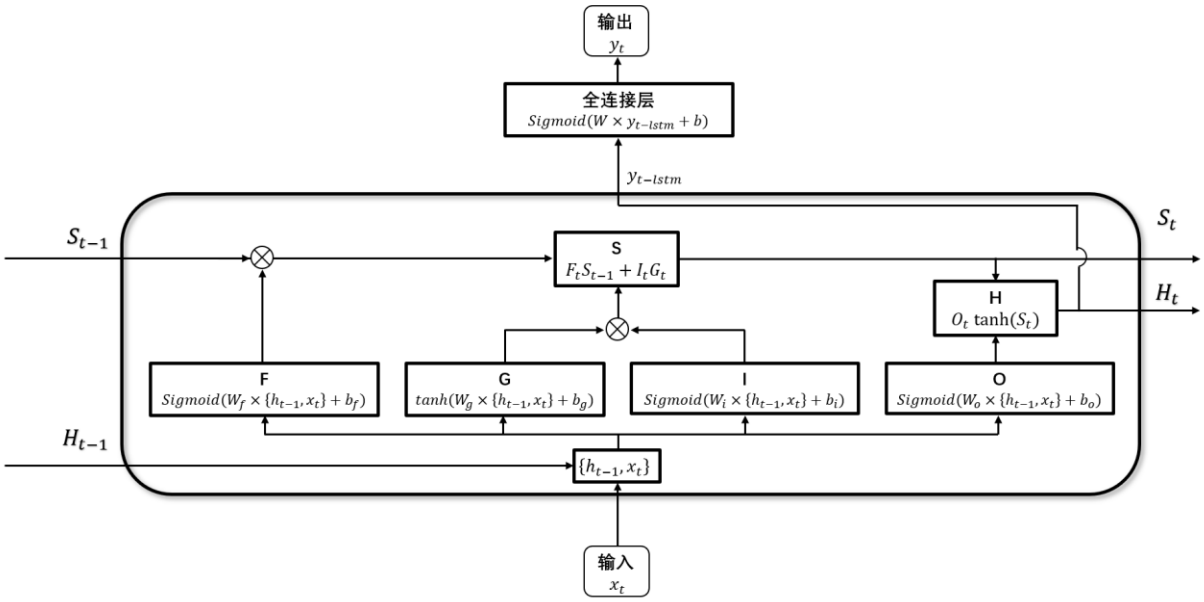


图 5-3 时间序列预测模型流程示意图

5.3 数据挖掘结果

5.3.1 原始样本时间序列

本章节以 BDXUS0963 徐水五站基站 HLHD-3 的 KPI 指标中上行总吞吐量时间序列数据为例，对本 *LSTM* 数据预测模型进行诠释和分析。该序列数据从 2018 年 7 月 23 日 0 点整开始，至 2018 年 8 月 5 日 23 点整为止，以小时为颗粒度共包含 336 个数据点。首先，绘制出徐水五站 HLHD-3 上行总吞吐量随时间分布的趋势图，如图 5-4 所示。

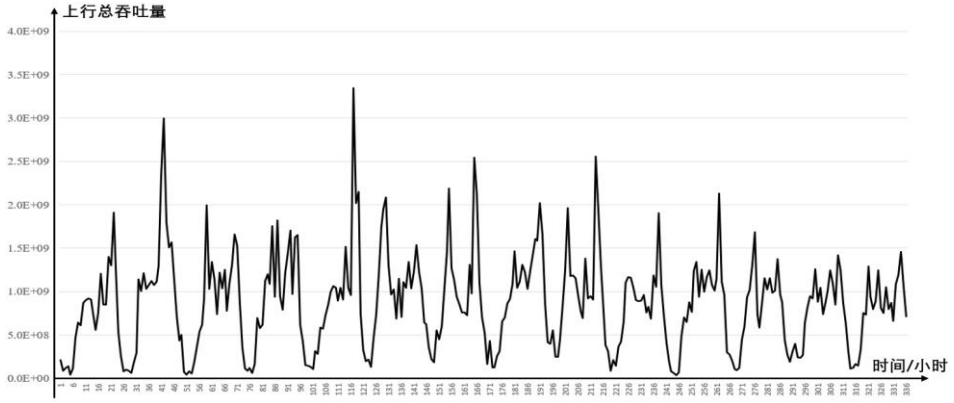


图 5-4 BDXUS0963 徐水五站 HLHD-3 之 KPI 上行总吞吐量的时间序列图

从上图中，我们可以看出，徐水五站上行总吞吐量时间序列数据根据不同时间区段，有明显的分布差异，在夜间区段从清晨 0 点开始吞吐量数据曲线快速下降趋近于零，并在清晨 0 点至 7 点处于波谷，然后在早上 7 点至 8 点快速升高至 10^9 吞吐量附近，并在早上 8:00 至晚上 23:00，上行吞吐量数据都处于通信相对活跃的范围，上述分析也符合绝大多数人的正常作息习惯。其次，徐水五站上行总吞吐量在下午 17:00 至晚上 22:00 区间内，呈现明显高于其他时间区段，并在下午 19:00 左右达到的每日的波峰极值，可以分析

认为这是下班后人们出于休闲娱乐的目的，对数据通信服务的需求大幅度增加而形成的结果，符合对现实的实际预期。故该时间序列数据，趋势正常，且存在随时间的规律信息，能够很好的作为检验我们 *LSTM* 预测模型的测试数据。

5.3.2 预测时间序列

本文对徐水五站上行总吞吐量时间序列数据，将前 250 个样本数据点作为训练集合，后第 250 至第 336 的样本数据点作为测试集合。然后，将数据进行预处理并归一化，使得其数据值分布在 0 和 1 之间以方便模型计算，最后，导入至本章节建立的 *LSTM* 时间序列预测模型进行分析和训练，计算得到并绘制出时间序列 *LSTM* 预测对比图，如图 5-5 所示。通过时间序列预测对比图，我们可以分析出，对于前面训练集合的 250 个样本点，本模型预测结果与原始数据曲线基本重合，说明随着训练次数的逐渐增加，我们的 LSTM 模型能够向原始时间序列曲线收敛，输出结果和训练集之间的损失数值能够逐渐缩小。因此，我们用 C++自己编写 LSTM 模型程序，能够有效并正确地实现训练阶段的预期目标。

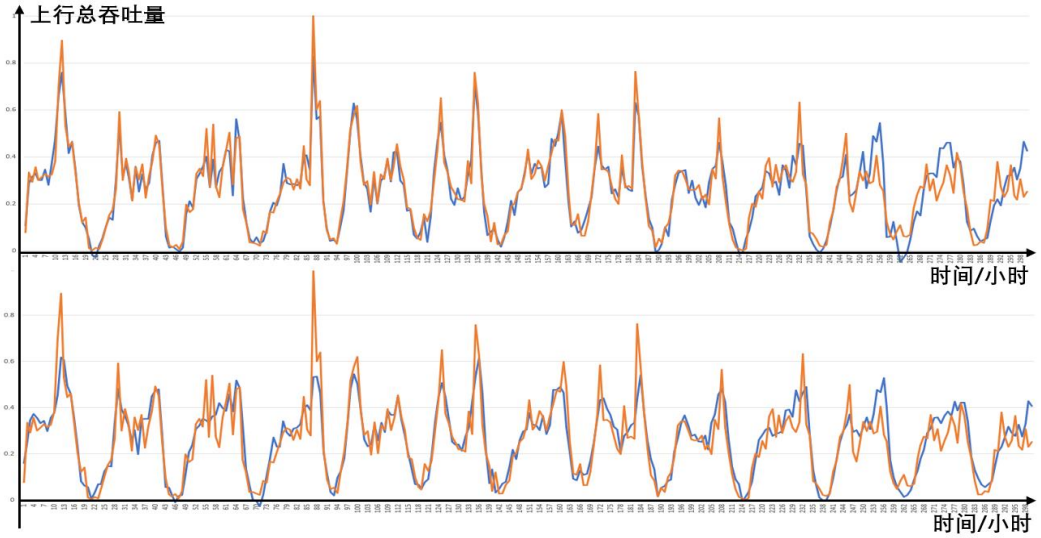


图 5-5 BDXUS0963 徐水五站 HLHD-3 之 KPI 上行总吞吐量的时间序列 LSTM 预测对比图
(红色:原始样本序列，蓝色: 预测序列，上:训练次数较多，下:训练次数较少)

对于测试集合的第 250 至 300 位的样本点，从大尺度来看，本模型预测的时间序列曲线能够很好的反映以天为周期的 KPI 上行总吞吐量的欺负变化，在清晨 0 点和上午 7 点的特殊时刻，预测曲线能够随。从以时间为颗粒的小尺度来分析，预测曲线基本和原始数据保持一致。综上所述，本文设计的 LSTM 模型，能够很好地实现对输入样本数据的未来时间的预测，能够学习并挖掘原始时间序列随时间分布的蕴含特征和演化趋势。

5.4 本章小结

本章节主要针对时间序列预测问题，基于 LSTM 算法提出了一种预测模型，以 KPI 上行吞吐量为例，训练 LSTM 网络参数收敛到理想状态，通过实验测试我们对该模型预测结果和原始时间序列曲线的异同进行了详实的分析，并评价了预测模型的性能。

6 KPI、KQI 综合关联评价模型

6.1 综合模型流程

本章节的目标是将上述第二、三、四和五章设计的几类机器学习模型进行整理、统一，构建一套针对 KPI、KQI 指标数据的综合模型，以实现实时无监督地监控和评价通信业务的运行质量和服务状态的目的。模型算法流传示意图，如图 6-1 所示。

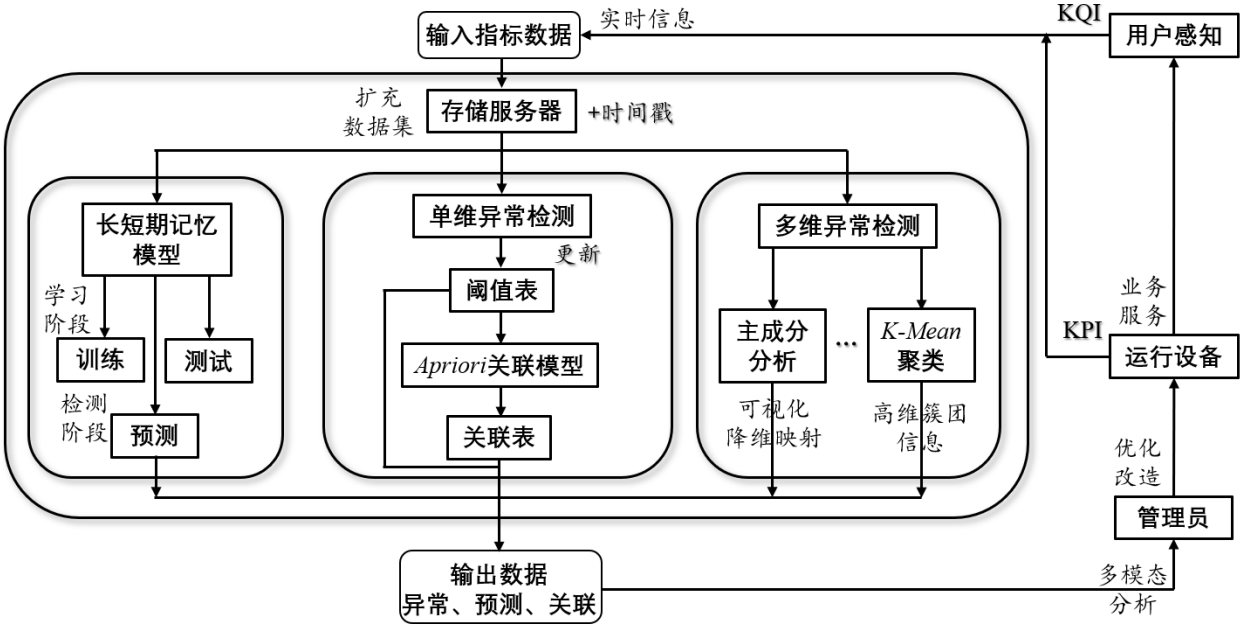


图 6-1 KPI、KQI 综合关联评价模型流程示意图

模型流程示意图显示，整套综合评价模型配合实际运营商通信服务设备和管理人员，形成了闭环负反馈且能够自我调节的信息流系统。首先，综合系统实时地采集每一时间颗粒度下 KPI、KQI 各个指标数据的测量，汇总并输入监控系统内，并标记上时间戳存储在运营商服务器里。然后，将该时刻输入数据连同过去时刻存储的数据一同输入单维异常检测-关联分析、多维异常检测和时间序列预测三个子模型中。其次，多个子模块对于每一时刻的样本输入将输出的不同类型的数据结果，能够帮助运营商管理员从不同角度多模态地对通信网络的情况进行分析，主成分分析的可视化空间显示，能更符合人日常感知习惯的分析数据在高维空间的分布结果，单维异常检测-Apriori 关联能快速判别哪些出现的 KPI 通信运行设备指标异常最可能显著影响到用户服务的感知体验，以更具准确、具有针对性地优化修复通信网络设备。而 LSTM 的预测数据输出，能帮助运营商对于数据情况进行合理预期，了解每个不同地区不同基站一天内通信数据流量随时间的趋势分布和潮汐效应，例如针对通信流量的工作高峰区段，管理人员可以为即将到来的数据量涨潮提前做好准备，也可以利用通信流量的潮汐效应更好地做好设备资源在不同时间的合理分配，提高当前通信资源能带来的最大经济效益并节省不必要的资源开支浪费。

同时，该综合模型所使用的算法基础均满足无监督的前提条件，不需要认为的对测量数据进行手动标记和处理，只需将每一时刻所测量到的 KPI、KQI 各个指标的数据信息输入至监控评价模型，该模型会自主学习输入样本数据集并自动挖掘 KPI、KQI 指标数据中有价值的特征信息输出反馈给管理人员。不仅如此，随着 KPI、KQI 样本数据资源的逐渐富集，能很好的帮助各个数据挖掘模拟迭代收敛至理想状态。随着样本数据的增多能使得样本频数直方图更加清晰，包络更平滑，单维异常检测也更容易挖掘和拟合样本分布所呈现的特征规律。其次，也能帮助不断丰富 LSTM 神经网络用于训练和测试的数据集，而神经网络有效性和准确性与数据集的大小关联紧密，于是网络能够因此收敛至更佳的状态，从而更好地预测未来的 KPI、KQI 指标信息，挖掘其在时间序列上的数据特征。

根据该综合系统的各模块并行运行的特点，我们可以在更广泛的阅读无监督数据挖掘的各类文献和算法模型和逐渐积累的实际经验后，为模型添加更有价值效率的新的子算法模型，而不影响系统原有的模块的正常运行。并且，我们也可以进一步对各个子模块的多模态的数据挖掘结果之间进行关联分析，优化简化挖掘结果的数据规模从而帮助管理人员更容易地理解数据信息。最后，本课题已经为图 6-1 的综合模型中所有用到的子模块算法编写并调试好了与之对应的 C/C++程序代码，可以方便地部署在实际的运营商通信业务服务器中或辅助并尽可能满足管理人员的分析需求。

6.2 本章小结

本章节将第二至五章的各数据挖掘模型整理、统一，构建一套针对 KPI、KQI 的综合模型，从而达到了实时且无监督地监控、评价通信运行质量和服务状态的目的，并分析了该综合模型的性能及特点，完成课题预期的研究目标。

7. 总结与展望

7.1 总结

本文实现了通信网络设备运行质量与用户对服务的感知和满意度评价之间的关联，针对 KPI 与 KQI 指标之间的相互映射问题构建了监控、检测系统模型并给出了详细的解决方案，能够帮助通信运营商更准确完善具有针对性的升级优化自己服务设备，以提高用户对通信业务服务的实际体验。

本文通过分别设计了单维异常检测-关联分析、多维异常检测和时间序列预测等多个子模型，实现了构建监控、评价通信业务的运行质量和服务状态的综合模型的研究预期目标，有条不紊地实现了课题的每一项计划和要求。本文主要成果如下：

(1) 基于指数分布 Gamma 分布、对数正态分布为代表的经典概率分布、极大似然估计、卡方假设检验、偏度-峰度检测方法等统计学算法，以及 Apriori 关联规则算法，完成对 KPI、KQI 样本数据本质的分布规律的数据挖掘工作，找到了部分指标的经典分布函数，计算并给出了 KPI、 KQI 指标内部和两者之间的关联程规则和程度。

(2) 利用主成分分析和 K-Means 聚类等多种机器学习算法，可视化地绘制出了数据在二维、三维的特征降维投影，得到了数据在高维空间聚集的簇心点坐标等多种有价值的特征信息。

(3) 基于 LSTM 算法构建了时间序列预测模型，且模型能够拟合原始样本并反映通信数据流量在一天中存在的潮汐效应。

(4) 整理并总结了一套监控、评价通信业务的运行质量和服务状态的综合模型，完成了对 KQI、KPI 的关联数据挖掘任务。

7.2 不足与展望

本文模型可以改进的地方有:

(1) 可以增加第二章单维异常检测模型的经典随机变量分布的分析案例，以更全面地挖掘拟合 KPI、KQI 的样本分布。

(2) 可以用其他关联规则算法来挖掘 KPI-KQI 之间的映射关联，同本文第三章的 Apriori 模型的结果与性能进行比较。

(3) 可以测试更多的机器学习算法对 KPI、KQI 的分布特征进行挖掘，并优化第四章的多指标异常检测模型。

本文作者才疏学浅，如有纰漏，欢迎斧正。

参考文献

[1] 赵刚. 彩铃系统中于 KPI 和 KQI 的性能分析系统的设计与实现[D]. 北京: 北京邮电大学, 2009.

[2] 倪萍. 流数据挖掘关键技术研究[D] . 北京: 北京邮电大学, 2010.

[3] Goldstein M , Dengel A . Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm[C]. KI-2012: Poster and Demo Track. 2012.

[4] Shyu, M.L., Chen, S.C., Sarinnapakorn, K. and Chang, L., 2003. A novel anomaly detection scheme based on principal component classifier. [C]. IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM'03). IEEE, 2003.

[5] 关健, 刘大昕. 基于主成分分析的无监督异常检测[J]. 计算机研究与发展, 2004(09):1474-1480.

[6] Portnoy L., Eskin E., Stolfo S. Intrusion Detection with Unlabeled Data Using Clustering[J]. acm workshop on data mining applied, 2001.

[7] 杨斌. 基于聚类的异常检测技术的研究[D]. 长沙: 中南大学, 2008.

[8] Malhotra P, Vig L, Shroff G, et al. Long Short Term Memory Networks for Anomaly Detection in Time Series[C]. 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2015: 89-94

[9] S. Hochreiter and J. Schmidhuber. Long short-term memory[J]. Neural computation, 9(8):1735–1780, 1997.

[10] Guochao Song, Wei Wang, Da Chen, and Tao Jiang. KPI/KQI-Driven Coordinated Multipoint in 5G: Measurements, Field Trials, and Technical Solutions[J]. IEEE Wireless Communications. October 2018: 23-29

[11] Hartigan J A, Wong M A. A K-Means Clustering Algorithm[J]. Applied Statistics, 1979, 28(1):100-108

[12] Shlens J. A Tutorial on Principal Component Analysis[J]. International Journal of Remote Sensing, 2014, 51(2).

[13] Song G, Wang W, Chen D, et al. KPI/KQI-driven coordinated multipoint in 5G: Measurements, field trials, and technical solutions[J]. IEEE Wireless Communications, 2018, 25(5): 23-29.

[14] 范荣明，胡博. 基于大数据的客户感知 KQI 与 KPI 关联研究[J]. 2017-2019 年“学术金秋”获奖论文集. 2020: 200-208

[15] 魏远伦. TD-LTE 网络中 KQI 指标优化的研究[J]. 通讯世界. 2017(08): 6-7

[16] 李钟瑞，戴明珠. 移动用户业务感知时延类 KQI 优化[J]. 软件导刊. 2019,18(02): 169-

173
[17] 刘冰. 基于用户感知的网络优化项目质量管理研究[D]. 北京：北京邮电大学，2012.
[18] 陈德金，郑成林，吴智恺. LTE 移动用户感知模型研究[J]. 电信技术. 2015(3): 14-16.
[19] 伏玉笋. 移动通信网络评价准则与解决方案[J]. 电信科学. 2020,36(11): 28-38
[20] 程卫华, 何肖嵘. 基于大数据的网页浏览质差分析方法研究[J]. 电信科学. 2020,36(11): 174-181
[21] 杨磊. 基于 FPgrowth 机器学习的影响用户感知无线根因问题的快速定位方法研究[J]. 江苏通信. 2019,35(02): 56-62

附录

Algorithm 1 常见分布: 分布函数、密度函数程序

```
inline double igamma_low(double x, double s, int N = 200) {
	double ans = 0;
	for (int i = 0; i < N; i++) ans += pow(s, i) / tgamma(x + i + 1);
	return ans * tgamma(x) * pow(s, x) * exp(-s);
}
inline double PoissonDistrib(int x, double mean) {
	double ans = 0;
	for (int i = 0; i < x; i++) ans += pow(mean, i) / NumberTheory::Factorial(i);
	return ans * exp(-mean);
}
inline double NormalDensity(double x, double mean = 0, double var = 1) {
	return 1 / sqrt(2 * PI * var) * exp(-pow(x - mean, 2) / (2 * var));
}
inline double NormalDistrib(double x, double mean = 0, double var = 1) {
	return 1.0 / 2 * (1 + erf((x - mean) / sqrt(2 * var)));
}
inline double ExpDensity	(double x, double mean) {
	return x <= 0 ? 0.0 : 1.0 / mean * exp(-x / mean);
}
inline double ExpDistrib	(double x, double mean) {
	return x <= 0 ? 0.0 : 1 - exp(-x / mean);
}
inline double GammaDensity(double x, double mean, double var) {
	double a = mean * mean / var,
		b = mean / var;
	return pow(b, a) / tgamma(a) * pow(x, a - 1) * exp(-b * x);
}
inline double GammaDistrib  (double x, double mean, double var) {
	double a = mean * mean / var,
		b = mean / var;
	return 1.0 / tgamma(a) * igamma_low(a, b * x);
}
```

致谢

华电四年光阴，弥足珍贵，这篇本科毕业论文，便是我这四年里所获能力的缩影与见证。

罗翔老师曾说，我们所得，并非全是自己努力，更多的是环境和他人给予了我们舞台和提点。很多人都很聪明，但却没有踏入道路的机会。我当感恩给予我机会的他者。

大一电路的葛玉敏老师，给予了我很大尊重，葛老师自己设计创新的教学模式，是真心想要教育好她的学生。在这门课上，我也总坐第一排努力参与同老师的交流互动，因为我高中通过物理竞赛有电路知识的基础，且课堂考试成绩较好，因此葛老师给我了大学唯一一个满分成绩，给予了我大学继续刻苦学习的自信。　　电信系的张卫华老师，给予我在算法竞赛的舞台，在算法编程整体实力相对薄弱的电信系，张老师和电信系仍然为学生申请诸如“蓝桥杯”比赛经费和资格，在给予的舞台上我也不负众望，拿过省奖晋升过国家级决赛，在 C++的算法编程能力上我不输专业的计算机系学生，但在这条漫长的学习道路上我仍要也仍会继续走下去。　　班主任张珂老师，在我的本科学习给予很大帮助，大二时组织大创课题，将我引领至计算机前沿的“神经网络”和人工智能领域，在这项目中我粗浅地了解了神经网络的历史、原理、数学基础和实现方法，并利用 Pytorch 框架逐个实现了主流的神经网络架构，在大四也自己用 C++重新底层实现了一遍。而该大创项目我也不负班主任期望拿下了国家级优秀的成绩，证明了自己的实力。

华电四年，还有很多很多给予我帮助的他者，尽心教育鼓励我的各位老师，电子设计俱乐部的学长学姐，带领我们电赛的郭以贺、姚国珍老师，数学建模竞赛同舟共济的各位队友，启蒙我算法编程能力的柳婼学姐。我华电四年，遇各位三生有幸，此生铭刻。

生于世，求索真理与永恒，在人类的智慧之海，和自然的本质谜题中，畅游理解前人的思想体系，明白自己身处的这个自然的本质，到底是什么。我生于自然，见过自然，理解自然，归于自然。足以。

愿这一生，玩得开心，尽我责任。