

杭州电子科技大学

硕 士 学 位 论 文

题 目：大型科研仪器机时智能计算的系统设计与实现

研 究 生 刘 剑

专 业 电子与通信工程

指导教师 陈科明 副教授

企业导师 曾 佳

完成日期 2022 年 5 月

杭州电子科技大学硕士学位论文

大型科研仪器机时智能计算系统的设计与实现

研 究 生：刘剑

指导教师：陈科明 副教授

2022 年 5 月

**Dissertation Submitted to Hangzhou Dianzi University
for the Degree of Master**

**Design and implementation of time intelligent
computing system for large scientific research
instruments**

Candidate: Liu Jian

Supervisor: Associate Prof. Chen Keming

May, 2022

杭州电子科技大学

学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。

申请学位论文与资料若有不实之处，本人承担一切相关责任。

论文作者签名：刘剑

日期：2022 年 5 月 31 日

学位论文使用授权说明

本人完全了解杭州电子科技大学关于保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属杭州电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署单位名称仍然为杭州电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密论文在解密后遵守此规定）

论文作者签名：刘剑

日期：2022 年 5 月 31 日

指导教师签名：陈明

日期：2022 年 5 月 31 日

摘 要

机时（关机机时、待机机时、工作机时）可以有效反映出仪器使用率、闲置情况等信息，在大型科研仪器开放共享与绩效考核中扮演着重要的角色。然而，目前各单位、机构仪器的机时数据还停留在人工统计阶段，不仅繁琐低效且存在漏记、错记等问题，无法满足当下构建规范化、专业化、网络化的大型科研仪管理体系的需要。

针对人工统计机时的方式不适用于当下大型科研仪器种类、数量众多的场景，本文设计并实现了一种大型科研仪器机时智能计算系统。基于 NB-IoT 的终端对仪器电流数据进行采集；每日电流数据上报后，调用后端服务进行机时自动计算；自动计算各仪器机时的流程为，首先使用易于操作的分箱与改进 K-means 算法构建各仪器的电流分类模型，之后使用模型对仪器每日上报的电流数据进行分类，统计分类结果得到机时。然后，针对其中机时误差大的仪器，再使用成本高但适用性更强的 DP-SVM 算法进行重新建模，从而确保机时计算结果的准确率；电流、机时等数据通过 Web 端进行展示，最终解决了人工统计机时的痛点，有效地提升了仪器管理的智能化水平。

机时计算的前提在于实现电流分类，本文分析了电流阈值法及其缺点后，提出了一种基于分箱与改进 K-means 的机时计算方法。首先对目标仪器历史电流分箱处理，选用均值和标准差构建样本，然后通过改进 K-means 算法聚类生成电流分类模型并存至云端，最后调用云端模型对其上报电流进行分类，统计各类总数得到机时。对多种仪器进行测试，结果表明本方法优于现有的电流阈值法，所测样本中 81.9% 的仪器机时基本符合实际。此方法无需标定数据、训练模型，可由系统自动执行，进而降低人工成本与实施难度。

在实际使用中发现，由于部分仪器的电流较为复杂，使用基于分箱与改进 K-means 算法得到的机时与实际机时间存在较大偏差。针对这类少量特殊的仪器，本文设计了一种基于动态规划与支持向量机（DP-SVM）的机时计算方法进行重新建模处理。在此方法的设计中，先是介绍了各仪器 SVM 电流分类模型的建立流程，然后针对线上电流分类前需要对电流序列进行切分，使用固定窗口导致效果不佳的问题，提出了一种基于 DP 的信号切分算法来提升电流分类及机时计算的准确性。测试结果表明，DP-SVM 可以用于机时的自动计算，并且效果要比人工设置固定窗口处理的方法更好，多台仪器机时平均相对误差不超过 5%，提高了系统的适用性。

关键词：物联网、大型科研仪器、电流分类模型、机时计算

ABSTRACT

Instrument-time (time of shutdown, time of standby, time of working) can effectively reflect the utilization rate and idle of instruments, and plays an important role in the open sharing and performance evaluation of large-scale scientific research instruments. However, at present, the instrument-time of various institutions still stay in the stage of manual calculation and summary, which is not only cumbersome and inefficient, but also has problems such as omission and wrong recording, which can not meet the needs of building a standardized, professional and networked management service system of large-scale scientific research instruments.

In view of the fact that the method of manual statistical instrument-time is not suitable for the scene with a large number of types of large-scale scientific research instruments. This thesis designs and implements a instrument-time intelligent computing system for large-scale scientific research instruments: Using the sensor which based on NB-IoT collect the current data of instruments; After daily data is reported, the backend service is invoked to automatically calculate instrument-time. The system realizes the automatic calculation of the instrument-time of each instrument. Firstly, the current classification model of each instrument is constructed by using the easy to operate box division and the improved k-means algorithm. Then, the model is used to classify the current data reported by the instrument every day, and the instrument-time is obtained by counting the classification results. Then, for the instruments with large instrument-time error, the DP-SVM algorithm with high cost but stronger applicability is used to re model, so as to ensure the accuracy of instrument-time calculation results; The current, instrument-time and other data are displayed through the web, which finally solves the pain point of manual statistical instrument-time and effectively improves the intelligent level of instrument management.

The premise of instrument-time calculation is to realize current classification. After analyzing the current threshold method and its disadvantages, this thesis designs a instrument-time calculation method based on box division and improved k-means. First the base note of target instrument is processed by the historical current box algorithm, and choose mean and standard as the features of samples and use improved K-means algorithm to get classification model. The reported data can be classified by the model of the cloud, and then the instrument-time can be calculated by counting the total number of every types. The test results show that this method can be used for automatic calculation of instrument-time, and is better than the method of using clustering to solve the threshold. 81.9% of the measured samples are basically in line with the reality. This method does not need calibration data and training model, and can be implemented automatically by the system, which can

effectively reduce the labor cost and implementation difficulty of the system.

In practical use, it is found that due to the complex current of some instruments, there is a large deviation between the instrument-time obtained by using box splitting and improved K-means algorithm and the actual instrument-time. For this kind of small number of special instruments need manual intervention and remodeling, this thesis designs a instrument-time calculation method with dynamic programming and support vector machine (DP-SVM). Firstly, introduce the establishment process of svm current classification model of each instrument. Then, aiming at the problem that the current sequence needs to be segmented before online current classification, and the algorithm effect is poor due to the use of fixed window, a signal segmentation algorithm based on DP is proposed to improve the accuracy of current classification and instrument-time calculation. The test results show that DP-SVM can not only realize the automatic calculation of instrument-time, but also has better effect than the method of manually setting a fixed window. The average relative error of instrument-time measured by multiple instruments is no more than 5%, which improves the applicability of the system.

Keywords: Internet of things, Large-scale scientific research instruments, Current classification model, Instrument time calculation

目 录

摘 要.....	I
ABSTRACT	II
目 录.....	IV
第 1 章 绪论.....	1
1.1 课题来源.....	1
1.2 研究背景及意义.....	1
1.3 国内外研究现状.....	2
1.3.1 机时计算研究现状.....	3
1.3.2 用电设备状态识别研究现状.....	3
1.3.3 信号分类研究现状.....	4
1.4 论文内容与结构安排.....	5
第 2 章 大型科研仪器机时智能计算系统.....	7
2.1 系统总体设计.....	7
2.1.1 需求分析.....	7
2.1.2 系统总体框架.....	7
2.2 采集层设计.....	8
2.2.1 硬件设计.....	9
2.2.2 软件设计.....	10
2.2.3 通信协议设计.....	13
2.3 设备接入平台设计.....	14
2.4 应用服务平台设计.....	15
2.4.1 数据存储中心.....	15
2.4.2 系统业务平台.....	18
2.4.2.1 服务开发框架.....	18
2.4.2.2 核心业务设计.....	18
2.4.2.3 业务平台架构.....	22
2.4.3 Web 端设计.....	23
2.5 整体效果展示.....	24
2.5.1 采集终端展示.....	24
2.5.2 AEP 云平台展示.....	25

2.5.3 Web 端展示	26
2.6 本章小结.....	29
第 3 章 基于分箱与改进 K-means 的机时计算	30
3.1 仪器电流数据特点.....	31
3.2 电流数据处理.....	31
3.2.1 数据分箱.....	32
3.2.2 特征提取.....	34
3.3 K-means 算法分析与改进	34
3.3.1 K-means 算法原理	34
3.3.2 K-means 常见缺陷	37
3.3.3 K-means 算法改进	37
3.3.3.1 初始聚类中心选取策略.....	38
3.3.3.2 最优聚类结果求解策略.....	38
3.4 基于分箱与改进 K-means 的机时计算流程	42
3.5 实验测试.....	43
3.5.1 实验准备.....	43
3.5.2 聚类结果与分析.....	45
3.5.3 电流分类对比及分析.....	49
3.5.4 机时计算测试.....	52
3.6 实际应用分析及错误处理.....	54
3.7 本章小结.....	56
第 4 章 基于动态规划与支持向量机的机时计算.....	57
4.1 引言.....	57
4.2 基于 SVM 的电流分类模型	57
4.2.1 支持向量机理论.....	57
4.2.2 核函数及参数选择.....	59
4.3 基于 DP 的信号切分算法.....	60
4.3.1 动态规划理论.....	60
4.3.2 基于 DP 的信号切分算法设计.....	60
4.3.3 基于 DP 的信号切分算法测试.....	62
4.3.3.1 测试准备.....	62
4.3.3.2 测试结果.....	64
4.4 基于 DP-SVM 的机时计算流程.....	66
4.5 实验测试.....	67

4.5.1 实验准备.....	67
4.5.2 测试结果与分析.....	69
4.5.3 电流分类对比及分析.....	72
4.5.4 机时计算测试.....	74
4.6 本章小结.....	76
第 5 章 总结与展望.....	77
5.1 总结.....	77
5.2 展望.....	78
致谢.....	79
参考文献.....	80
附 录.....	84

第 1 章 绪论

1.1 课题来源

本课题来源于浙江省科学技术厅的委托项目—浙江省大仪设备管理服务协作平台。该项目的目标是：对全省近万台大型科研仪器进行管理，动态更新与汇聚仪器设备的基本信息、收集包括电流、运行状态、机时在内的现场数据，协助推进大型仪器设备开放共享工作，加快构建规范化、专业化、网络化的大仪管理体系。

1.2 研究背景及意义

人类文明进程的推进与科技的发展息息相关，科技除了给社会生产和人民生活方式带来巨大的影响，同时也成为了当下经济发展的决定性因素及衡量国家综合国力强弱的重要标志。大型科研仪器作为突破核心关键技术、突破科学前沿的重要物质基础，在当今科技发展中扮演着重要的角色。随着我国大量资源投入，大型科研仪器与科研设施规模不断增长，覆盖的领域、种类不断拓展，创造的效益日益显现，当下大型科研仪器大类划分如图 1.1 所示。

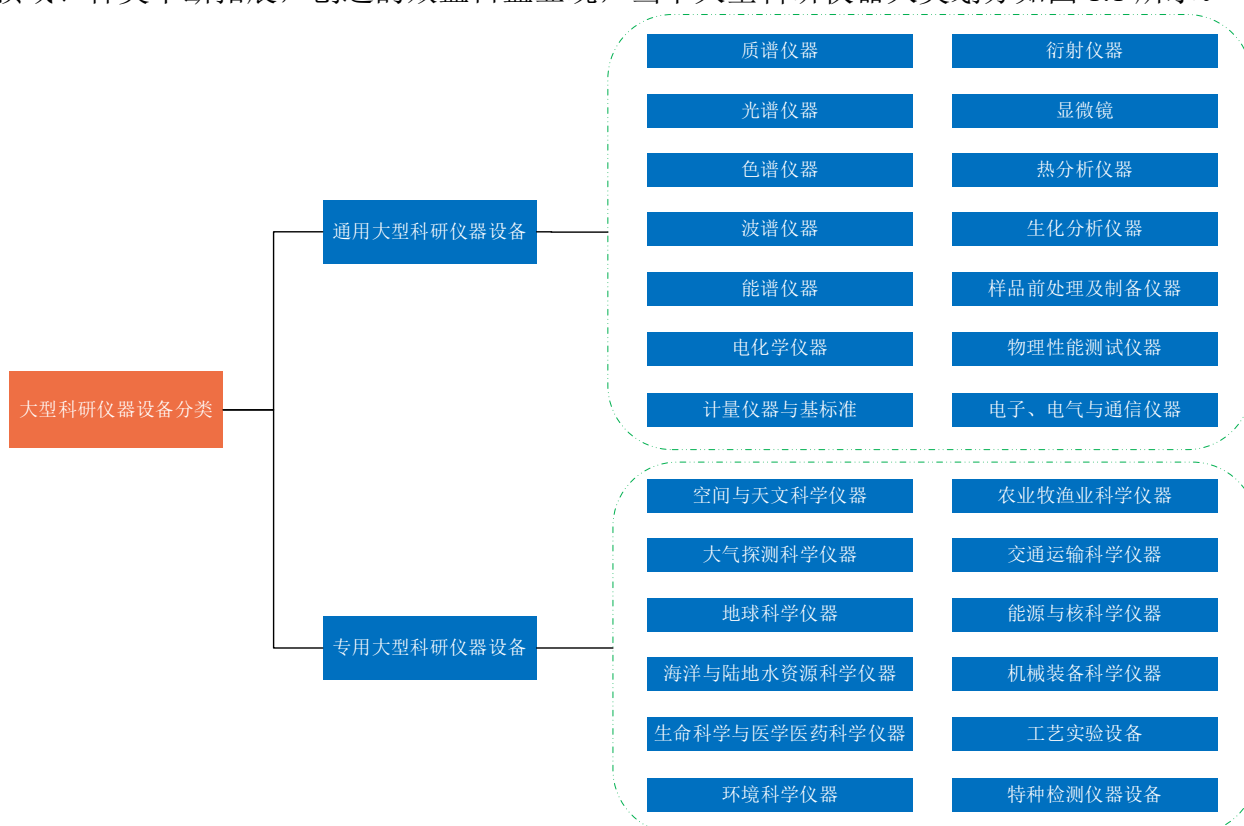


图 1.1 大型科研仪器大类划分

据统计，2014 年底我国 2200 余家高校和科研院价值 50 万元及以上大型科研仪器共计

61251 台（套），合值 868.8 亿元，价值 500 万元以上的大型科研仪器数量达到 1438 台（套），较 2008 年增长了 1.51 倍^[1]。2018 年，我国 50 万元及以上的大型科研仪器在册数共 10 万余台（套），总值 1500 多亿元，仪器规模已经位居世界前列^[2]。从全国范围来看，华东区域占有全国大型科研仪器总量的 33%，华北区域占 31%，中南区域占 14%^[3]，如图 1.2 所示。

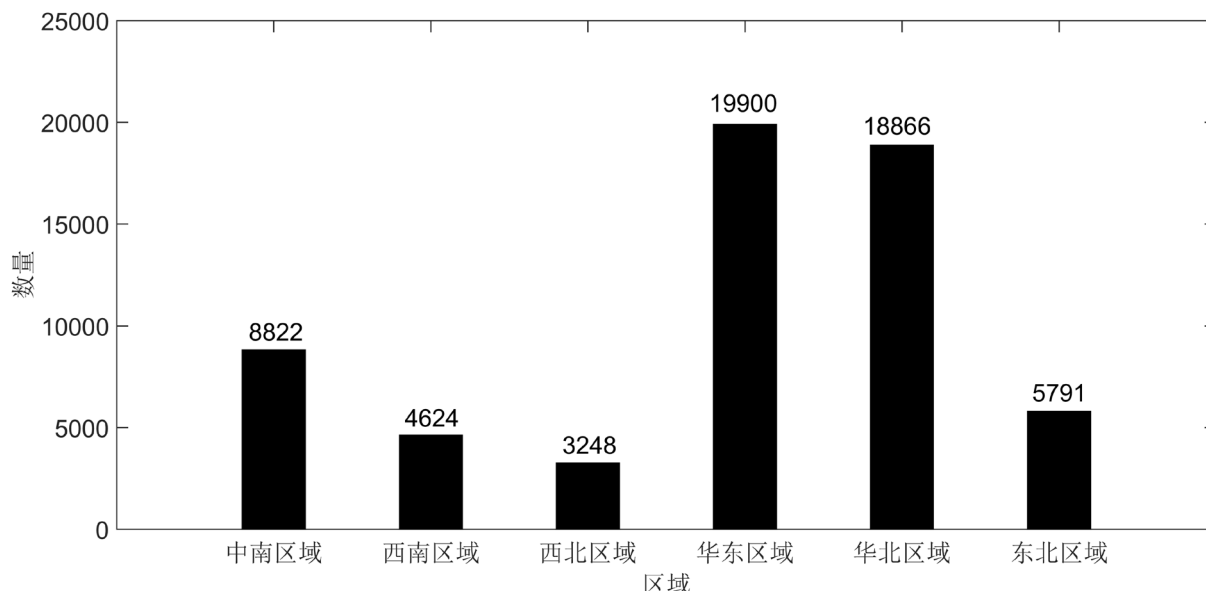


图 1.2 我国大型科研仪器分布情况

虽然我国仪器规模已位居世界前列，但仪器管理水平相较于国外还存在一定的差距，存在着许多问题^[4-6]。比如：大型科研仪器重复购买，且存在部门化、单位化、个人化的倾向，闲置浪费问题比较严重；仪器使用情况不够明确，大部分仪器的使用情况还停留在纸质记录阶段，存在漏记、错记、记录丢失等情况；仪器的科研潜力没有被充分挖掘，对科技的服务与支撑作用没有充分发挥，仪器管理制度落实不到位导致仪器丢失、损坏仪器得不到及时维修等。

为了解决上述问题，国务院在《关于国家重大科研基础设施和大型科研仪器向社会开放意见》中指出，力争在三年内基本建成覆盖各类仪科研设施与仪器的完善统一的管理服务体系，将价值在 50 万元以上的各类科研仪器纳入管理，按照功能实行分类开放共享，挖掘现有科研仪器潜力^[7]。机时作为仪器管理中的重要数据之一，不仅可以反映出大型科研仪器使用率、闲置情况等信息，也是当下仪器共享、绩效考核的重要参考指标。然而，目前机时统计还依赖于人工，存在许多问题。因此，本文基于物联网与机器学习等相关技术，结合对当前场景的实际考量，设计并实现了一种智能的大型科研仪器机时计算系统。该系统通过对大型科研仪器的电流进行采集分析，不仅可以实现对仪器运行状态的识别，而且可以自动计算、统计、汇总各接入仪器的机时数据。同时，该系统提供了一个规范化的平台，打破了各机构仪器信息壁垒，所有接入仪器的相关数据都进行了可视化展示，并可以通过外部接口调用，为管理者提供决策参考与支持。

1.3 国内外研究现状

1.3.1 机时计算研究现状

目前, 仪器机时的统计方式还较为传统, 大多数高校实验室、机构单位还仅停留在人工填写仪器使用情况表阶段。这种方式较为粗略, 不仅会存在漏记、误记、记录丢失等问题, 而且想要了仪器在较长时间内的使用情况, 还需要人工对这些文件记录进行查阅归纳, 十分繁琐。虽然部分大型科研仪器在出厂时配备运行情况记录软件, 但该软件记录内容仅限于该仪器具体模块运行起止时间, 也没有直观的机时计算结果, 收集、统计还是依赖于人工。在一些较为先进的仪器管理系统中, 虽然没有专门对仪器机时进行统计, 但仪器使用前后需要刷卡, 要分析仪器使用情况时可以参考刷卡记录。这种方式虽然实现了数字化, 但还是不够直观, 而且仪器的待机时长、工作时长并不明确。

总体而言, 目前计算仪器机时这一工作还很依赖于人工, 传统纸质记录的方式存在较多问题, 且现有的仪器管理领域中也忽略了对这方面的关注。

1.3.2 用电设备状态识别研究现状

虽然目前对于如何实现机时自动化计算的直接研究较少, 但从本质上来说, 机时是仪器在各运行状态下持续时长的统计, 只要能够准确识别出仪器不同的运行状态, 机时自动化计算就成为了可能。随着数据分析、机器学习等技术的发展, 国内外学者在用电设备状态识别领域取得了许多的成果。Yu-Hsiu Lin^[8]以电器的周波电流的峰值、平均值、均方根作为特征, 通过 BP-ANN 建立了电器开关识别模型, 从而实现电器运行状态的判断。Liang^[9]等人通过提取电流、有功功率、谐波分量等多个特征参量, 设计了一种基于多特征的电器状态在线检测算法; Lin^[10]等人提出了基于改进的时频分析的非侵入式电器状态识别方法, 通过短时傅里叶变换提取特征方案与多维背包算法完成了对家中各电器的状态识别; 王晓玲等人^[11]将仪器分为两类, 对于电流较为“简单”的仪器, 可以通过设置阈值加上一定的误差识别状态, 由于操作简单, 这也是实际中应用最广泛的方法。而对于电流“复杂”仪器, 则通过 SVM 模型对仪器各状态下的电流信号进行分类来实现状态的识别。刘克冰^[12]通过交叉小波变换降低了放电信号中的噪声的影响后生成局部放电时域图像, 然后通过改进的 FasterR-CNN 模型提取图像特征, 采用逻辑回归和分类器完成局部放电状态的检测, 最终实现了较高精度的电力设备状态识别。邢晓莹^[13]通过 AR 模型提取设备正常工作的电流特征后, 将其存入云端, 再对实时采集到的电流波形做同样解, 通过比对云端特征与实时特征的欧式距离, 来判断仪器当前处于故障状态还是正常工作状态。燕续峰^[14]定义了家用电器的暂态与稳态, 然后分析了电器状态的功率成分, 针对隐马尔可夫模型的弊端提出了多种状态转移矩阵, 然后分别为 8 种家用电器生成了电器运行状态模型, 在实际测试中得到了较好的效果。

电流信号在用电设备的状态识别中扮演着重要的角色, 大型科研仪器同样属于用电设备一类, 也可以尝试对其电流进行分析。但目前对于电力设备状态识别的研究方向主要是针对

故障检测这一领域，即使有运行状态检测也还停留在专用设备或家用电器等较小的应用范围内，对于实际工程中应用还缺乏有效的解决方案。

1.3.3 信号分类研究现状

从研究方法论的角度出发，仪器状态分类与识别本质上是通过对仪器不同状态下信号进行分类，常使用的方法包括聚类算法和分类算法。其中聚类算法的核心思想为自定义的相似度，然后根据相似度将对象进行划分^[15, 16]。相较于分类算法而言，聚类算法属于无监督的领域，无需标注数据、无需事先训练模型，操作简单。发展至今，聚类算法大致有 5 类^[17]，分别是基于划分、基于层次、基于密度、基于网格和基于模型的聚类。

基于划分的聚类通过划分将数据集分为多个簇，同属于一个簇的样本属于同一类^[18]，代表算法有围绕中心（Partitioning Around Medoid, PAM）、K-means、模糊均值（Fuzzy C-means, FCM）算法等^[19-22]；基于层次的聚类按照执行方式分为凝聚或分裂两种，代表算法有鲁棒链接算法（Robust Clustering Using Links, ROCK）等^[23, 24]；基于密度的聚类关注样本空间密度，密度越高的区域越有可能形成类，代表算法有 DBSCAN、OPTICS 等^[25, 26]；基于模型的聚类需要首先假设样本符合一种或多种分布模型，然后在聚类过程中去搜索与该模型匹配的样本，代表算法有 EM 算法、高斯混合模型（Gaussian Mixture Models, GMM）等^[27-29]；基于网格的聚类算法将空间划分成网格，以网格为单位聚类。CLIQUE（Clustering In Quest, CLIQUE）、信息统计网格聚类（Statistical Information Grid, STING）都属于该类算法^[30, 31]。

分类算法的核心思想是根据已知样本来建立映射规则，进而推测未知样本类型。因此，分类算法需要人工标注样本，分类的结果也肯定是人工设置的标签之一。这类算法的优势在于，一个训练良好的分类器具备一定的泛化能力，面对未知的数据时也能具备较高的准确性。劣势在于收集数据、标注数据成本较高，对于预置类别以外的新类无法识别。目前较为经典的分类算法有决策树^[32]、朴素贝叶斯^[33]（Naive Bayes, NB）、支持向量机^[34]（Support Vector Machine, SVM）、神经网络^[35]等。

聚类、分类算法应用在信号分类相关领域后，出现了许多成果。李诗雨^[36]对光纤振动信号进行预处理后，选用短时平均能量、短时平均幅度、峰均比等时频参量作为样本特征，然后通过 K-means 与 DBSCAN 算法实现了对不同事件下振动信号的区分。王文亮^[37]通过对管道泄露声波信号进行小波去噪后进行区间划分，然后以每个子区间的峰值绝对值为样本特征采用基于凝聚的层次聚类算法聚类，区分出了正常声波子信号与异常声波子信号，实现了在一帧声波信号中异常子信号的检测与提取。王江月^[38]对脉象信号样本进行降维处理，根据类内聚合性与类间分离性选择了四种脉象信号特征作为样本，再通过 KNN 和 SVM 算法实现了脉象信号的基本分类。Kevric 等人^[39]提出了一种算法，通过主成分分析与小波包分解对脑电图（Electroencephalogram, EEG）信号进行处理后，结合 KNN 实现了运动想象分类，并且在 BCI 竞赛 III 的公开数据集 Iva 上各个任务取得了平均 92.8% 的准确率。

总体而言，信号分类分为两个步骤：信号的特征处理和算法选择。其中，特征选择上除了均值、标准差等时域特征以外，还可以通过小波、傅里叶变换等方式提取频域信息。在对算法进行选择时，需要结合实际场景进行考虑，没有最好的算法，只有最合适的数据。

1.4 论文内容与结构安排

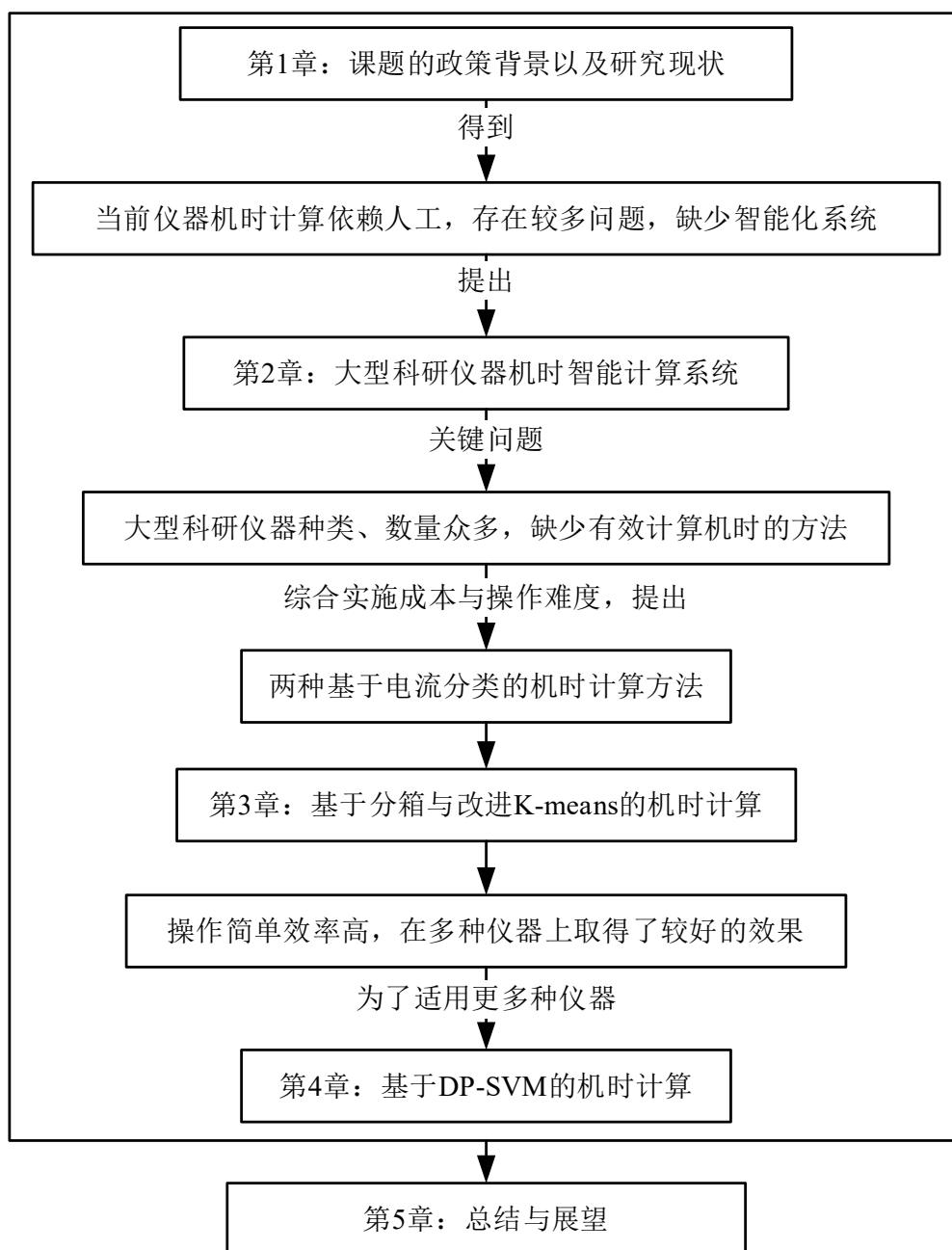


图 1.3 全文写作思路

图 1.3 展示了全文的写作思路及流程。为了协助浙江省大型科研仪器管理平台的搭建，以及为仪器开放共享策略的实施、新常态下仪器绩效考核提供机时数据支撑，本文提出了一种智能的大型科研仪器机时计算系统。该系统基于物联网架构，采集终端采集仪器的电流数据后，通过 NB-IoT 模块上报至电信 AEP 平台，然后调用相关模块、服务对其进行处理分析，

并在 Web 对相关数据进行了可视化展示；针对传统机时计算依赖于人工，无法满足当下大型科研仪器的种类与数量繁多的场景，在本文在对众多大型科研仪器电流分析的基础之上，综合了实施难度与成本，提出了基于分箱与改进 K-means 以及基于 DP-SVM 的两种机时方法来实现机时计算的自动化。

全文共分为 5 章，结构安排如下：

第 1 章：绪论。阐述了本课题的背景及研究意义，然后对当前各机构大型科研仪器机时统计的现状以及实现机时计算自动化这一关键内容的国内外相关技术进行了介绍，引出了本文的研究内容。

第 2 章：大型科研仪器机时智能计算系统。首先分析了系统需求，提出了系统总体框架，在该框架中系统被分为采集层、网络层、设备接入平台、应用服务平台，然后依次对其中关键部分的技术选型、方案制定进行了详细的介绍，最后对系统实现效果进行了展示。

第 3 章：基于分箱与改进 K-means 的机时计算。为了实现多种仪器机时的自动计算，本章提出了一种操作简单、无需标定数据、训练模型的机时计算方法。首先，分析了多种仪器的电流数据特点后，提出了一种分箱算法用于电流预处理。然后介绍了 K-means 算法的原理并对其进行改进后用于构建电流分类模型。最后，介绍了基于分箱与改进 K-means 的机时计算方法的详细流程，并选取了多种不同的仪器进行了测试，测试结果以及实际应用结果均表明了方法有效性。

第 4 章：基于动态规划与支持向量机的机时计算。针对基于分箱与改进 K-means 无法处理的仪器，提出了一种基于 DP-SVM 的机时计算方法。通过 SVM 算法训练出各仪器的电流分类模型，针对在线使用时固定窗口切分的电流序列导致分类结果较差的问题，提出了一种 DP 算法提高了切分的精度从而保证机时计算更加准确。最后构建了基于 DP-SVM 的机时计算模型，并在多种仪器的测试结果中表明了模型的有效性，可以提升系统的适用性。

第 5 章：总结与展望。总结了本文为实现大型科研仪器机时智能计算所做的主要工作，然后对所设计的大型科研仪器机时智能计算系统的不足之处进行了分析，并对后续系统建设的方向提出了展望。

第2章 大型科研仪器机时智能计算系统

当前仪器机时计算还处在人工统计阶段，这种方式不仅十分繁琐且容易出错。因此，本章设计了一种大型科研仪器机时智能计算系统来满足当下需要。

2.1 系统总体设计

2.1.1 需求分析

当前大型科研仪器机时统计主要存在这两个问题：绝大多数仪器的机时数据是由人工统计得出的，执行效率低而且仪器使用起始需要人工处于现场；统计得到的机时数据没有统一的录入平台，各机构的机时信息相互隔离，管理者无法进行全面把控。因此，对大型科研仪器机时智能计算系统的主要需求有以下几点：

1、实现自动化机时计算，利用数据分析、机器学习等相关技术来替代人工获取各仪器的机时数据，除了工作机时以外，还要计算仪器的关机机时、待机机时。机时数据要准确，不仅要知道仪器在某一段时间处于什么状态，还要知道各状态具体持续时长。

2、采集仪器相关数据时不能破坏仪器原有结构，系统的设计要考虑到灵活性和拓展性。系统对外暴露出接口的接口要设计规范，为用户的定制化需求提供相关数据。系统要做到高可用，能够支撑大量的仪器接入以及用户的请求。

3、机时计算涉及模型以及相关参数可以调整，操作方式做到简单明了，调整模型后要提供历史机时重新计算功能。数据实现可视化，满足用户查看的需求。

2.1.2 系统总体框架

基于当前场景考虑，本系统在传统物联网模型^[40]的基础上，将系统划分为采集层、网络层、设备接入平台、应用服务平台四个部分。其中，应用服务平台再细化为数据存储中心、系统业务平台、Web 端，具体的：

1、采集层：负责采集仪器的电流数据以及上报。

2、网络层：负责将现场数据传输到云平台。

3、设备接入平台：管理海量的采集终端，负责将采集层上报的数据流进行接收与存储，以接口的形式与外部服务进行交互。

4、数据存储中心：系统的数据支撑，负责定义数据的存储结构与方式，为上层提供数据来源以及执行增删改查相关的操作。

5、系统业务平台：系统的核心，负责对系统整体逻辑进行实现，主要功能包括数据持久

化、电流数据分析、模型生成、机时计算等。

6、Web 端：提供人机交互的功能，负责将仪器的电流波形、机时数据等相关内容进行可视化展示。

系统总体框图如图 2.1。

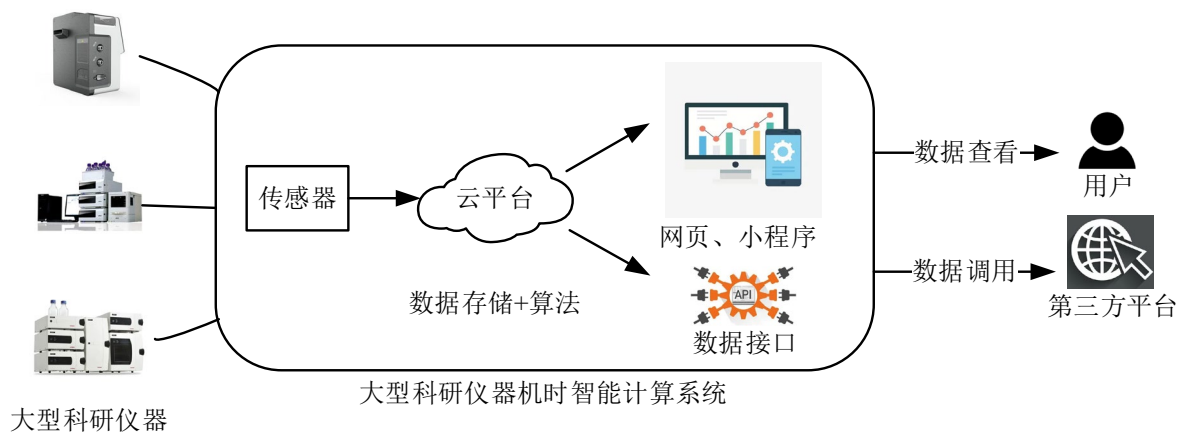


图 2.1 系统总体框图

系统各层次结构如图 2.2 所示。

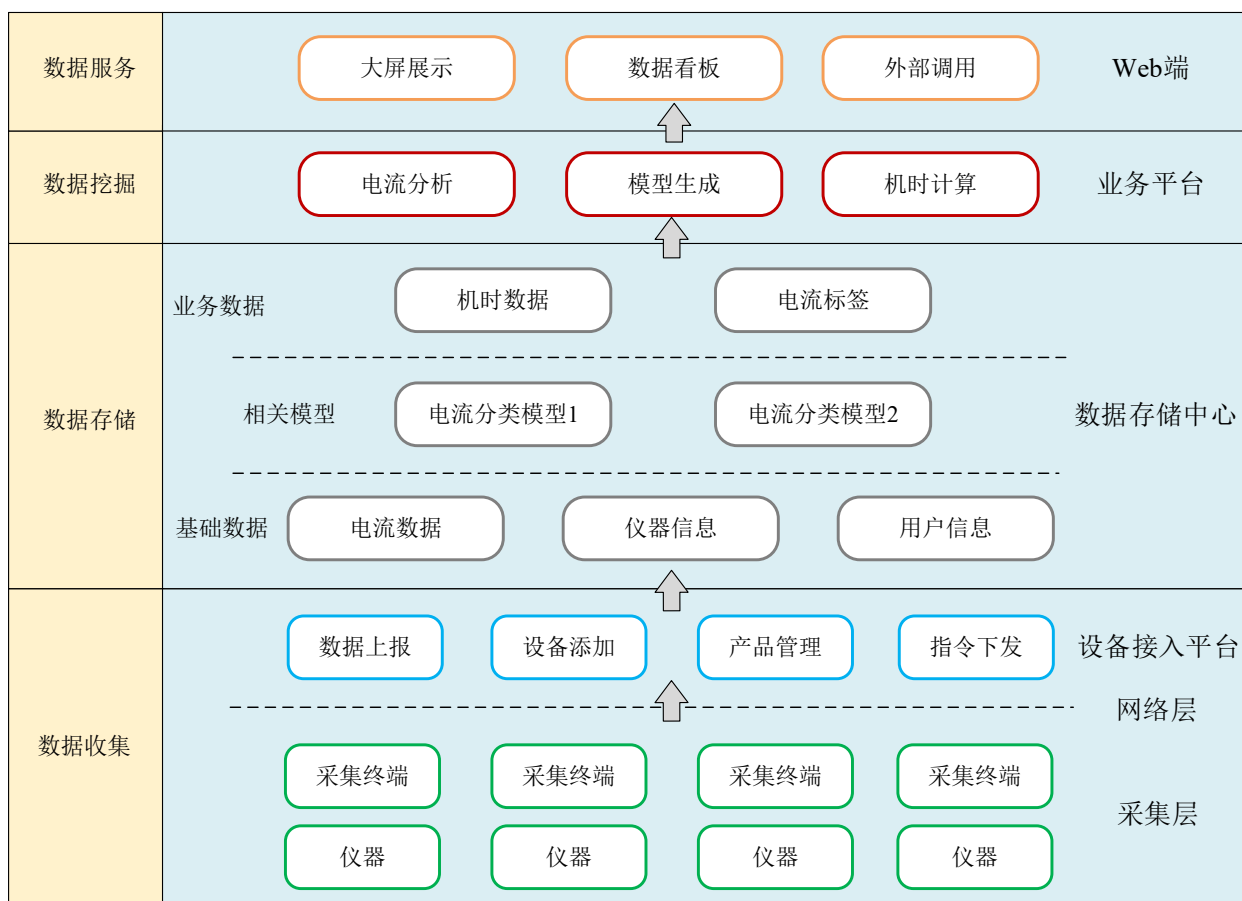


图 2.2 系统层次结构

2.2 采集层设计

采集层为整个系统提供数据原料。在设计时，要充分结合现场环境和实际业务需求，确保数据来源通畅，为后续功能实现夯实基础。

2.2.1 硬件设计

采集终端硬件框图如图 2.3 所示。

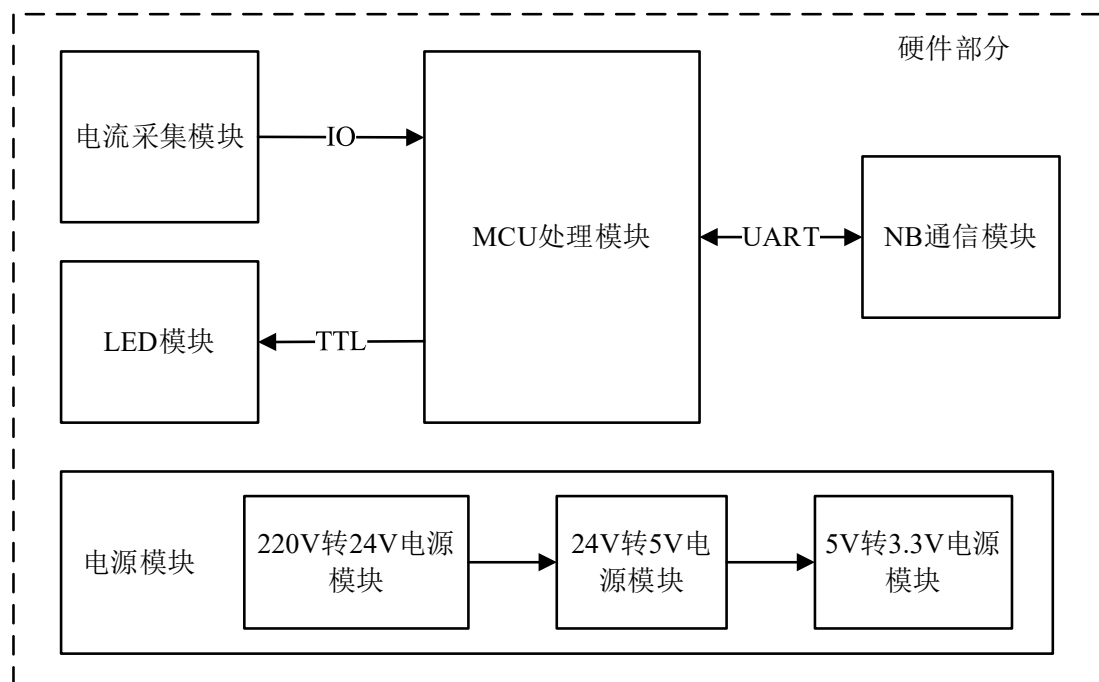


图 2.3 硬件设计框图

采集终端的 MCU 处理模块选用 STM32F103C8T6 微处理器，该处理器基于 ARM Cortex-M 内核架构、拥有最高 72MHz 的系统时钟以及 64KB 的 FLASH 与 20KB 的 RAM，计算速度快，存储空间大，可以很好的满足数据处理、任务调度的需要。

NB-IoT 是一种基于 E-UTRAN 的通信技术，因其具有广覆盖、低功耗、低成本等特性，常用于智慧小区、抄表、智慧城市等物联网应用场景^[41-43]。相比于 Lora，NB-IoT 无需自己组网，国内三大运营商已经搭建好了通信网络，入网方便，频段公开。相比于 Wi-Fi、ZigBee，当前场景更多是面向长距离通信，NB-IoT 不需要考虑网络覆盖等问题，还能满足海量设备连接的需求。除上述优点外，NB-IoT 在国内已经有了许多大规模应用的实例，技术方案相对成熟，故此处选用 NB 作为终端的入网方式。在 NB 模组的选型上，本设计选用移柯通信的 NB-IoT BCL620 模组，该模组采用 LCC 封装，体积小，性能稳定，尺寸上满足采集设备的需求，在 3GPP 下 R13/R14 标准下涵盖全频段，内部集成 TCP、UDP、LwM2M 等通信协议，能够满足构建稳定通信网络的需要。

图 2.4 展示了 NB-IoT 的基本网络架构，其中包含了 NB 模组、NB 核心网络、NB 接入平台以及应用服务器。

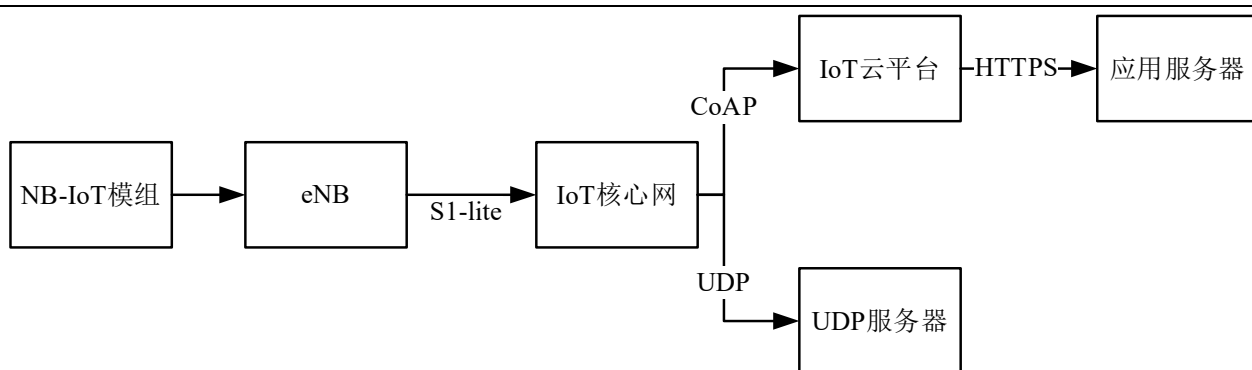


图 2.4 NB-IoT 网络基本架构

与工厂采集或小区采集的场景不同，面对大型科仪器电流采集场景，除了基本的采集功能以外，对采集模块还需要提出以下要求：

1、非侵入式，采集终端不能破坏仪器原有结构。

2、强兼容性，仪器种类众多，电流、电压等运行参数跨度较大，存在单相、三相供电的两类仪器，需要满足对各类仪器电流采集的任务。

3、安全性，采集模块在工作时不能对仪器正常运行产生干扰，更不能损坏仪器。

按照上述要求，采集模块选用电流互感器进行实现。电流互感器基于电磁感应原理，在采集仪器电流数据时无需破坏仪器电源线，就可以将仪器电源线端的大电流信号按照一定倍数转换成输出端的小电流信号，满足非侵入式采集的需求。由于采集的小电流信号存在一定的噪声，需经 LC、RC 滤波后再由 V9260S 芯片存储至内部寄存器供 MCU 读取。V9260S 电流检测芯片可以计量多种用电参量，误差精度小于 0.1%，可以满足系统的需要。

2.2.2 软件设计

采集终端的软件部分主要是主控模块、数据处理模块、网络通信模块这三部分，采集终端软件功能如图 2.5 所示。



图 2.5 采集终端软件功能图

主控模块负责对数据采集模块、网络通信模块的调度，是整个软件的“主干”，决定了系统整体执行流程。终端通电后，主控模块首先对网络通信模块进行初始化，执行入网测试。当 NB-IoT 模块入网成功后，即开启数据采集任务，当采集数据已经符合上报条件以后，按照既定通信协议对上报数据进行处理封装，通过串口发送至网络通信模块进行数据上报。上报

流程不会干扰数据的采集，主控模块调度流程如图 2.6 所示。

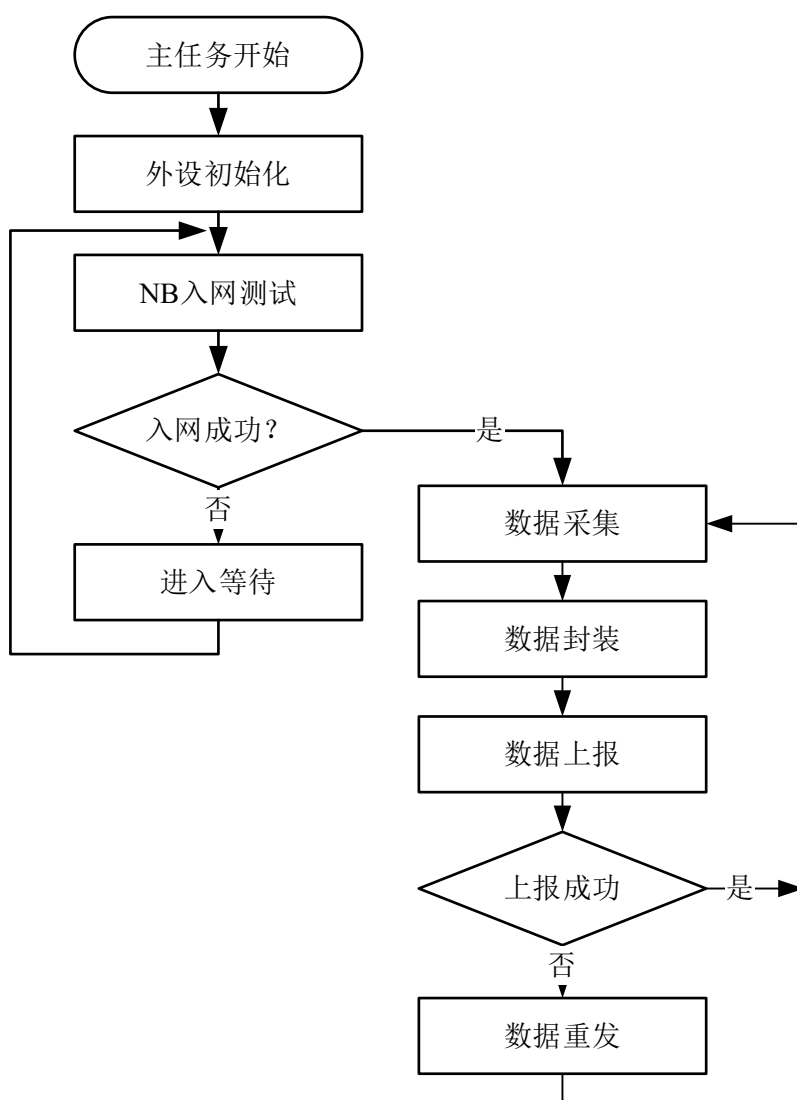


图 2.6 主任务调度流程

数据采集任务开始后，MCU 以一定频率去读取 V9620S 芯片内部寄存器中的电流数据，再存入本地缓存中。采集模块在采集电流数据时，采集频率越高，信号损失越小，越符合真实形状，但高频率读取也就意味着整体数据量上升，在进行数据上报时，会对网络和云平台造成更大的压力。降低读取频率虽然可以减少数据量，对网络和云平台的压力更小，但数据中蕴含的信息将损失更多，更难发现内在的规律。读取频率过高或者过低都不好，经综合考虑，本设计决定将采集频率设置为每分钟 1 次。上报数据时，考虑到数据上报流量消耗及云平台的负荷，规定当缓存中存在 30 个数据后，就对数据进行封装上报，上报成功后再清空缓存。

网络通信模块通过串口与 MCU 通信，负责对电流数据封装上报，是采集终端对外沟通的桥梁。网络通信模块初始化后，MCU 首先向串口发送 AT 指令进行入网测试，测试内容主要包括握手检测、SIM 卡检测、网络注册状态检测等。NB 获得入网授权后，将通过基站连接云平台并发送自身唯一标识 ID，云平台认证通过后，就可以正常的进行数据的收发了，流程

如图 2.7 所示。

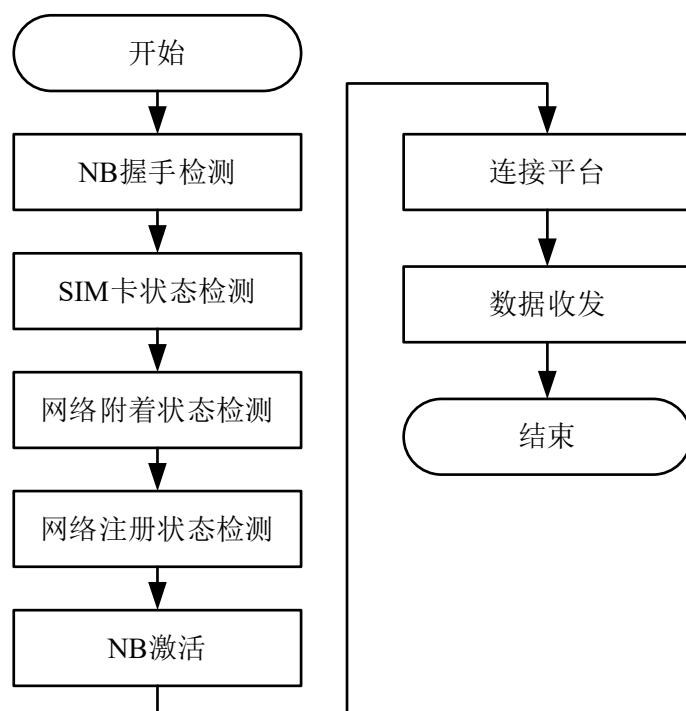


图 2.7 通信模块初始化流程

为了确保在弱网的情况下传输的数据尽量不丢失，需要一套重发机制，该机制执行流程如图 2.8 所示。

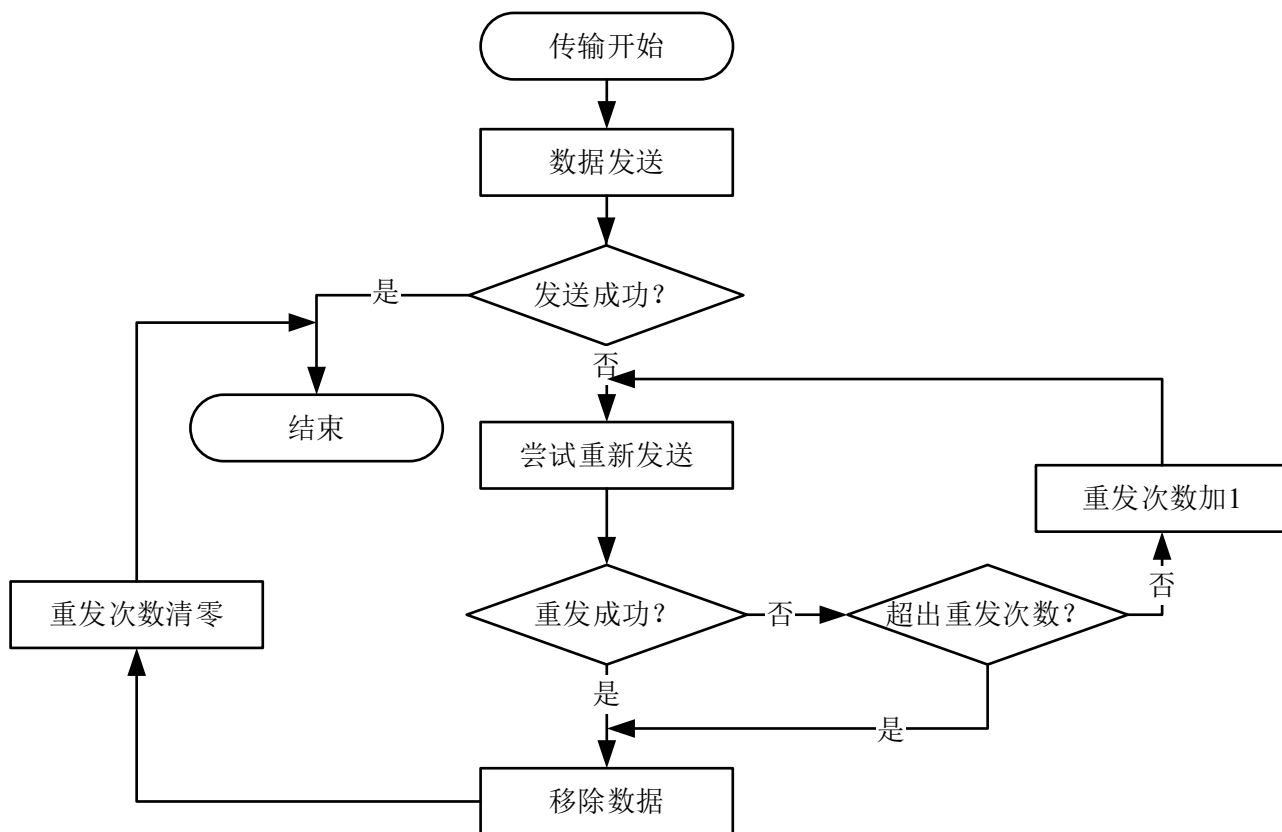


图 2.8 消息重发流程

当数据首次发送失败时会进入重发。在重发机制中，当数据发送超时或者失败时，会在 10 秒后尝试继续发送，直至数据重发失败超过 3 次、或者发送成功这 2 种情况下才将该数据删除。

2.2.3 通信协议设计

NB 模块入网后，选用轻量级物联网协议（Lightweight Machine to Machine, LwM2M）与云端通信。该协议由开发移动联盟（Open Mobile Alliance, OMA）提出，底层采用 UDP/SMS 协议传输，在物联网中具有广泛的应用^[44]。LwM2M 中所有信息都以资源的方式来表示，资源属于对象。每种对象都可以有多个实例，每个实例可以有多个资源，其内部结构如图 2.9 所示。

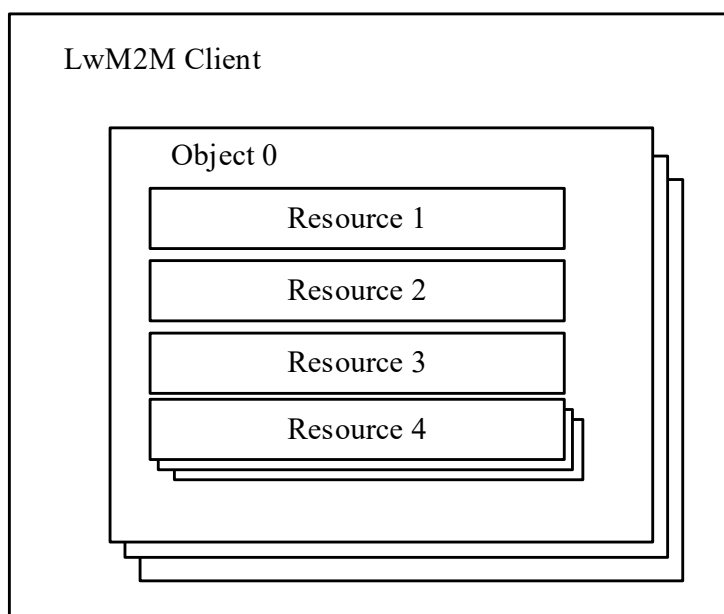


图 2.9 LwM2M 资源结构

采集终端上报事件主要有两种类型，分别是终端联网初始化上报以及正常数据上报。按照 LwM2M 协议的设计，终端上报的两种数据主要内容详情如表 2.1 和表 2.2 所示。

表 2.1 初始化数据主要内容

参数项	类型	内容示例
1	功能码	INIT_ID
2	设备唯一识别码	869858036950666
3	当前信号强度	15

终端通电后首先进行联网，此时将上报初始化数据。主要内容包括：功能码、设备唯一识别码（International Mobile Equipment Identity, IMEI）、当前信号强度（Received Signal Strength Indicator, RSSI）。其中，功能码用于表示上报数据类型，固定值为 INIT_ID。IMEI 表明当前

终端的唯一身份，通常由 15 位数字组成，是云平台能够正常接收数据的前提。RSSI 表示当前采集终端所处环境网络的好坏，若该值大于 12 时就表明当前环境下网络质量较好，终端可以正常工作。

表 2.2 周期上报数据内容

参数项	类型	内容示例
1	功能码	M_NORMAL_REPORT_ID
2	设备唯一识别码	869858036950666
3	电流数据	{0.48, 0.48, ..., 0.51}

终端联网成功后将正常开始工作，此时开始定时上报电流数据。数据内容包括 IMEI、电流数据、功能码等。平台在接收到数据后，会自动添加接收时间并向终端响应 ACK，这个过程通常不会超过 1 分钟。

按照上述方式，每台仪器仅与一台采集终端绑定，这将仪器抽象成了一个个数据节点，节点会不断向云平台发送数据流。通过对些数据流的内容进行解析，可以知道仪器在什么时候产生了多大的电流，并可以知道一共持续了多长的时间。

2.3 设备接入平台设计

设备接入平台为采集层的上层，除了接入采集终端以外，还需要具备数据接收、解析、存储等能力。仅靠个人设计一套安全、稳定、支持高可用、高并发的设备接入平台的难度太高，后续运维也难以继，因此考虑选用国内相关公司的物联网云平台作为设备接入平台的框架。

云平台是一种由硬件基础设施与软件资源构成的服务平台，在这个概念中，计算机的所有资源是虚拟化的商品，用户可以根据自身需求进行购买。云平台有三种模式^[45]：基础设施即服务（infrastructure as a service, IaaS）、软件即服务（software as a service, SaaS）与平台即服务（platform as a service, PaaS）。IaaS 主要为用户提供虚拟计算机，用户可以自选 CPU 核数、内存和磁盘大小以及操作系统，国内成熟的云服务器（Elastic Compute Service, ECS）厂商有阿里、华为、腾讯等。SaaS 主要为用户提供现成软件服务，软件运行在云端而非本地，如云相册、云盘等。PaaS 面向开发者，其自带的运行环境以及分布式系统极大提供了开发效率。

国内三大运营商凭借着得天独厚的优势，对云平台的研究起步较早，已经颇具成果^[46]。随着国家新基建需要，三大运营商加入了对工业物联网平台的研发中，推出了 OneNET、天翼物联等优秀产品。考虑到开发成本、现有生态及相关领域参考方案等，本设计决定采用电信 AEP 作为设备接入平台。

在接入设备前，首先需要在 AEP 云平台中创建产品，相关操作包括添加产品名称、选择

产品分类、选择节点类型、选择接入方式等。其中，根据产品的功能定义中的服务列表和属性列表为产品添加功能，根据节点类型定义其入网方式、通信协议、设备型号等。产品创建完成后就可以添加设备了，此处需要执行填入设备名称、IMEI 号、IMSI 号，选择是否订阅等操作，这样一来就在云平台创建了设备的实体。此时，采集终端再使用 IMEI 作为鉴权信息时，云平台就可以识别设备了。创建产品、添加设备界面如图 2.10 所示。

图 2.10 产品创建、设备添加界面

2.4 应用服务平台设计

2.4.1 数据存储中心

考虑到 AEP 平台数据读写存在诸多限制，不利于后续自定义功能开发，故搭建了数据存储中心来存储系统中的所有数据。数据存储中心需要存储仪器基本信息、相关模型、以及机时电流等数据。

不同类型的数据，适用的数据库不同。仪器基本信息可以存到 MySQL 这样的关系型数据库，因为这一类数据不会太多，而且通常较为固定，不会发生频繁的变化。但对电流数据、机时数据来说，MySQL 就不太合适了。由于这类数据的数量太大，而且几乎所有业务场景下的查询都是范围查询，查询结果还需要进行排序，即使通过设置主键和索引，查询速度也是很慢的，若涉及到长事务执行就容易造成操作阻塞，可能会导致数据库崩溃从而影响系统可用性。面对此类问题，业界常用的做法是采用 Sharding-Proxy、ShardingSphere 等工具进行分表分库的方式来降低单个节点的压力，对 IMEI 进行取余、求 Hash、设置 Range 划分是常用

策略。不过，考虑到可能存在的跨节点查询、分页排序、分表将引入更多的复杂性、后续拓展维护困难等问题，因此考虑采用 MongoDB 做电流数据的持久化。

MongoDB 是一种基于分布式文件系统的非关系型数据库^[47]，内部粒度按照大小分为数据库、集合、文档，存储结构如图 2.11 所示。

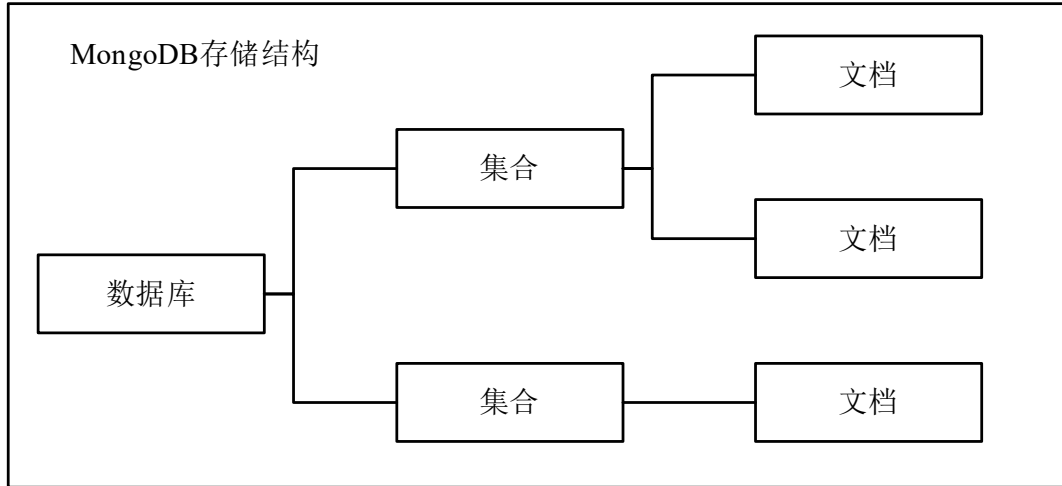


图 2.11 MongoDB 存储结构

MongoDB 支持动态扩展，当读写请求压力大的时候可以将其部署为主从模式，其故障自动转移功能可以保证整个数据库的高可用。随着电流数据量持续增加，横向拓展也无法缓解存储压力时，MongoDB 也提供了分片机制对数据库进行切分，保证在面对海量数据时数据库的使用效率。

由于仪器往往会长时间保持一个状态，采集终端采集的电流数据很可能在一段时间内是不变的，如果全量存储这些数据必然会造成大量冗余。考虑到当前场景下电流数据的特点，本设计将采用压缩算法对数据进行处理，减轻存储压力。行程长度压缩算法^[48]（Run Length Encoding, RLE）是一种简单的无损压缩算法，该算法将目标数据按照顺序分为连续重复数据块以及不重复数据块进行压缩。这种处理方式使得数据压缩与解压的过程都十分简单，算法时间复杂度为 $O(n)$ ，对线上使用友好，模拟过程如图 2.12 所示。

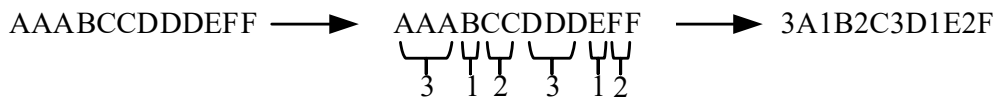


图 2.12 RLE 压缩模拟过程

压缩电流的具体细节为：读入电流数据后，以第一个数据为标定值，然后遍历后续数据，当后续数据与标定值相同时，数据长度加 1，然后继续向后遍历；当后续某个数据与标定值不同或与相邻数据的时间间隔相差超过 1 分钟时，记录当前标定值、标定值对应的时间戳、以及该段数据长度，然后更新标定值，直至所有的数据都压缩完毕。算法的执行流程如图 2.13 所示。

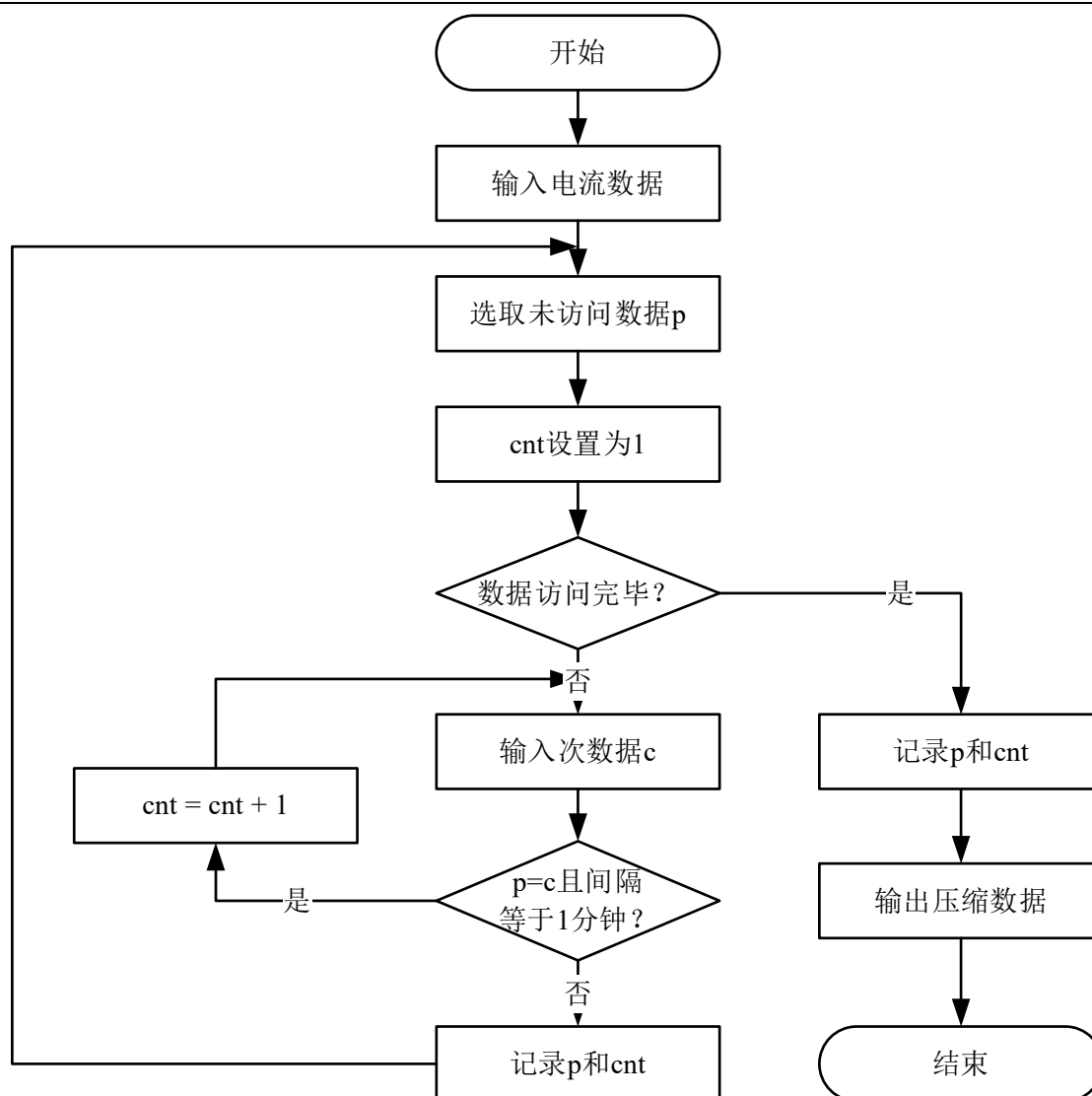


图 2.13 行程长度压缩算法流程

使用该算法压缩后，MongoDB 中电流数据的存储结构定义如表 2.3 所示。示例为 2021-10-06 00:00:00 开始，持续时长 454 分钟，值为 0.02 的一段电流数据压缩后的情况。

表 2.3 压缩数据存储格式

字段名	类型	说明	示例
time	Int64	第一个值的时间戳	1633449600000
value	Double	值大小	0.02
minutes	Int32	持续时长	454

压缩后的电流数据被存入 MongoDB 的电流集合 processedData 中，为了存储结构清晰及后续分析方便，规定一台仪器一日的电流数据为一个文档。此外，鉴于 MongoDB 文档结构的优点，在 MongoDB 中还建立了 deviceProperties、model、deviceUsage 集合分别用于存储仪器电流特征、相关模型、机时等数据。

除了 MySQL 与 MongoDB 以外，数据存储中心还使用了 Redis 作为缓存提高性能。Redis

作为一种基于内存的数据库，底层单线程操作加上非阻塞 I/O 多路复用机制使得其读写性能十分的出色。许多数据，如电流特征、历史机时、模型等在生成之后几乎不会改变，如果将其放入缓存中，那么读的速度远高于从数据库中读取。最终，建立的数据存储中心结构如图 2.14 所示。

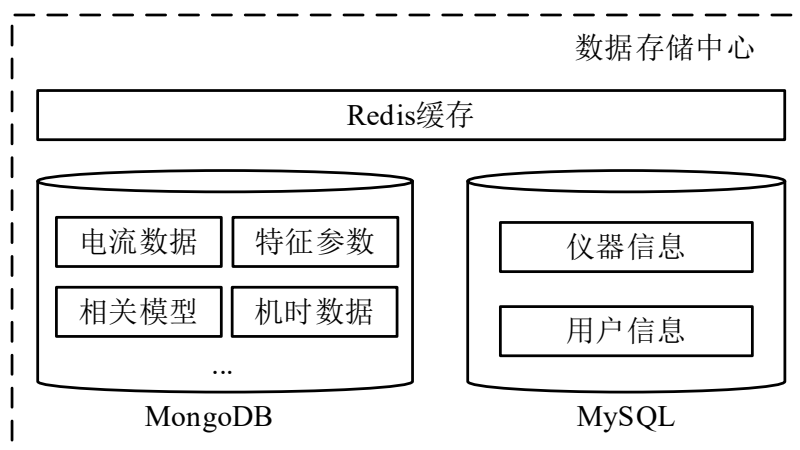


图 2.14 数据存储中心结构

2.4.2 系统业务平台

业务平台搭建在阿里云 ECS 之上，代表着整个系统的执行逻辑。虽然用户无法感知业务的实现，但是良好的业务逻辑设计和合理的后台架构可以切实提升用户体验。

2.4.2.1 服务开发框架

当下业务中核心业务包括数据持久化、模型生成、机时计算等，由开发者自行编写。非核心业务包括日志收集、监报告警等，可以自己设计，也可以调用第三方成熟的服务。考虑到低耦合、可重用、易维护的设计要求，本设计将业务拆分成多个服务，服务间通过应用程序接口进行交互。

SpringBoot 框架内嵌了 Tomcat 容器可以自动处理外部请求，内置 Maven 实现依赖自动化管理，通过简单的配置文件就可以快速搭建起一个 Web 服务。并且 SpringBoot 还提供了 MyBatis、RedisTemplate、MongoTemplate 等常用套件来方便开发者处理各类数据库，安全策略简洁^[49]。使用 SpringBoot 框架可以在快速对系统进行构建，因此本设计选其作为各服务的基本开发框架。

此外，由于目前 Java 对于机器学习的接入较少，即使存在 JAVA-ML、DJL 等工具，但开发起来还是较为困难，所以除了 SpringBoot 以外，还设计了一个基于 Python 的 Web 服务，借助其中的 sklearn 工具包来建立涉及电流分析、机时计算的所需模型。

2.4.2.2 核心业务设计

核心业务包括数据持久化业务、模型生成业务、机时计算业务，如下图 2.15 所示。

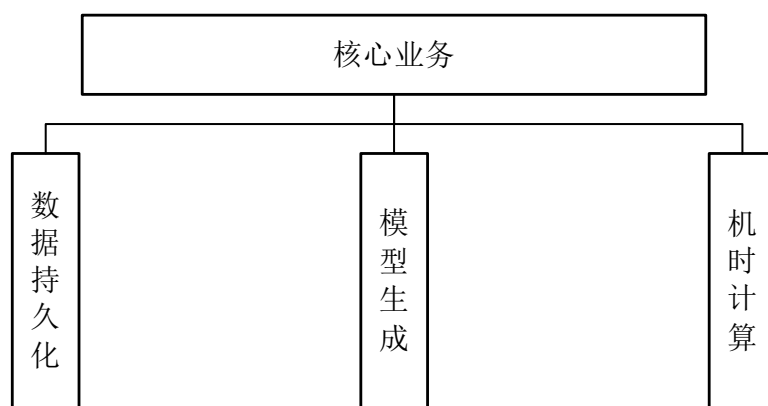


图 2.15 核心业务框图

1. 数据持久化业务

该业务主要负责将 AEP 平台中存储的电流数据转存至 MongoDB。由于电流数据持续上报，按照每日一个文档的设计，要将每日的电流数据转存首先需要在系统中添加具有定时功能的模块。在 Java 中，已经有许多成熟的定时器工具类，这里选用 Hutool 中封装好的 CronUtil 定时器进行实现。具体步骤为：首先在项目的 pom 文件中添加对应依赖，然后在执行入口方法中添加 CronUtil.setMatchSecond(true)与 CronUtil.start()两行代码，最后在项目文件夹下添加 cron.setting 文件，并设置定时器执行方法所属的类以及定时规则。程序触发定时规则后，将通过反射开启一个线程去调用相应任务。按照目前的执行速度，定时任务在凌晨 2 点开始执行，在凌晨 4 点前就可以执行完毕。设计好定时器后，接下来对数据持久化业务进行设计。具体内容为：首先从 MySQL 中读取全量仪器 IMEI 进行遍历，然后根据当前 IMEI 读取 AEP 平台存储的前一日数据，数据内容包括上报的电流数据以及数据的上报时间，再按照上报时间为各电流数据打上时间戳，时间戳精确到分钟即可，最后将上述数据压缩后持久化至 MongoDB 中。在这个过程中有两个问题需要解决，分别是数据跨天、重复数据。

1、数据跨天处理：由于采集终端上报时间无法统一，很多情况下会出现一条上报数据中存在跨天的数据，比如从某个上报数据中第一个数据为前一日的 23 点 45 分采集的，最后一条数据为当日 0 点 14 分采集的。为了形成每日一个文档，系统在每次执行该任务时会额外多读取前后 1 个小时的数据，即读取 26 个小时的数据，之后再根据时间点进行切割，使得每日的第一条数据肯定是从 0 点 0 分开始，最后一条数据到 23 点 59 分结束。数据切割完毕之后，就可以按照压缩格式写入电流集合中。

2、数据重复处理：由于存在重发机制，那么当网络情况较差时，上报数据到达云平台耗时较长导致采集终端重发，此时云平台会收到两条相同数据。故而在获取到全部数据后，需要进行去重处理。去重的规则为：由于电流数据的时间戳精确到分钟，当出现多条数据时间戳的分钟位重复时仅保留第一条数据。

经过上述处理，数据库中的文档就变得十分规整，业务上清晰明朗。每处理完一台仪器，就在 Redis 中记录该仪器 IMEI 以及当前日期，标记已经转存完毕。这样做到好处是，当该服

务崩溃后再次重启时，可以直接通过 Redis 缓存内容找到哪些仪器没有处理完毕，避免出现大量无意义的重复动作，影响系统性能，持久化业务执行流程如图 2.16 所示。

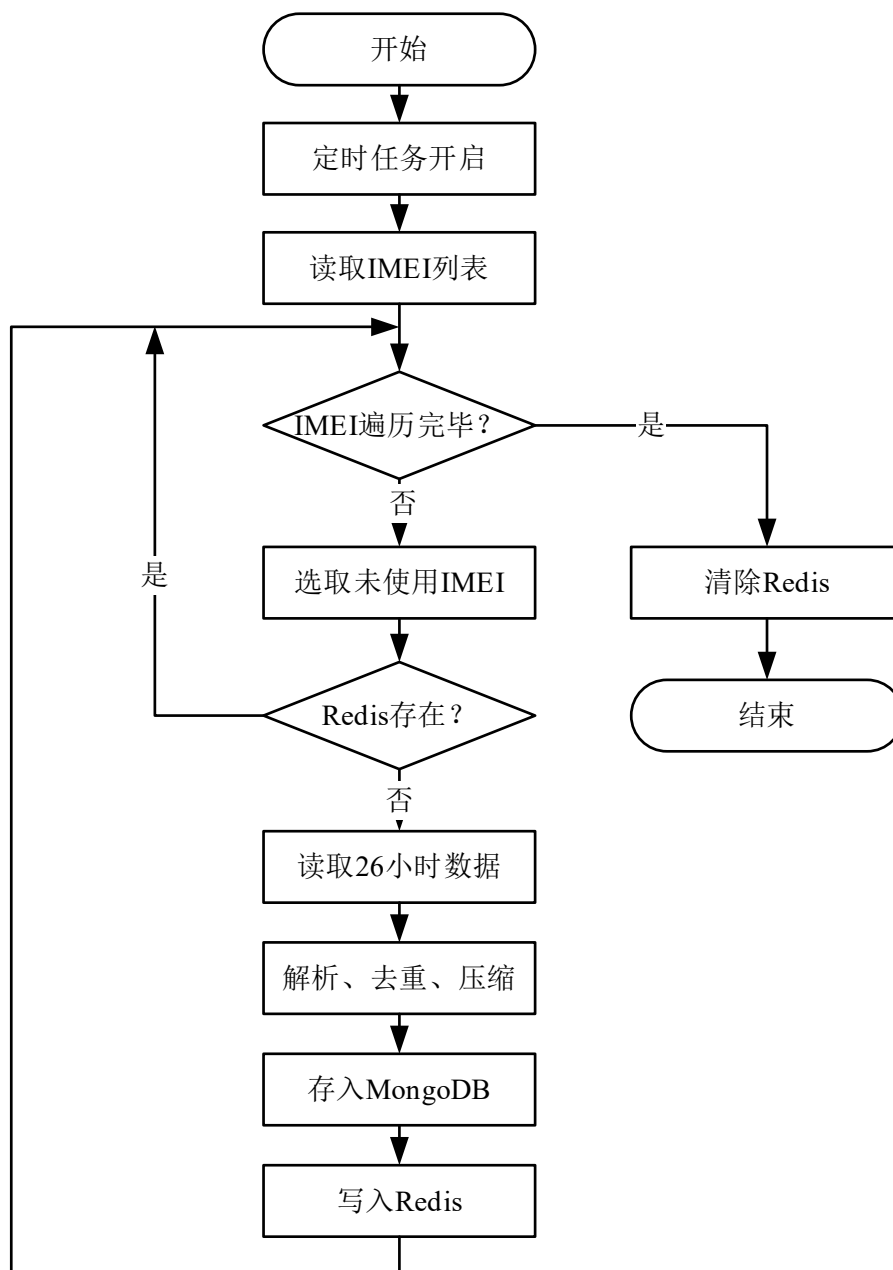


图 2.16 持久化业务执行流程

2. 模型生成业务

考虑到当前场景仪器种类、数量庞大，要计算各仪器机时的实施难度以及操作成本较高，故本文共设计了基于分箱与改进 K-means 及基于 DP-SVM 两种机时计算方法，这两种方法的核心是其中的电流分类模型。其中，基于分箱与改进 K-means 中电流分类模型在生成时可以操作简单，可以在不用标注数据的情况下解决多种仪器的电流分类问题，其中电流分类模型的生成流程如图 2.17 所示。

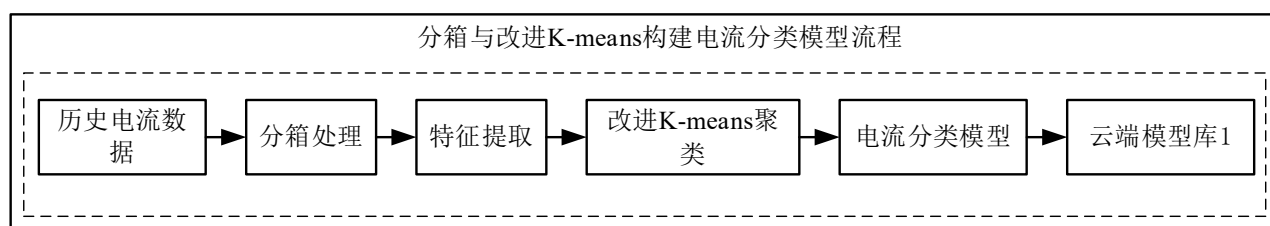


图 2.17 电流分类模型生成流程

当仪器历史数据量达到一个月时，就会触发上述流程。由于基于分箱与改进 K-means 构建的电流分类模型的核心是其中的电流特征，聚类输出的仅仅是数字标签，因此需要设计一套电流特征与实际状态对应的规则。参考专家的建议，目前状态对应规则设计为：当执行结果中簇的个数为 1 时，认为仪器状态不全，暂定为关机；当执行结果中簇的个数为 2 时，电流特征中均值小的为关机，电流特征中均值大的为工作；当执行结果中簇的个数为 3 时，以电流特征中均值最小的簇对应的状态为关机、电流特征中均值次小的簇对应的状态为待机、电流特征中均值大簇对应的状态为工作。按照该规则处理完成后，当聚类结果中簇个数大于 1 时，将簇心对应样本的电流特征以及该特征的推断标签（关机、待机、工作）存入 deviceProperties 集合，若簇个数为 1 时，将仪器标记为数据不足，并记录该仪器下一次聚类时需要的数据量。标记仪器会被定期调用，读取对应数据量执行聚类，直到簇个数不为 1 了再取消标记。

基于分箱与改进 K-means 构建的电流分类模型对某些电流复杂的仪器可能会失效，因此提出了基于 DP-SVM 的机时计算方法作为补充。该方法中核心是 DP-SVM 电流分类模型，其中 SVM 训练完成后以仪器 IMEI 命名，然后使用 Joblib 工具存储至云端，流程如图 2.18 所示。

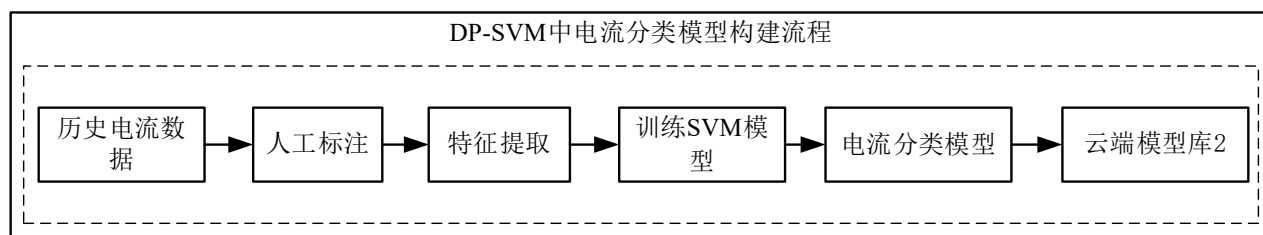


图 2.18 电流分类模型生成流程

3. 机时计算

此业务负责计算仪器机时。每日电流数据持久化后：首先调用接口将 processedData 中指定的压缩电流数据进行解压，然后根据 IMEI 执行相应的方法，最后分别统计标记为工作、待机、关机的电流数据个数作为机时计算结果保存，单位为分钟。

此处有一个地方需要注意：实际中若发生断电的情况，会导致电流数据解压后在时间上会出现间隔，即前一个压缩块解压后的最后时刻和下一个压缩块解压后的最早时刻的间隔超过 1 分钟。所以在计算机时的时候，需要建立一个与电流数据个数相同大小的状态数组，将缺失时间点对应的位置上标记为-1，电流分类后，状态数组各位置的标签为该时刻电流的标签。之后，按照电流压缩的方法对状态数组进行压缩保存，这样一来，不仅在读取机时数据

时只需要统计各压缩块记录的时长，而且也可以知道具体时刻仪器的运行状态。

由于存在不同机时计算方法，数据库中将基于分箱与改进 K-means 的机时计算方法记为 1，基于 DP-SVM 的机时计算方法记为 2，两种方法的使用逻辑如图 2.19 所示。

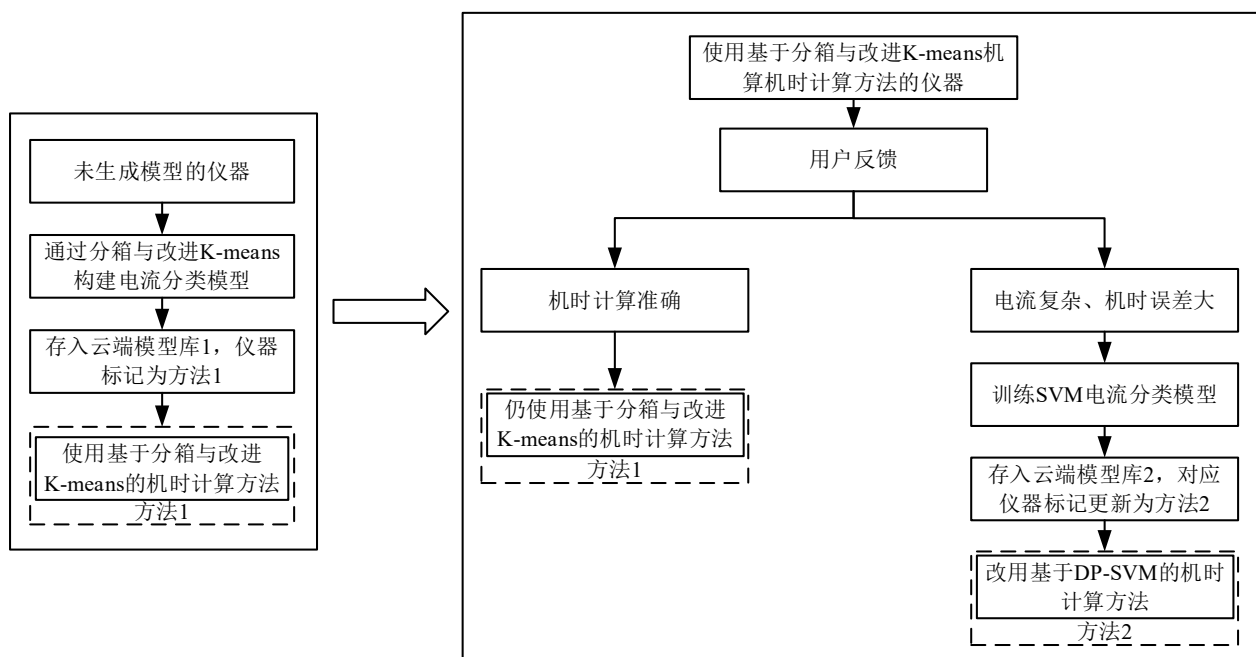


图 2.19 机时计算方法使用逻辑

对于未建立模型的仪器，首先使用基于分箱与改进 K-means 进行建模然后求解机时，当用户反馈机时计算误差大后，若无法调整则该仪器改用于基于 DP-SVM 的机时计算方法。由于有两种机时计算方法，为了统一处理，断电状态的标签统一为-1、关机状态的标签为 0、待机状态的标签为 1、工作状态的标签为 2。机时计算完成后将数据存至 MongoDB 中并在 Redis 中进行记录，机时计算示意图如图 2.20 所示。

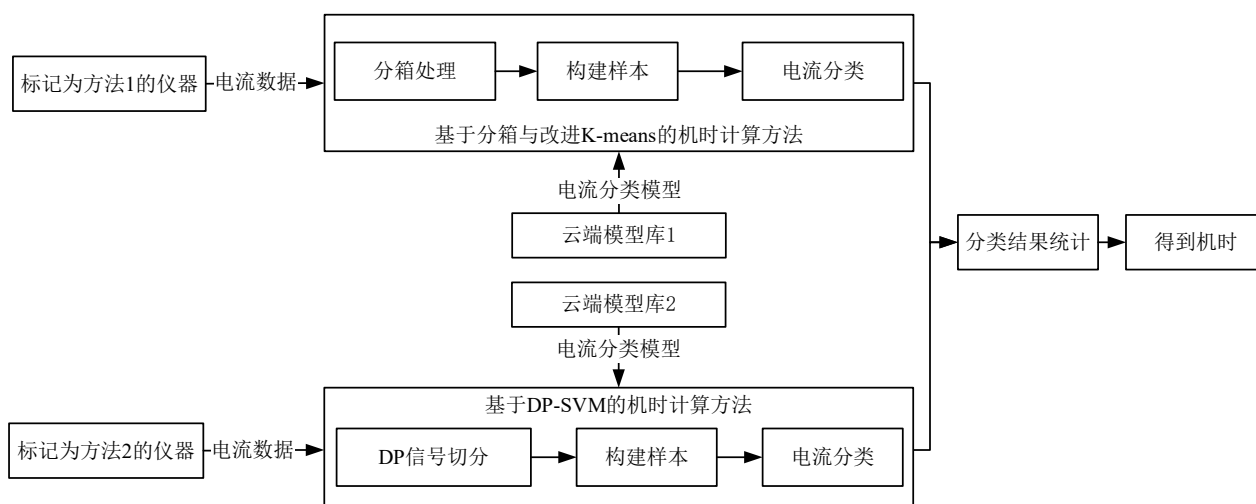


图 2.20 机时计算示意图

2.4.2.3 业务平台架构

用户的访问请求通过前置 Nginx 负载均衡，分发到指定的服务器上。MongoDB 配置成集群模式来减轻访问压力，提高可用性。为避免仪器数量太多导致持久化、计算机耗时过长，配成多机集群后，采用对 IMEI 取余的方式实现集群同时处理不同的仪器。Rdies 在系统中负责缓存模型以及配合持久化业务，为提高其可用性，将采用哨兵模式进行配置，无需人工干预。业务架构如图 2.21 所示。

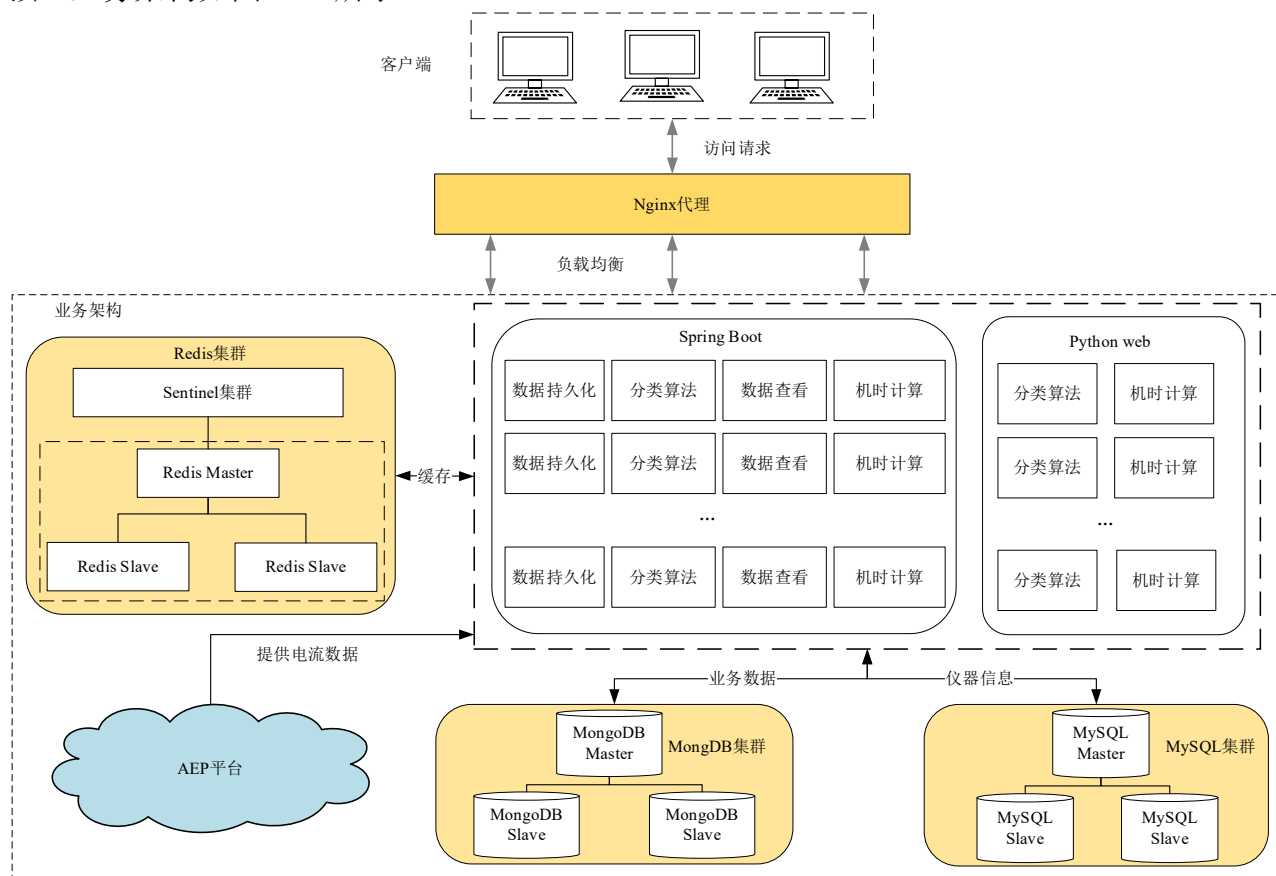


图 2.21 业务架构图

2.4.3 Web 端设计

Web 端是用户与系统交互的主要方式，Web 端功能如图 2.22 所示。

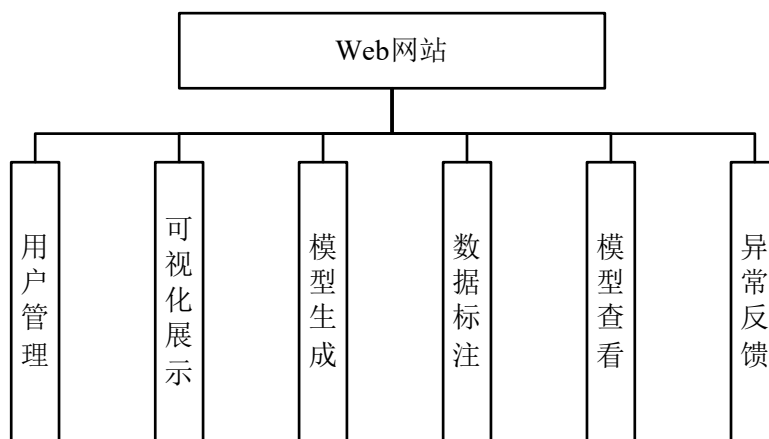


图 2.22 Web 端功能图

Web 端的详细功能包括用户管理、数据可视化展示等，具体细节的如表 2.4 所示。

表 2.4 Web 详细功能表

条目	内容
用户管理	用户注册登录、管理员注册登录
可视化展示	仪器基本情况（电流波形、机时等）
模型生成	生成相关模型
数据标注	对电流数据进行标注
模型查看	查看模型类型、参数等
异常反馈	用户提交仪器损坏、机时错误等反馈

仪器机时和模型密切相关，在涉及到模型修改方面需要慎重处理。因此，本系统参考基于权限的角色控制模型（Role-Based Access Control, RBAC）对用户权限进行设计，将权限划分为多个等级。普通用户注册时需要绑定所属机构，登录系统后只能查看当前机构的所有仪器详情、电流、机时等内容，无法对模型作出修改；管理员分为机构管理员与超级用户，机构管理员可以看到本机构下所有仪器的情况以外，还可以修改模型；超级管理员具有最高权限，可以查看所有仪器的信息，也可以修改模型。

2.5 整体效果展示

2.5.1 采集终端展示

采集终端如图 2.23、图 2.24 所示，三相采集终端结构与单相类似，仅互感器单独引出。采集终端接入大型科研仪器的现场如图 2.25 所示。

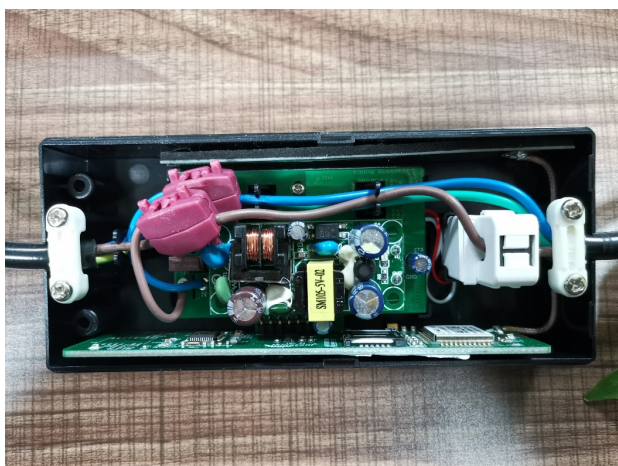


图 2.23 单相采集终端主板



图 2.24 单相采集终端实物

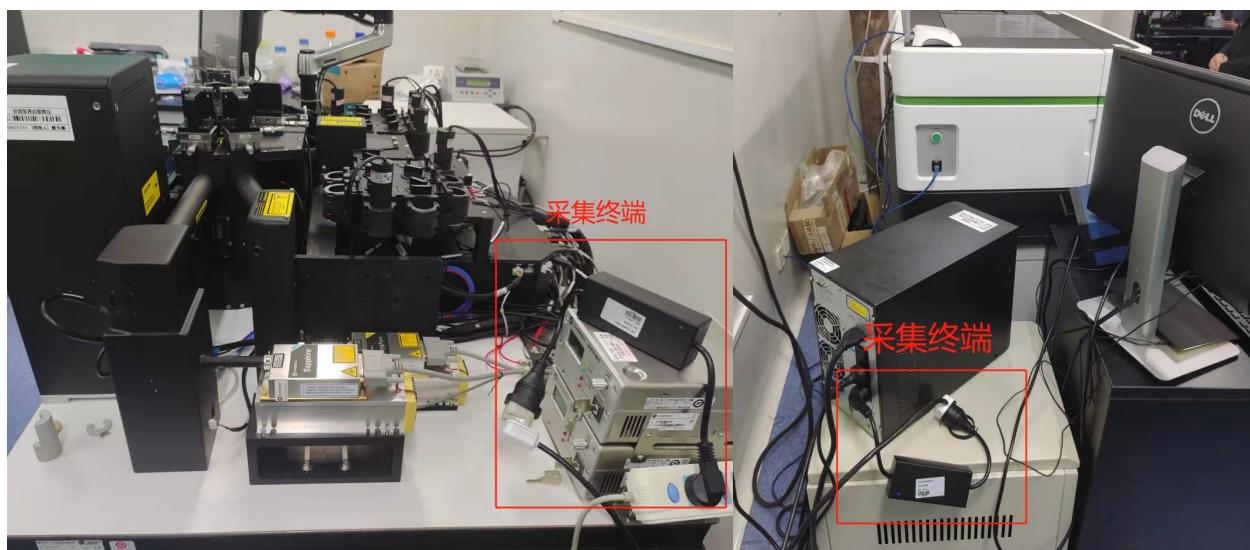


图 2.25 采集终端现场接入情况

2.5.2 AEP 云平台展示

登入 AEP 平台后，设备管理界面如图 2.26 所示，可以查看设备名称、IMEI 等信息。

产品概况

设备管理

数据查看

指令下发日志

订阅管理

远程升级管理

消息跟踪

批量任务

请输入设备名称、设备ID、IMEI

Q

在线状态

设备状态

批量添加模板

添加设备

批量删除

批量添加

导出设备

	设备名称	设备ID	IMEI	IMSI	创建时间	最后上线时间	最后离线时间	状态	操作
☐	杭州电子科技大学-2...	16b717dc71040f19...	864823041537606	460113118530284	2020-10-23 17:41:58	2021-10-21 16:39:56	2021-10-25 20:17:18	已激活	
☐	浙江海洋大学-20204...	3a1dd798efc422ab...	864823041533464	460113118530279	2020-10-23 17:40:56	2021-10-08 06:15:35	2020-10-26 14:55:18	已激活	
☐	杭州师范大学-20121...	e190e984ebd743ac...	864823041532870	460113118508799	2020-10-23 17:40:15	2021-08-05 22:47:08	2020-10-26 10:37:56	已激活	
☐	湖州食品药品研究...	8bfc1e4480044e93...	864823041545088	460113118508800	2020-10-23 17:39:32	2021-10-20 00:45:27	2021-10-20 00:37:28	已激活	

共 5097 条

10条/页

< 1 2 3 4 5 6 ... 600 >

前往

3 页

图 2.26 AEP 平台设备管理界面

设备详情界面如图 2.27 所示。

设备详情	事件上报	数据查看	指令下发日志	Object对象	消息跟踪
设备详情 设备ID16700ee8c99c4f68908dac2ca2463a94 IMEI864823041542309 secret-- 生命周期状态已激活 创建时间2020-10-23 17:51:31 设备名称湖州师范学院-1803176S IMSI460113118530287 固件版本-- 在线状态在线 最后上线时间2021-11-29 17:01:23					
认证信息 设备ID16700ee8c99c4f68908dac2ca2463a94 IMEI864823041542309 认证方式IMEI认证 通信协议LWM2M					

图 2.27 设备详情界面

图 2.28 数据上报界面

图 2.29、图 2.30、图 2.31、图 2.32 展示了多种仪器的电流波形以及仪器的运行状态。电流数据和仪器状态是严格对齐的，可以使用鼠标滚轮间来方法缩小查看细节。

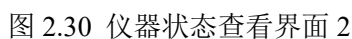
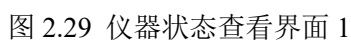




图 2.31 仪器状态查看界面 3



图 2.32 仪器状态查看界面 4

仪器的机时如图 2.33 所示, 该仪器在查询时间内一共工作了 152 分钟。

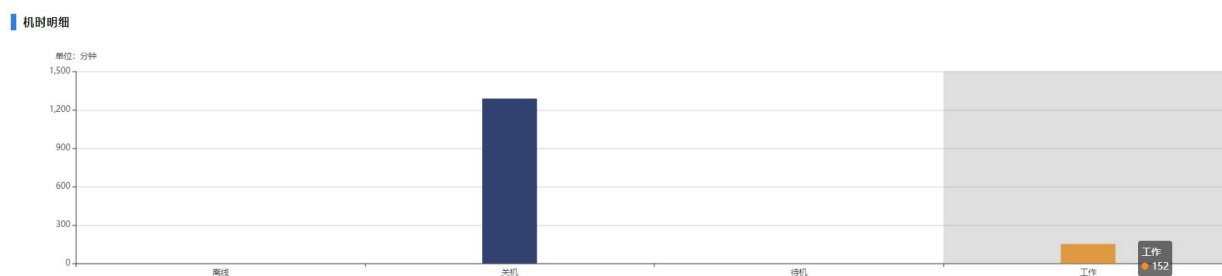


图 2.33 机时数据

除了这些以外, 电流分类使用的电流特征也可以供用户查看, 如图 2.34 所示。

特征值

类型	均值	标准差
关机	0.0	0.0
待机		
工作	1.616754030161198	0.09446171405323692

图 2.34 仪器的电流特征

若电流特征质量不佳，那么也提供了手动聚类求解功能，界面如图 2.35 所示。用户可以选择该仪器任意一段时间内的电流数据，然后直接进行手动聚类求解电流特征，聚类结果是实时展示的，可以协助用户判断其质量，聚类后的电流特征会自动更新。

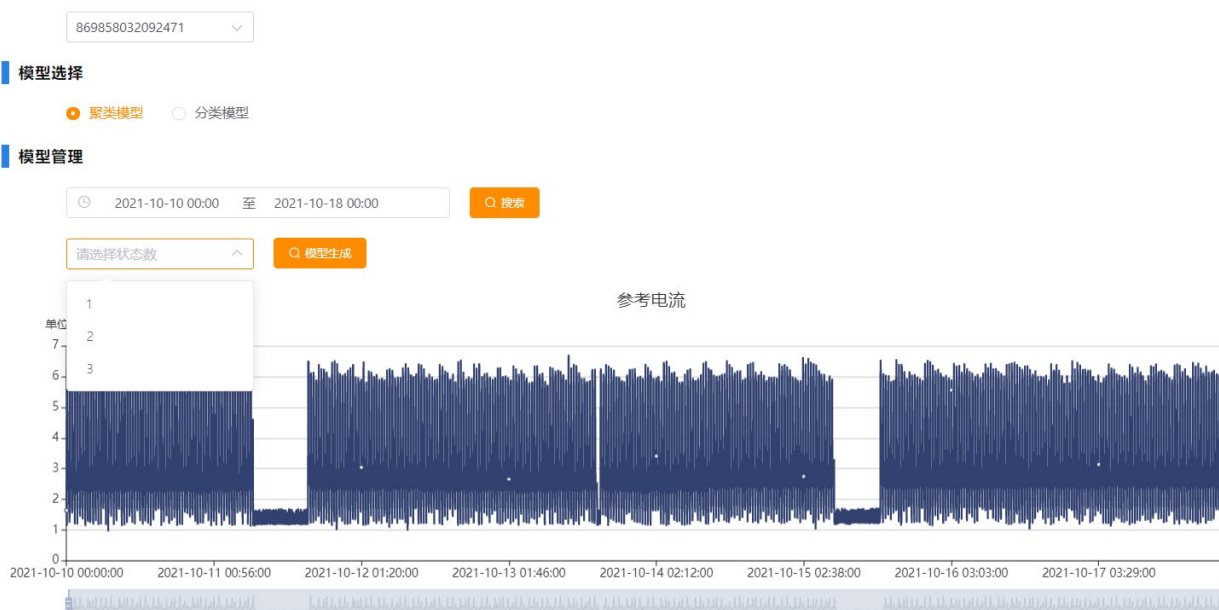


图 2.35 手动聚类界面

目前，系统中接入仪器共 8848 台，浙江省大型科研仪器开放共享平台中的仪器机时相关的数据就是调用本系统的接口所得，该平台首页如图 2.36 所示。



图 2.36 用户 Web 端

进入仪器使用界面，可以看到各高校机构的仪器数量、仪器总价值、采集终端的安装率以及仪器平均使用机时等内容，如图 2.37 所示。

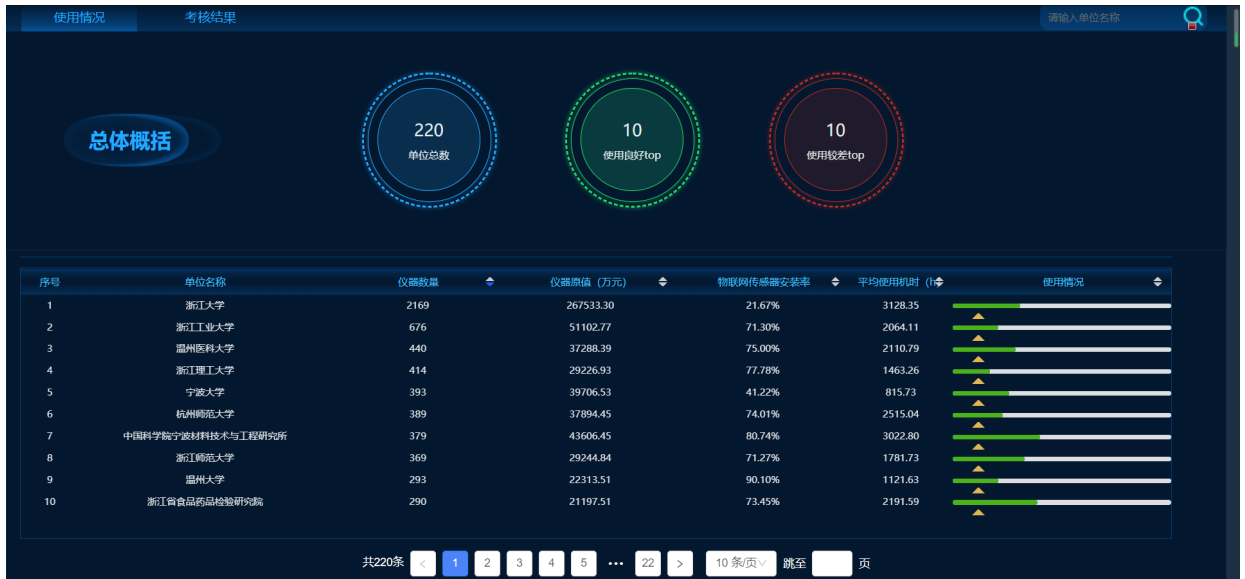


图 2.37 仪器使用界面

登录管理员账号进入展示界面，可以看到仪器年机时、有效机时等内容，如图 2.38 所示。

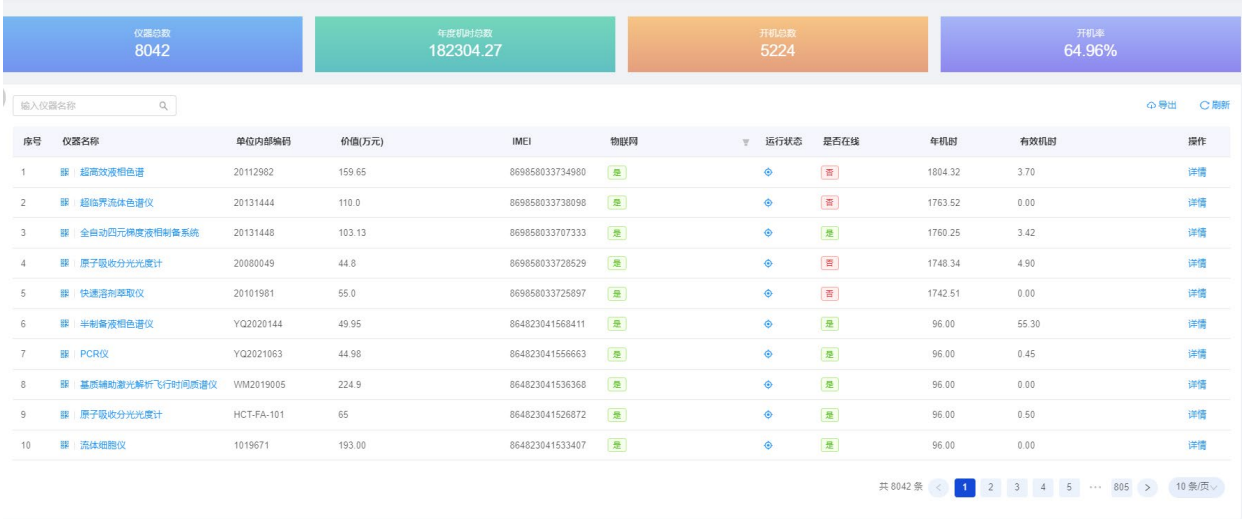


图 2.38 仪器机时展示界面

2.6 本章小结

本章主要阐述了大型科研仪器机时智能计算系统的设计实现过程。首先分析了现实需求，提出了系统的总体设计。然后对系统各部分的搭建过程进行了详细介绍，包括：采集终端的硬件选型、软件及通信协议的设计；设备接入平台的选型、设备接入流程介绍；数据库选型以及电流、机时等数据压缩存储结构的设计；电流转存、模型生成、机时计算等业务的设计，并给出了业务架构图；最后展示了系统实现效果。

第3章 基于分箱与改进 K-means 的机时计算

传统机时计算依赖人工，存在着较多的问题，亟需一种自动化的方法进行处理。按照前文分析，若可以通过电流区分出仪器不同的运行状态，那么对不同状态下的电流数据进行分类后，统计出各类电流数据的个数就可以达到获取机时的目的。电流阈值法是实际工程中区分仪器运行状态最常用的方法，其步骤为：首先，人工查看仪器一段时间内的运行电流，比对后为该仪器设置几个合理的电流阈值，然后根据阈值与实际电流的大小关系来判断仪器的运行状态。该方法虽然操作简单，但当仪器数量与种类数量增加时，人工一一设置太过繁琐且效率太低，在工程中使用不太实际。为了使该方法能在大量仪器的场景中应用，有学者研究后发现，同一仪器在相同状态下的电流往往都在一个幅度内波动，不同的状态恰好对应聚类概念中的不同簇，通过聚类计算出各仪器的电流特征再处理后就可以得到阈值^[50]。在实际中应用后，发现所述方法在信号分析仪等一些电流较为“简单”的仪器上求出的阈值较为合理，最后计算的机时与人工统计结果基本一致。但随着接入仪器种类的增多，该方法在凯氏定氮仪、气相色谱仪等多种仪器上出现了较大的偏差。对这些仪器进行分析后发现，造成误差的主要原因是这类仪器待机和工作时的一部分电流大小较为接近，存在工作电流小于等于待机电流的情况，这时就出现了阈值求解不够合理，以及无法通过阈值来判断仪器运行状态的问题。电流阈值法失效后，按照现有的其它方法就需要人工介入建立更复杂的模型才能达到区分状态的目的，但这样一来系统的操作难度与建设成本也随之上升。

由于上述方法在许多电流特征较为明显的仪器上表现较差，为了在满足更多种仪器机时计算需要的同时尽可能降低系统实施成本，提高处理效率，因此本章提出了一种基于分箱与改进 K-means 的机时计算方法，总体思路如图 3.1 所示。

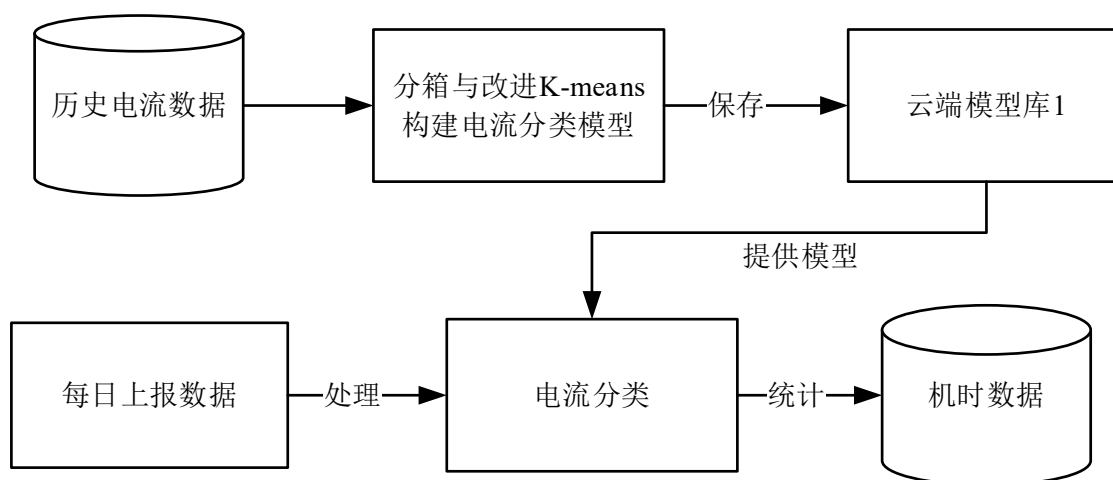


图 3.1 总体思路

对于某一仪器，首先通过基于分箱与改进 K-means 算法构建出电流分类模型并存至云端。

后续执行机时计算时，再调用该仪器云端的电流分类模型进行电流分类，最后将分类结果统计后得到机时保存至对应数据库。

3.1 仪器电流数据特点

电流数据是采集终端从仪器侧采集而来，不论仪器处于关机（非断电）、待机还是工作状态下都会存在。经过现场调研和查看大量仪器的电流数据后，发现虽然各种仪器的电流波形不尽相同，但还是可以找到一些普遍特点：

1、虽然仪器种类众多，但同种仪器在不同的型号下电流波形也可能有不同，但同一台仪器在相同状态下的电流波形是十分相似的，或者说同一状态下电流数据分布基本一致。

2、电流数据是一种时间序列，往往一段时间内电流数据都对应同一个仪器状态，或者说无论仪器处于何种运行状态，这种状态通常都会持续一段时间，很少出现频繁的变化。

3、仪器关机状态下的电流值普遍低于待机状态下的电流值，待机状态下的电流值普遍低于工作状态下的电流值。这是由于仪器关机时，除了按键检测等一些基本模块以外没有其它模块在工作，几乎不消耗任何电流。待机相比工作也是同理，工作时往往有更多模块参与，这会消耗更多的电流来支持仪器的运作。

4、关机状态下电流基本不会变化，但不一定是 0A。虽然在普遍的认知中，仪器关机后电流值应该为 0A，但随着软关机的仪器越来越多，一些仪器处于关机状态但是电源按键检测模块可能还在工作，那这时还是存在着一定的电流。而待机电流和工作电流根据仪器类型的不同，可能平稳，也可能存在一定的波动，甚至在电流值上还会存在重叠。

图 3.2 展示了从 8 台不同的大型科研仪器上采集的电流波形，涉及时长为一天。为了方便起见，横轴的坐标按照电流出现的时间顺序来进行设置，最早出现的电流的横坐标越小，单位为分钟。

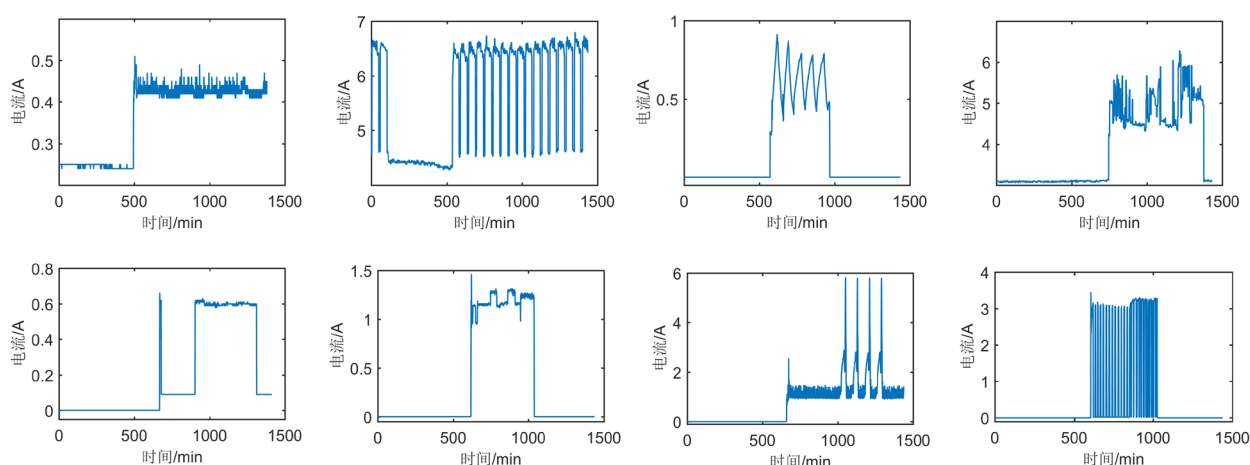


图 3.2 8 台仪器的一日电流波形

3.2 电流数据处理

3.2.1 数据分箱

电流阈值法效果不佳的根本原因在于仅通过电流值这单一维度对仪器状态进行判断，许多被误判的仪器实际上在电流波形的特征上还是较为明显的，若能够提取到这些特征就能够进行有效区分。此外，考虑到由于输入的电流数据往往包含许多状态，后续分类时不能人工事先进行切分，因此就需要一种方法将电流数据划分为电流量列。

分箱算法是一种简单的数据处理方法，按照一定的规则将原始数据分成多组，这样处理后可以降低算法对异常数据的敏感度以及运算的复杂度，对线上使用较为友好。为了实现电流的自动切分以及特征提取，基于电流波动特点以及仪器的使用特性分析后提出了一种分箱算法。通过该算法将电流量列切分成为多个电流子序列再进行分析，具体过程为：假设输入电流量列为 $X = \{x_1, x_2, \dots, x_n\}$ ，首先构造 X 的电流值序列 Y 用于表示 X 的值域分布情况， Y 中数据按照从小到大的方式进行排序，可以记作 $Y = \{y_1, y_2, \dots, y_m\}$ ，并且两者的长度大小关系为 $1 \leq m \leq n$ 。其中， m 为值不重复的电流个数，当 X 中所有数据的电流值都相同时， $m = 1$ ，当 X 中没有重复电流值时， $m = n$ 。以序列 $X = \{1, 1, 2, 3, 4, 5, 8, 8, 9, 11, 1, 1, 1, 1, 2, 2, 4\}$ 为例，那么 X 的值序列 $Y = \{1, 2, 3, 4, 5, 8, 9, 11\}$ 。

在获取到电流值序列 Y 以后，以 Y 中最小的幅值 y_1 为起点，对 X 进行分箱处理。由于当前采集终端的采集精度为 $0.01A$ ，这意味这电流的最小分辨率为 0.01 ，因此分箱的策略设计如下：

以 y_1 作为标定点开始向后遍历，对于下一个电流值 y_2 ，检查 y_2 与前一个电流值 y_1 的差值是否为 0.01 ，如果成立，继续向后遍历 y_3 ，再判断 y_3 和 y_2 的关系。如果不成立，那么记录下第一个标定点以及当前点的上一个点作为分箱的范围，更新标定点为 y_2 ，继续下一次范围求解直至所有电流值遍历完毕。最终，得到的多个分箱的范围 P 以式 (3.1) 进行描述。

$$P = \begin{cases} [y_1, y_i] \\ [y_{i+1}, y_j] \\ \dots \\ [y_{l+1}, y_k] \end{cases} \quad (3.1)$$

分箱范围确定之后，记 y_1 为下边界 l ，依次取每个范围的最大值作为上边界 h ，形成边界范围 $[l, h]$ ，然后去切分电流量列 X 。切分逻辑设计为：首先搜索 X ，若存在某段未被访问过的电流子序列的电流值处于上下边界范围内，且该子序列持续长度超过 T ，那么就将这段子序列标记为访问并放入集合 S 中。若当前边界范围内已经不存在满足条件的子序列时，就更新上边界，直至所有上边界使用完毕。

最终可以获得电流子序列集合 $S = \{s_1, s_2, \dots, s_k\}$ ，且 S 中所有子序列长度之和会小于等于 X 的长度。在 S 中，每个子序列是仪器一段时间内的电流波形。相对于原来的单点聚类，这样处理完之后不仅可以获得更多的电流特征，又可以降低算法对异常点的敏感度，提升其稳定性，还将聚类的数据量减少了多个量级，数据分箱执行流程如图 3.3 所示。

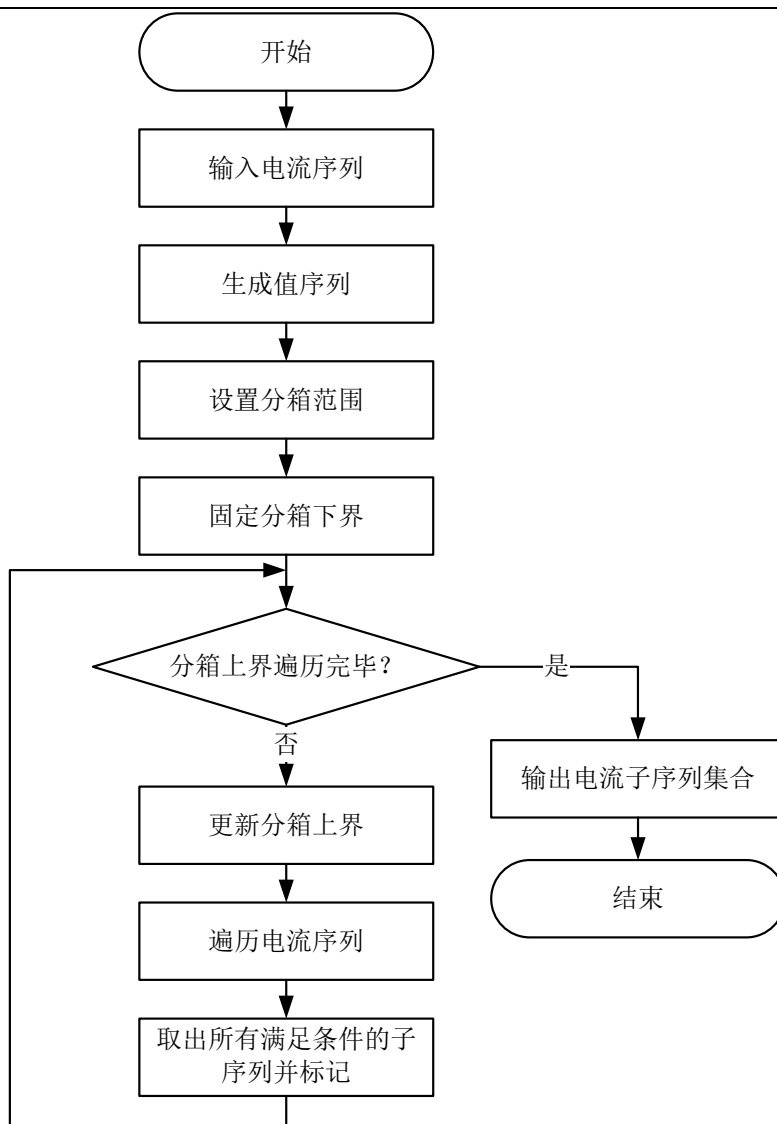


图 3.3 电流数据分箱流程

以差示量热扫描仪为例，原始电流序列中包含两个运行状态下的电流数据，经分箱处理过后的结果如图 3.4 所示。

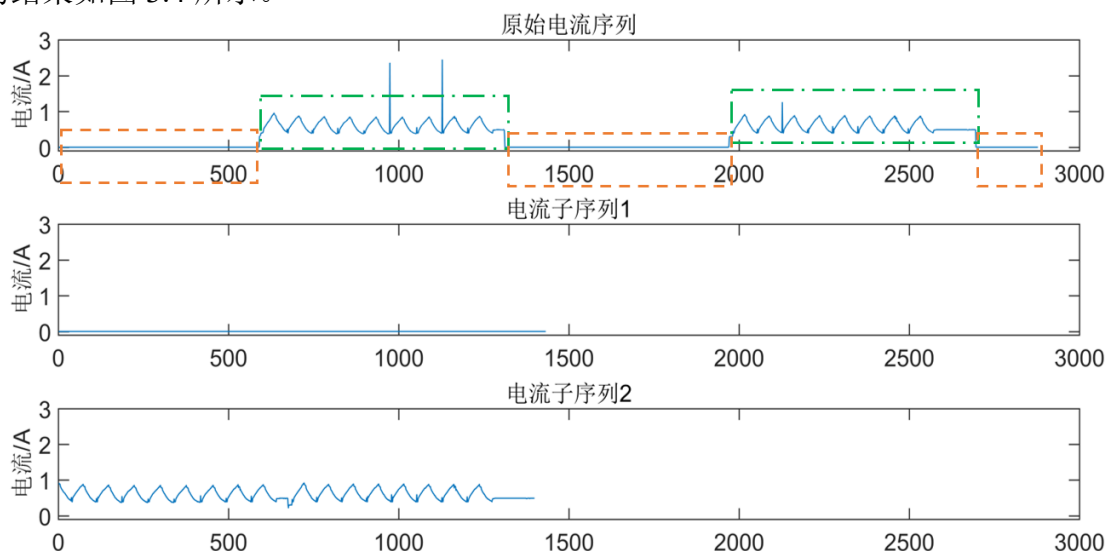


图 3.4 差示量热扫描仪分箱结果图

3.2.2 特征提取

在未分箱之前，电流阈值法仅提取了电流值这一特征来对仪器状态进行区分，导致了一些问题。在对电流数据进行分箱之后，仪器各状态的电流特征除了电流值以外还可以提取到一些其它的典型特征参量，如最大值、有效值、峰均比、短时能量、均值、方差等，这些特征分别从时域上对各状态电流的强度以及能量进行描述。除此之外，还有一些用于描述形状的特征参量，如峰值系数、裕度因子、形状参数等，以及从频域上进行描述一些特征参量。如峰值频率、中心频率等。

通过观察发现，仪器不同状态下的电流数据在时域上的差别大多较为明显，从通用性上考虑，此处选取均值、标准差作为各电流子序列的特征，计算公式如下：

均值：

$$\mu(i) = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad (3.2)$$

标准差：

$$\sigma(i) = \sqrt{\frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \mu(i))^2} \quad (3.3)$$

其中， x_{ij} 为第 i 个电流子序列中第 j 个电流值， $\mu(i)$ 为第 i 个电流子序列的均值， $\sigma(i)$ 为第 i 个电流子序列的标准差。

均值可以表示电流子序列的平均值大小，标准差可以描述电流子序列的离散程度。通过特征变化后，可以将电流子序列集合 S 转换成一个具有 2 维特征的样本集合 $F = \{f_1, f_2, \dots, f_k\}$ ，其中 $f_i = (\mu_i, \sigma_i)^T$ 。为了避免不同特征之间的量级差异过大，导致聚类的结果取决于更大量级的特征，因此此处使用最大最小归一化方法进行处理，将样本的每个特征值都压缩到 0 和 1 之间，归一化公式如下：

$$y_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (3.4)$$

其中， x_i 为归一化前特征， $x_{\min}$ 为所有样本该特征的最小值， $x_{\max}$ 为所有样本该特征的最大值， y_i 为归一化之后的特征。

3.3 K-means 算法分析与改进

3.3.1 K-means 算法原理

K-means 算法是一种基于划分的聚类算法，因其具有实现简单、复杂度低、处理大量数据时也能保证良好伸缩性等优点，在现实多种场景中得到了广泛的应用。

该算法的主要思想是：先随机取出 k 个样本作为初始簇心（初始聚类中心），以相似度的

高低为其它样本分配其所属的簇心，形成初始的 k 个簇。然后重新计算所有簇心，按照上述方法再次分配每个样本实现簇的更新。当各簇的簇心不再变化时，说明聚类过程已经完成，此时的簇心就是最终的簇心^[51]。具体而言，K-means 算法的流程可以用图 3.5 进行描述，主要由两个步骤组成：

1、选出 k 个初始的簇心。 k 值往往按照人工经验或者是现有先验知识进行设置，这也决定了最终聚类结果中会存在几个簇。虽然初始簇心往往都不是最佳簇心，但会影响到后续簇变化的趋势。

2、不断迭代簇心直至稳定。获得了初始的簇心后需要分配其它非簇心样本从而形成簇，再通过各簇样本来更新簇心的位置，并对所有非簇心样本进行再分配直到簇心样本不再变化。由于样本的归属取决于样本与簇心的相似度，因此，相似度定义决定了后续簇以及簇心的变化过程，并对最终聚类的效果造成影响。

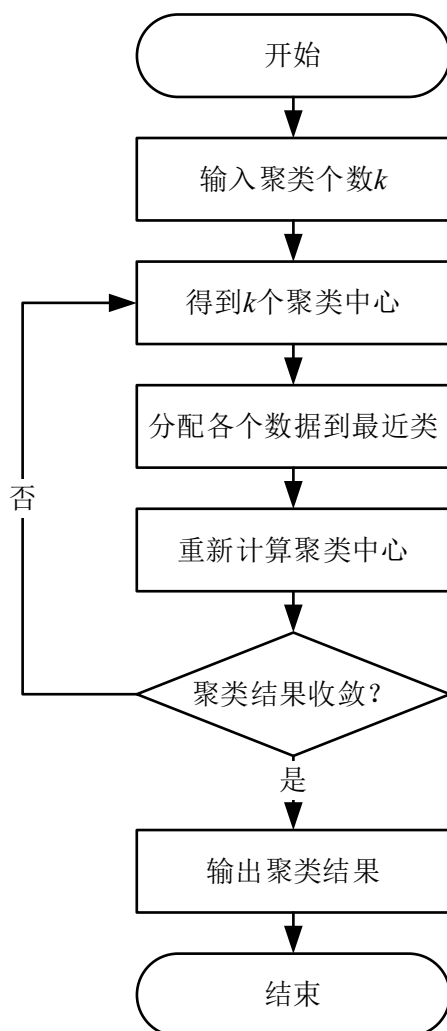


图 3.5 K-means 聚类流程

样本间相似度通常采用距离或者相似性系数来进行描述。常见距离有欧式（Euclidean）距离、曼哈顿（Manhattan）距离、切比雪夫（Chebyshev）距离等。这些距离是对样本在特征空间内距离的表述，从本质上来说都可以由闵可夫斯基（Minkowski）距离变形得到。例如，

对于包含 m 个特征的任意样本 $x = \{x_1, x_2, \dots, x_m\}$ 之间的闵可夫斯基的公式表示为:

$$d(x_i, x_j) = \left(\sum_{k=1}^m |x_{ik} - x_{jk}|^q \right)^{\frac{1}{q}}, q \geq 1, (i, j = 1, 2, \dots, n) \quad (3.5)$$

当 q 值为 1 时变形为曼哈顿距离, 式 (3.5) 可以写作:

$$d(x_i, x_j) = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (3.6)$$

当 q 值为 2 时变形为欧式距离, 式 (3.5) 可以写作:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2} \quad (3.7)$$

当 q 无穷时变形为切比雪夫距离, 式 (3.5) 可以写作:

$$d(x_i, x_j) = \max_{1 \leq k \leq m} |x_{ik} - x_{jk}| \quad (3.8)$$

相似度系数是对样本相似度的另一种表述, 取值范围为-1 到 1, 样本间相似度系数的绝对值越大表明其相关性越大。常见的有余弦相似度及相关系数等^[36], 这两者的计算公式如下所示:

余弦相似度:

$$C(x_i, x_j) = \frac{|\sum_{k=1}^m x_{ik} x_{jk}|}{\sqrt{\left(\sum_{k=1}^m x_{ik}^2\right) \left(\sum_{k=1}^m x_{jk}^2\right)}} \quad (3.9)$$

相关系数:

$$\gamma_{ij} = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\left\{ \left[\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2 \right] \left[\sum_{k=1}^m (x_{jk} - \bar{x}_j)^2 \right] \right\}^{\frac{1}{2}}} \quad (3.10)$$

若待分析的是时间序列, 还可以使用动态时间规整算法^[52] (Dynamic Time Warping, DTW) 来判断其相似度。该算法基本思想是利用动态规划迭代寻找全局最优解, 最初用于解决语音信号的模板匹配问题。对于两段时间序列 $X = \{x_1, x_2, \dots, x_m\}$ 和 $Y = \{y_1, y_2, \dots, y_n\}$, 相似度的计算方式为: 首先构造一个大小为 $n \times m$ 的矩阵, 求出任意坐标 (i, j) 处的规整距离 $DP(i, j)$, 然后找到两段序列整体的最短规整距离作为相似度。两段序列任意处的规整距离计算公式如下:

$$DP(i, j) = D(i, j) + \min \begin{bmatrix} DP(i-1, j) \\ DP(i-1, j-1) \\ DP(i, j-1) \end{bmatrix} \quad (3.11)$$

通常 $D(i, j)$ 取值为序列 X 第 i 个元素 x_i 与序列 Y 第 j 个元素 y_j 的欧式距离, 动态规划搜索

从坐标(1, 1)开始，到坐标(m, n)结束，在 i, j 都单调递增的约束下，寻找一条最优路径。

以序列 $A = \{1, 1, 3, 3, 2, 4\}$ 与序列 $B = \{1, 3, 2, 2, 4, 4\}$ 为例，这两个序列的动态规整距离矩阵以及求解的路径如图 3.6 所示。

		A(1)					
B(1)		0	0	2	2	1	3
	2	2	0	0	1	1	
	1	1	1	1	0	2	
	1	1	1	1	0	2	
	3	3	1	1	2	0	
	3	3	1	1	2	0	

图 3.6 序列 A、B 的动态时间规整求解路径

其中，颜色较深路径为动态时间规整的最优路径，颜色较浅的为传统信号计算距离的路径。序列 A、B 的距离按照传统方法计算的结果为 6，而经过动态时间规整计算距离为 0。

总的来说，欧式距离目前是 K-means 中最常使用的相似度计算方法，一来欧式距离可以直观描述样本空间距离，二来欧氏距离不会因为样本的特征空间平移或旋转而发生改变。

3.3.2 K-means 常见缺陷

由 K-means 的原理可知，该算法存在的缺陷主要可以概括为以下两点：

1、聚类开始时初始簇心选择具有随机性，常常出现同一类型的样本被同时选为簇心。而且，即使聚类簇心是不同类型的样本，也会出现初始聚类中心不是最佳导致算法陷入局部最优，导致结果与实际情况不符。

2、需要提前知道数据集实际包含的种类，即 k 值。尤其是在当前场景下，如果只有几类仪器的话可以通过一定的先验知识设置 k 值，但随着仪器种类增多，再由人工为每台仪器设置 k 值就不太现实了，因此需要系统在执行过程中能够自行找到最佳 k 值。

为了使算法可以由系统执行，且系统执行时也可以得到质量较高的电流分类模型，尽量减少后续人工调整，因此接下来对算法进行改进。

3.3.3 K-means 算法改进

3.3.3.1 初始聚类中心选取策略

针对初始聚类中心随机选取的问题, D.Arthur 等人在提出 K-means++ 算法中做出了改进。其思想为: 在已有 n 个聚类中心的情况下, 在选择第 $n+1$ 个聚类中心时, 与前 n 个聚类中心距离越远的样本成为新的聚类中心的可能性就越高^[53]。当然, 该方法在选取第一个聚类中心时还是随机的。Alex Rodriguze 和 Alessandro Laio 在 2014 发表的文章^[54]中提出, 适合作为聚类中心的对象的局部密度往往较大, 与等高线的形式类似。

受到上述思想的启发, 此处首先对初始聚类中心的选取过程进行改进。由于往往仪器会在某个状态下持续很长的时间, 因此, 对于第一个聚类中心的选择, 可以根据每个电流子序列的长度与总序列长度的占比来考虑, 一段电流子序列的持续时间越长, 那么它越可能成为一个优质的聚类中心, 如下式所示:

$$p_i = \frac{\text{len}(S_i)}{\text{len}(S)} \quad (3.12)$$

其中, p_i 是电流子序列 S_i 被选为第一个聚类中心的概率, $\text{len}(S_i)$ 为该电流子序列长度, $\text{len}(S)$ 为所有电流子序列长度之和。

后续聚类中心的选取, 除了按照其它样本与聚类中心的距离作为参考依据以外, 也额外引入样本对应的电流序列长度作为参考依据, 然后将长度占比与聚类中心的距离占比按照一定的权重进行融合, 形成一个新的被选概率。按照轮盘法, 第 j 个样本成为第 $n+1$ 个初始聚类中心的概率计算公式如下:

$$p_j = w \frac{\min d(j, n)}{\sum_{k=1}^m \min d(k, n)} + (1-w) \frac{\text{len}(S_j)}{\text{len}(S)} \quad (3.13)$$

其中 m 为样本的个数, w 为融合权重, $\min d(j, n)$ 为样本 j 到前 n 个聚类中心的最小距离, 所述距离计算公式如下:

$$d(j, n) = \sqrt{\frac{\sum_{i=1}^{\text{dim}} (f_i - f_n)^2}{\text{dim}}} \quad (3.14)$$

其中 dim 为样本的特征维度。

3.3.3.2 最优聚类结果求解策略

按照 3.3.3.1 小节聚类中心选取方式, 可以使得初始聚类中心的质量更高, 但还是存在一定的概率使得两个属于同类的样本被同时选为聚类中心。这样一来, 即使聚类的过程优化得再合理, 聚类的结果依旧偏离实际的, 可能需要进行多次的重复聚类才能得到更好的聚类结果。如何从多次聚类结果中选取最优的结果, 通过人工观察的方式显然是不能令人满意, 因

此，就需要引入一种聚类结果指标来判断当前结果是否已经达到全局最佳，常见的评价指标包括 CHI 指标、DBI 指标、SSE 指标和轮廓系数 SC 等^[55]。此处，以图 3.7 所示的 4 类模拟数据，以 SSE、SC 这两种指标为例说明评价指标的具体作用。

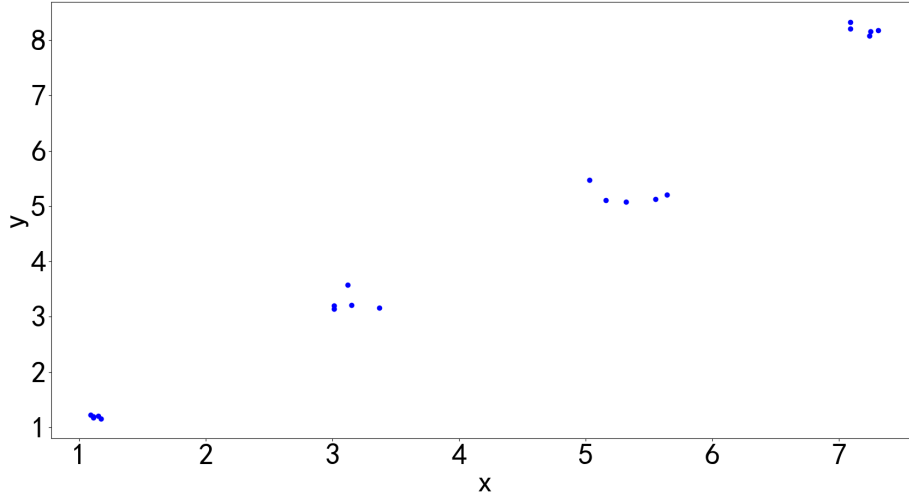


图 3.7 模拟数据

SSE 指标度量聚类结果好坏的方式为：计算每个样本与其所属簇的簇心距离，然后对这些距离进行求和得到 I_{SSE} ，公式如下所示。

$$I_{SSE} = \sum_{i=1}^k \sum_{x \in c_i} d^2(c_i, x) \quad (3.15)$$

轮廓系数评价依据可以总结为一句话：“簇内样本距离越近，不同簇样本距离越远”。也就是说簇内密度越高，簇间越分散，那么就认为聚类的效果就越好。样本 i 的轮廓系数计算公式为：

$$S(i) = \frac{(b(i) - a(i))}{\max(a(i), b(i))} \quad (3.16)$$

其中， $a(i)$ 为样本 i 与同簇其它样本的平均距离， $b(i)$ 为样本 i 与不同簇样本距离的最小值。

每个样本的轮廓系数都按照式 (3.16) 进行求解，然后取平均值作为当前聚类结果的轮廓系数，求解公式如下：

$$\bar{S} = \frac{1}{k} \sum_{i=1}^k S(i) \quad (3.17)$$

最终求解的轮廓系数的值必然在 $[-1, 1]$ 的范围内，当轮廓系数越接近 1，说明聚类结果中各簇的簇内紧凑度和簇间分散度达到一个最佳的状态，反映出来的效果越好。反之，轮廓系数越接近 -1 则说明簇内结构越分散，簇间区分度不高，聚类效果越差。

在已知样本种类为 4 的情况下，算法执行 30 次，分别用上述两种指标对结果进行评价。两种指标变化如图 3.8、图 3.9 所示。

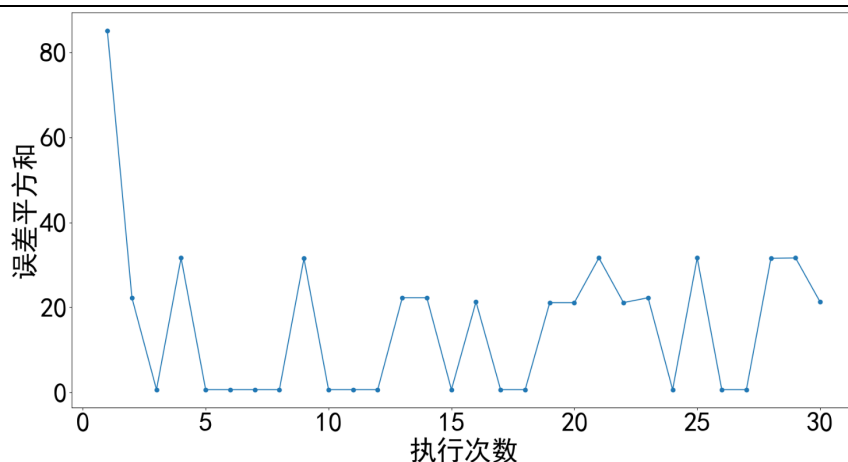


图 3.8 执行 30 次的误差平方和

在第 3、5、6、7、8、10、11、12、15、17、18、24、26、27 次执行时，SSE 都为最小值，此时得到的聚类结果是所有结果中最佳的。

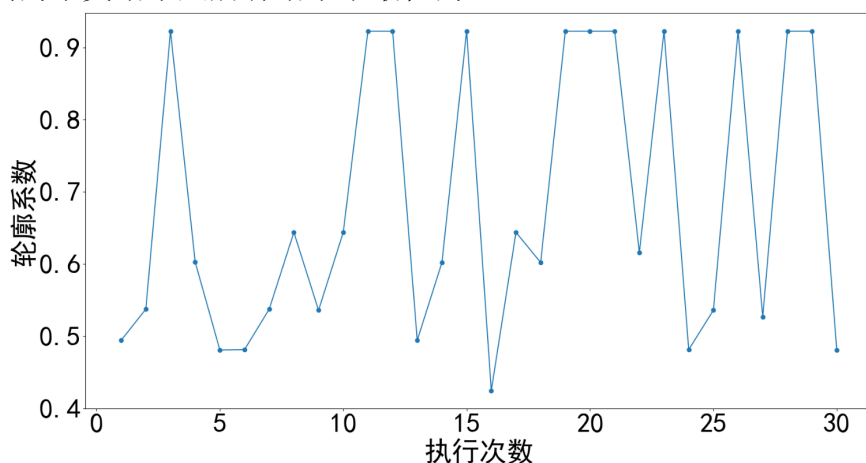


图 3.9 执行 30 次的轮廓系数

在第 3、11、12、15、19、20、21、23、26、28、29 次执行时，轮廓系数取得最大值，说明得到的聚类结果是所有结果中最佳的。

最终，通过两种方式求得的佳聚类结果一致，如图 3.10 所示。

k=4 最佳聚类结果

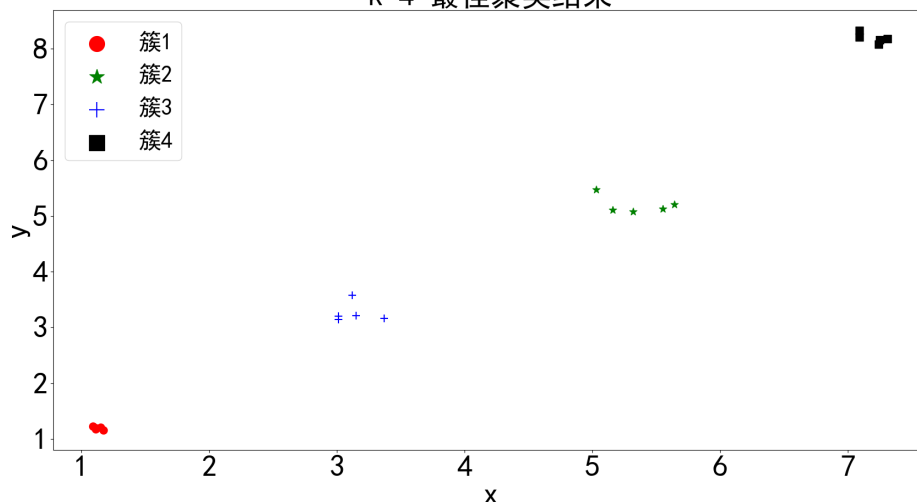


图 3.10 最佳聚类结果

同样的，在缺少一定先验知识的时候，最佳的聚类个数也可以借助评价系数求解。具体实施方式为：将聚类的个数 k 设定在一个范围内，然后从小到大依次选取 k 值进行聚类，每个 k 值都需要迭代多次，最后取出所有聚类结果中评价指标最优的那一个作为最终聚类的结果。以模拟数据为例， k 值范围设置为 2-6，每个 k 值执行 30 次，取最优结果的评价结果作为当前 k 值的结果，效果如图 3.11、图 3.12 所示。

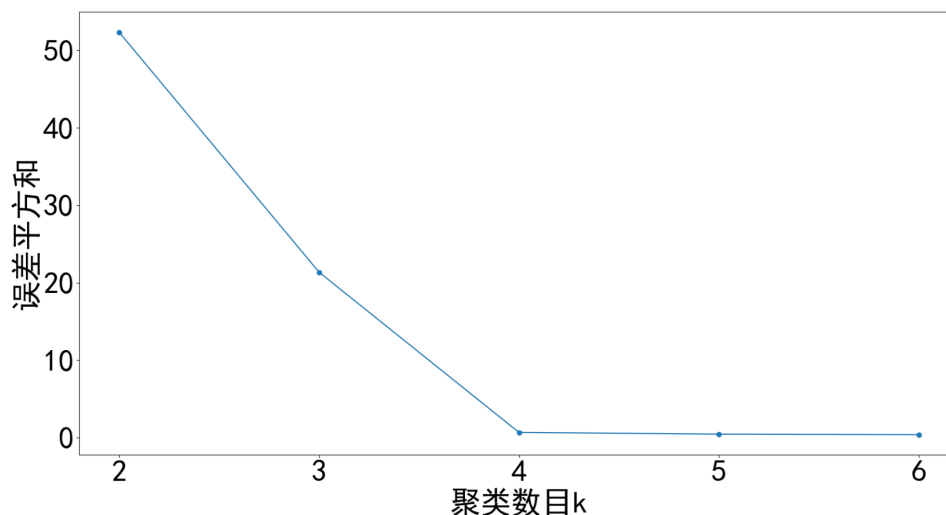


图 3.11 误差平方和随 k 值变化曲线

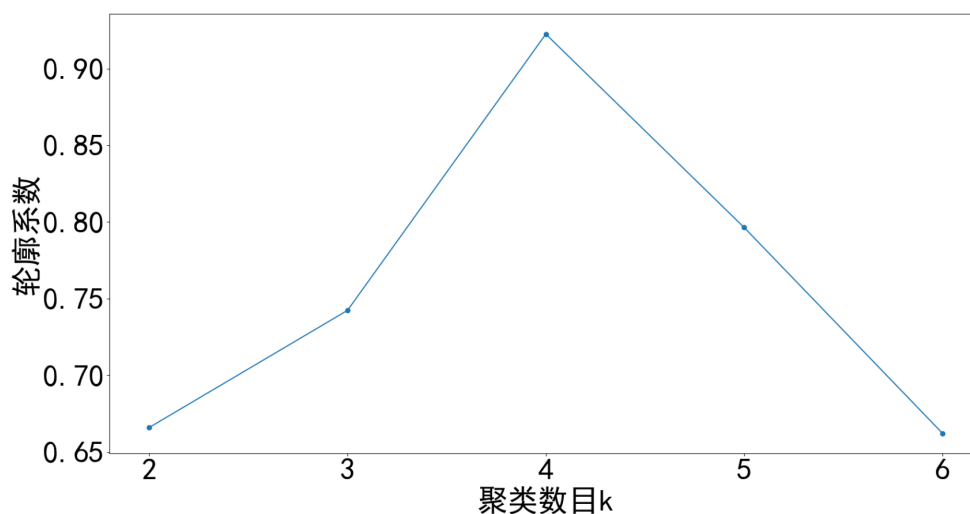


图 3.12 轮廓系数随 k 值变化曲线

可以看到，当 $k=4$ 时，SSE 曲线出现了明显的转折点，再增大 k 值聚类效果也不会有很大的变化，按照肘方法的思路此时的 k 值就是最佳的了。同样的，在 $k=4$ 时，轮廓系数获取了最大值，因此同样求得最佳 k 值也为 4，这两种方法求出的 k 值都符合实际。在实际使用中，由于仪器的运行状态数仅定义为关机、待机、工作这三种，且部分仪器状态仅包括关机、工作这两种，因此 k 的个数不会太多，通过这些评价指标可以避免每台仪器的 k 值需要人工设置。

从计算速度上来看，SSE 指标最快，轮廓系数较慢。但从聚类效果上来看，当簇划分得越细，SSE 指标的值就会越小，往往需要人工寻找 SSE 变化曲线的“肘部”。而轮廓系数因为

有一个离散度限制，只要寻找最大值即可，更适用在系统中应用。所以，从自动化的角度考虑，此处选择轮廓系数来作为聚类效果评价标准，求解最佳聚类结果的具体执行流程如图 3.13 所示。

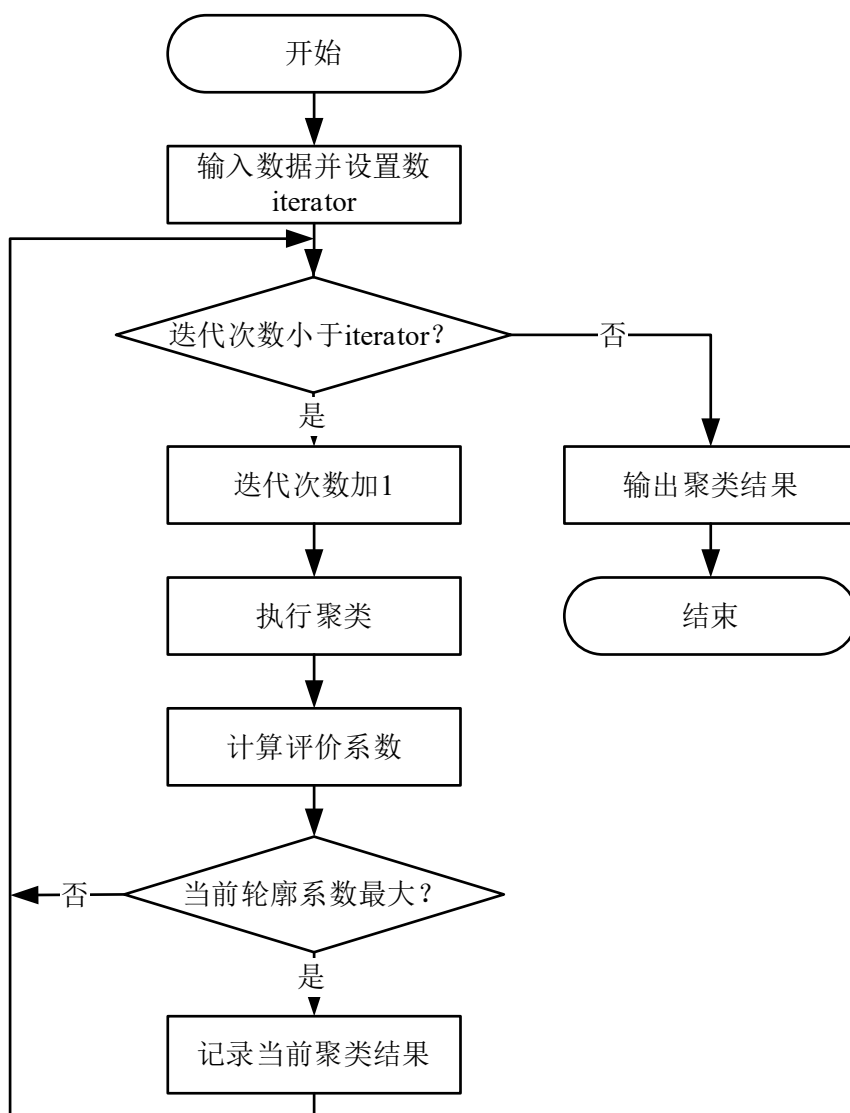


图 3.13 最佳聚类结果选取流程

详细来说就是在聚类过程中设置一个当前迭代次数和最大迭代次数，每次聚类结束后都使当前迭代次数加 1，并计算当前聚类结果的轮廓系数，若该值大于历史所有值，那么记录当前的聚类结果，否则进行下一次聚类。等到迭代结束后，记录的聚类结果就认为是最佳聚类结果。

3.4 基于分箱与改进 K-means 的机时计算流程

由于还有许多仪器不适用电流阈值法，而其它方法较为复杂，因此本章提出了分箱与改进 K-means 方法来降低成本、提升效率。首先针对电流特点设计了一种分箱算法，不仅可以电流的波形中去发现不同状态下更多特征的差异，而且大大减少了聚类的样本个数。该方

法的详细的流程如图 3.14 所示。

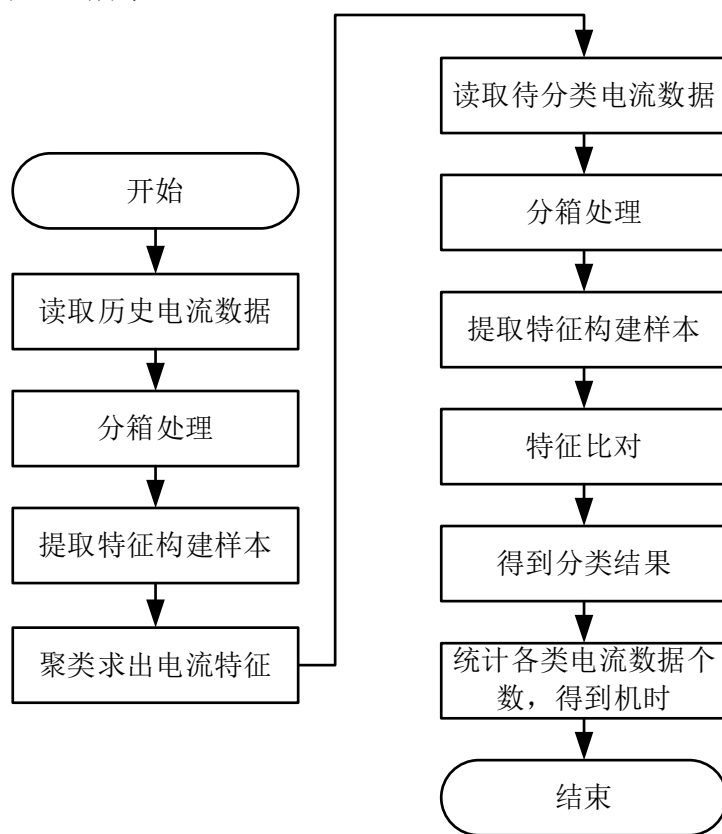


图 3.14 基于分箱与改进 K-means 的机时计算执行流程

具体实施过程如下：

- 1、输入历史电流数据，分箱处理成电流子序列，提取均值、标准差后进行归一化处理，生成样本集。
- 2、通过改进 K-means 算法聚类得到每个簇心样本的电流特征，将所述电流特征反归一化处理后按照均值从小到大的顺序打上标签后保存，得到电流分类模型。
- 3、对于一段未知电流数据，通过分箱处理后，剩余那些未被分入箱中数据也需要按照同样的方式转换为样本，但不进行归一化处理。
- 4、使用电流分类模型对步骤 3 所得样本集进行分类，分类过程的本质是比对样本与各状态电流特征的相似度，得出各样本对应状态标签，每个样本中所有电流数据的标签与该样本相同，完成对电流数据的分类。
- 5、统计各标签电流的个数，得到机时。

3.5 实验测试

本章涉及到的所有算法均采用 Java 编写，开发工具为 IntelliJ IDEA 2021.1.1 x64。

3.5.1 实验准备

为验证基于分箱与改进 K-means 的机时计算方法（下文简称本章算法）的效果，选取了 15 个大类的仪器共 105 台作为样本，以其中 6 台典型的仪器为例对算法执行过程进行展开说明，相关信息如表 3.1 所示。

表 3.1 实验数据信息

序号	IMEI	仪器名称	状态种类
1	864823041535253	多功能细胞分析系统	3
2	869858032105018	凯氏定氮仪	3
3	869858032092471	气相色谱仪	3
4	869858032109051	万能材料试验机	2
5	864823041521121	液相色谱仪	2
6	869858032081045	网络分析仪	3

接下来对表 3.1 中各仪器不同状态下的电流波形特点进行介绍，各仪器电流波形如图 3.15 所示。

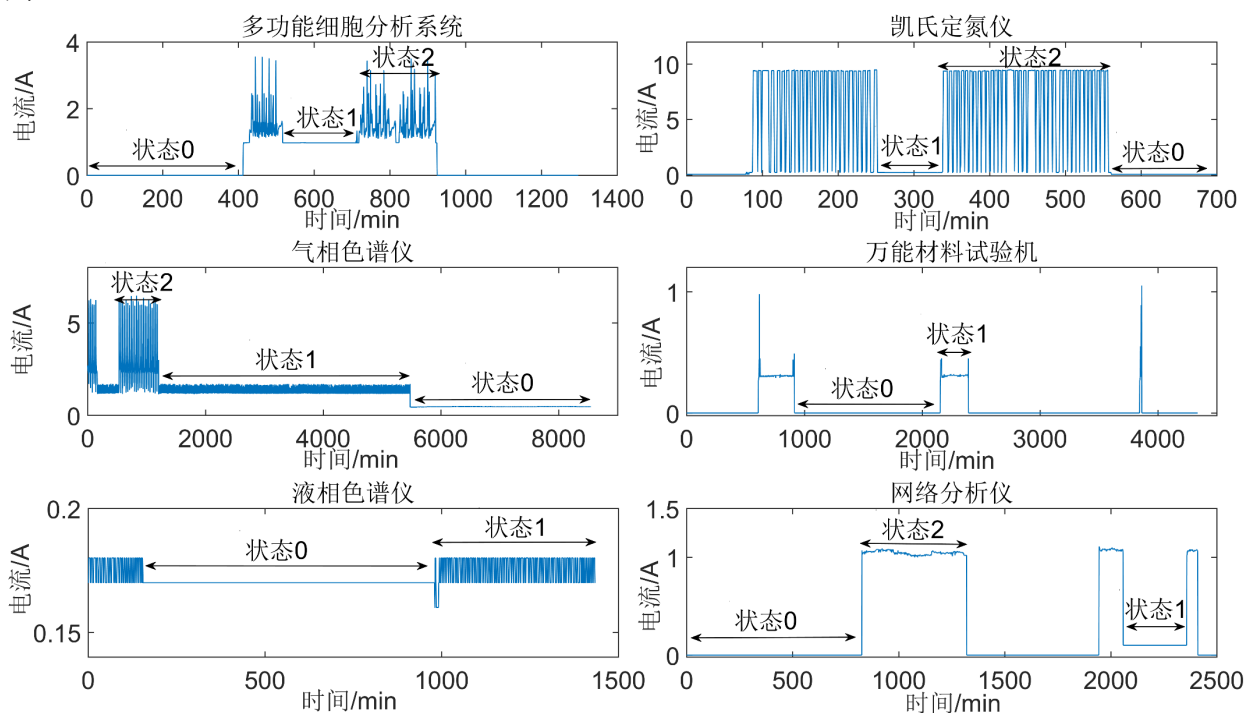


图 3.15 各仪器电流波形

其中，横轴代表电流数据出现的时序，具体的：

- 1、多功能细胞分析系统共包含 3 状态，其中第 0, 1 种状态的电流值为 0A, 0.97A，基本保持不变，第 2 种状态电流值在 1.1A 至 3.9A 间上下波动；
- 2、凯氏定氮仪包含 3 种状态，第 0, 1 种状态电流值分别为 0.02A, 0.18A，基本保持不变，第 2 种状态电流值在 0.18A 至 9.8A 间上下波动；
- 3、气相色谱仪总共包含 3 种状态，其中第 0 种状态电流值为 0.46A 左右，几乎不变，第

1 种状态电流值在 1.15A 至 1.81A 间波动，第 2 种状态电流值在 1.15 至 6.5A 间上下波动；

4、万能材料试验机共包含 2 种状态，分别是第 0 种状态下值为 0A 的电流以及第 1 种状态下值为 0.3A 的电流，在仪器状态切换时电流可能存在跳变；

5、液相色谱仪包含 2 种状态，电流值在 0.16A 至 0.18A 间上下波动，但第 0 种状态电流稳定，而第 1 种状态的电流值会频繁变化；

6、网络分析仪包含 3 种状态，第 0 种状态电流值为 0A，第 1 种状态电流值为 0.1A，第 2 种状态电流值在 0.95A 至 1.04A 之间波动。

为了更好地体现本章算法的优势，实验中以电流阈值法作为参照进行对比。其中，电流阈值法的阈值采用前文所述聚类方案进行求解，聚类算法采用 K-means++，本章算法中 T 值设为 60， w 值设为 0.7，两者 k 值按照实际输入，每台仪器重复执行各算法 15 次来获取最佳结果。测试数据为实际采集数据，聚类结果中每个电流数据的标签代表其所属的状态，用 0、1、2 进行表示。

3.5.2 聚类结果与分析

为了使各算法执行过程更加直观，下列各图中均包括原始电流数据、K-means++ 聚类结果、分箱处理后重组的电流数据及本章算法聚类结果。

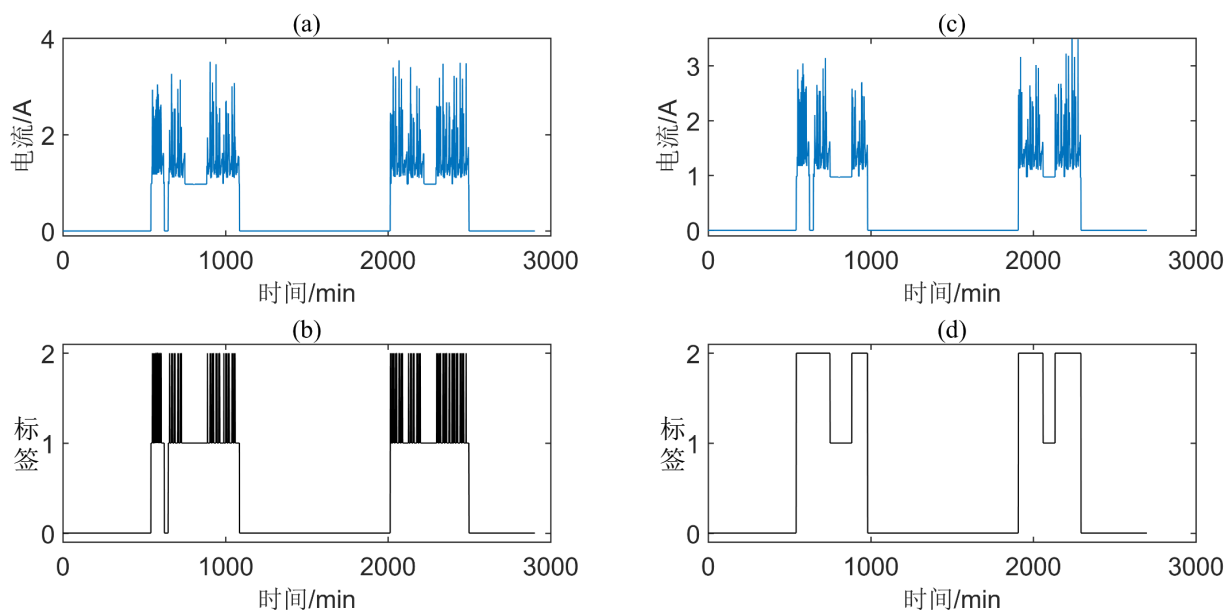


图 3.16 两种算法对原始电流数据处理、聚类的结果

图 3.16 展示了从多功能细胞分析系统采集的电流数据经两种算法处理、聚类后的结果。图 3.16(a)为原始电流数据，图 3.16(b)为使用 K-means++ 算法聚类结果，图 3.16(c)、(d)为本章算法分箱处理后的电流数据与聚类结果。聚类结果中的标签描述了对应时刻电流数据对应的状态。由于该仪器状态 2 的部分电流与状态 1 电流更接近，使用 K-means++ 算法聚类时这部分电流数据被错分到状态 1 中了，故此处电流的标签在 1、2 之间频繁变化，导致图 3.16(b)中该部分电流聚类结果与实际偏差较大，而在本章算法则不会被此影响，聚类结果与实际情况

基本一致。

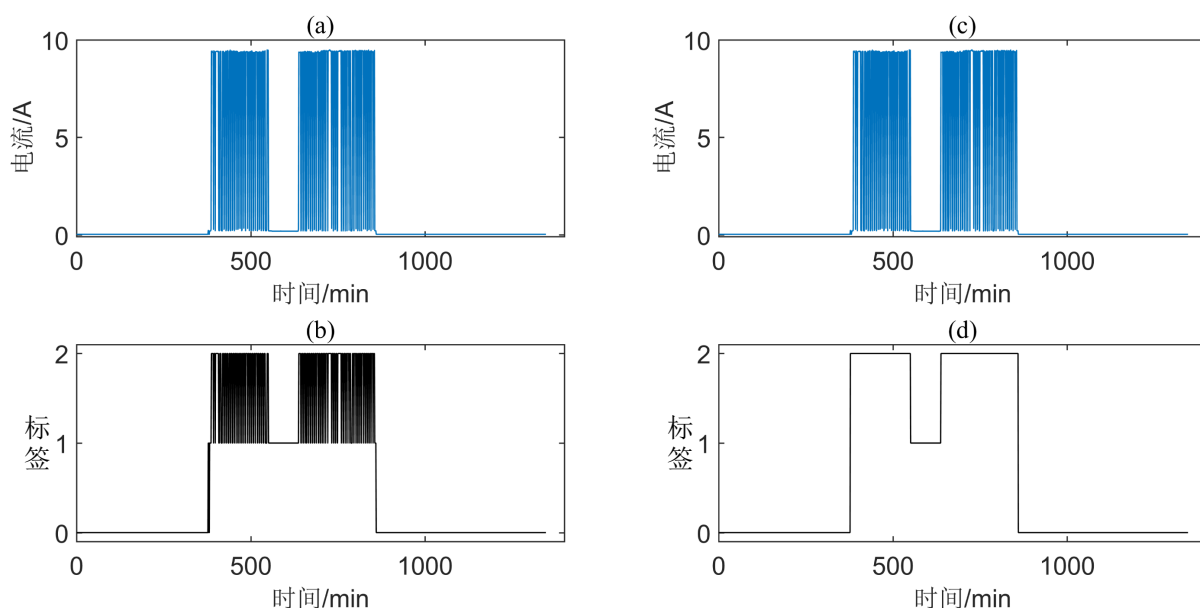


图 3.17 两种算法对原始电流数据处理、聚类的结果

图 3.17 展示了从凯氏定氮仪采集的电流数据经两种算法处理后的结果。图 3.17(a)为原始电流数据，图 3.17(b)为使用 K-means++算法聚类结果，图 3.17(c)、(d)为本章算法分箱处理后的电流数据与聚类结果。由于该仪器状态 1 对应的电流值为 0.02A，而状态 2 的电流值也存在 0.02A，因此使用 K-means++聚类时这部分电流数据将会被分为一个簇，造成误判。本章算法聚类结果更符合实际，求出的电流分类标准质量更好。

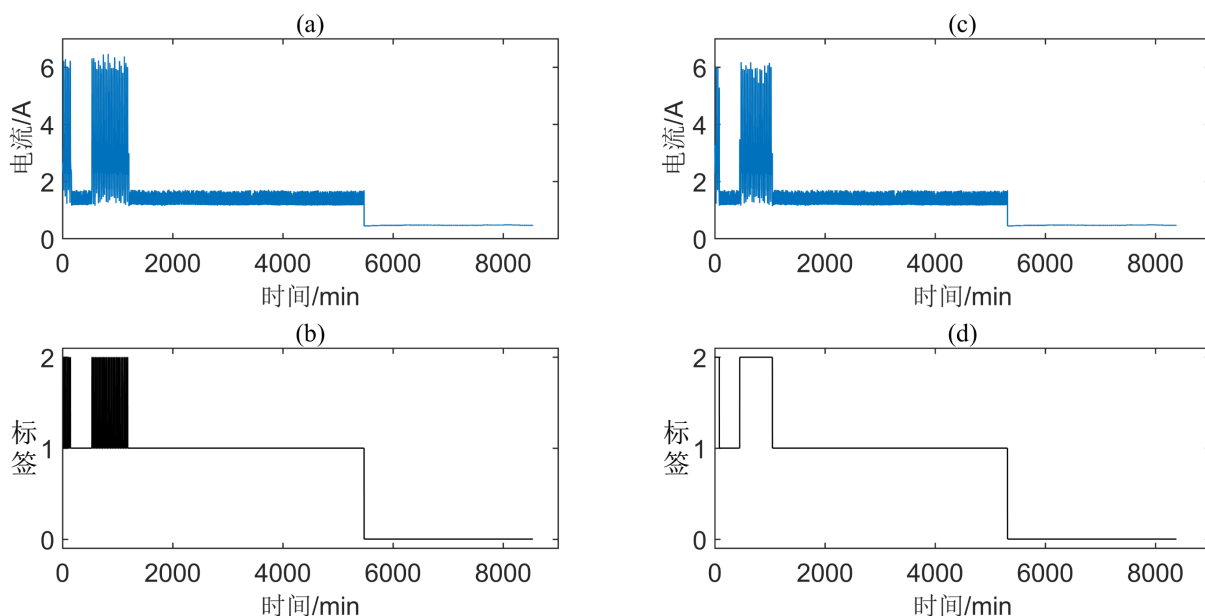


图 3.18 两种算法对原始电流数据处理、聚类的结果

图 3.18 展示了从气相色谱仪采集的电流数据经两种算法处理后的结果。图 3.18(a)为原始电流数据，图 3.18(b)为使用 K-means++算法聚类结果，图 3.18(c)、(d)为本章算法分箱处理后的电流数据与聚类结果。可以看到使用 K-means++算法同样区分状态 1、2 的电流数据时有大

量的误判，而本章算法的效果则好得多。

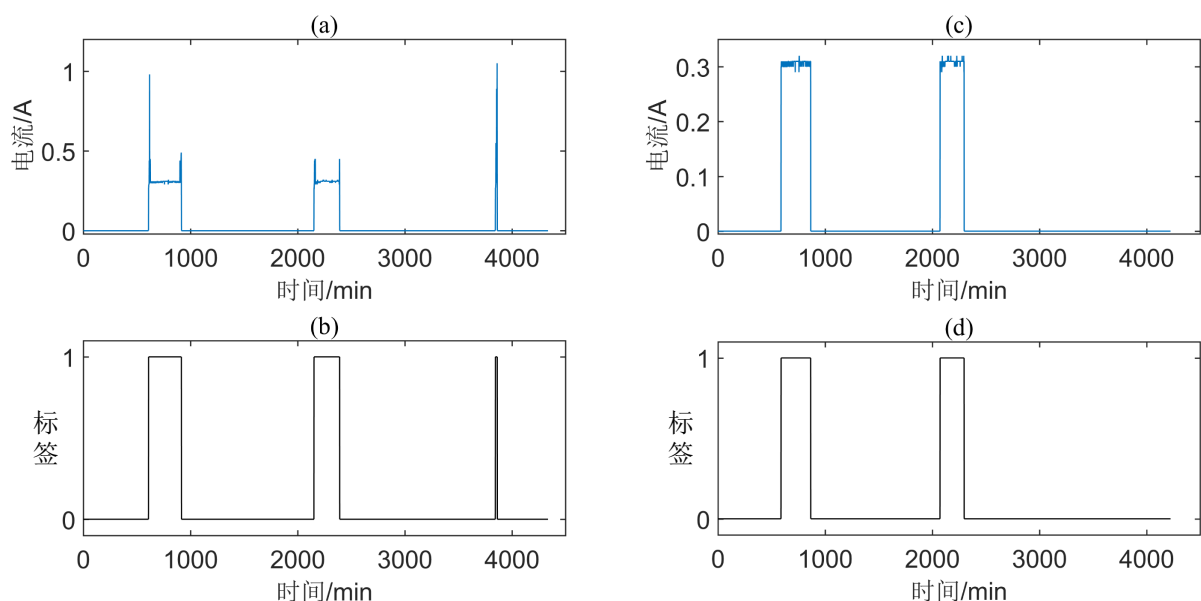


图 3.19 两种算法对原始电流数据处理、聚类的结果

图 3.19 展示了从万能材料试验机采集的电流数据经两种算法处理后的结果。图 3.19(a)为原始电流数据，图 3.19(b)为使用 K-means++ 算法聚类结果，图 3.19(c)、(d)为本章算法分箱处理后的电流数据与聚类结果，两种方法的聚类结果都较为符合实际情况。

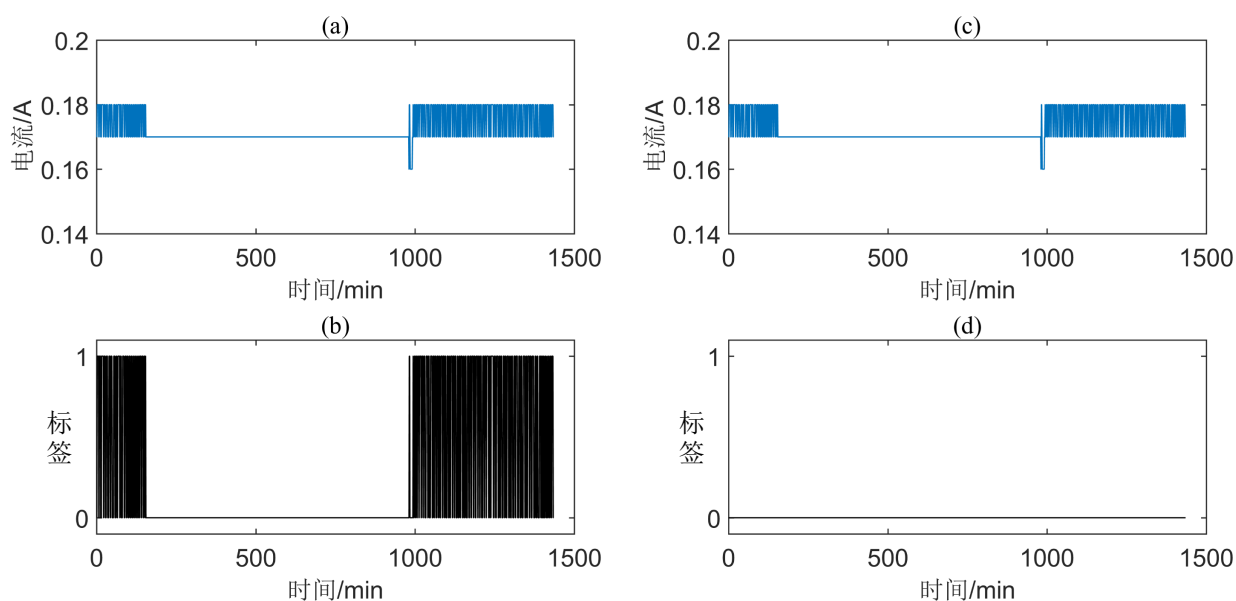


图 3.20 两种算法对原始电流数据处理、聚类的结果

图 3.20 展示了从液相色谱仪采集的电流数据经两种方法处理及聚类后的结果。图 3.20(a)为原始电流数据，图 3.20(b)为使用 K-means++ 算法聚类结果，图 3.20(c)、(d)为本章算法分箱处理后的电流数据与聚类结果。使用 K-means++ 算法出现的问题与凯氏定氮仪与多功能细胞分析系统类似，存在大量的误判。本章算法由于分箱的原因，导致这些电流数据都被放入一个箱中，所以无法正确聚类，因此该仪器被判定为只有一种状态。

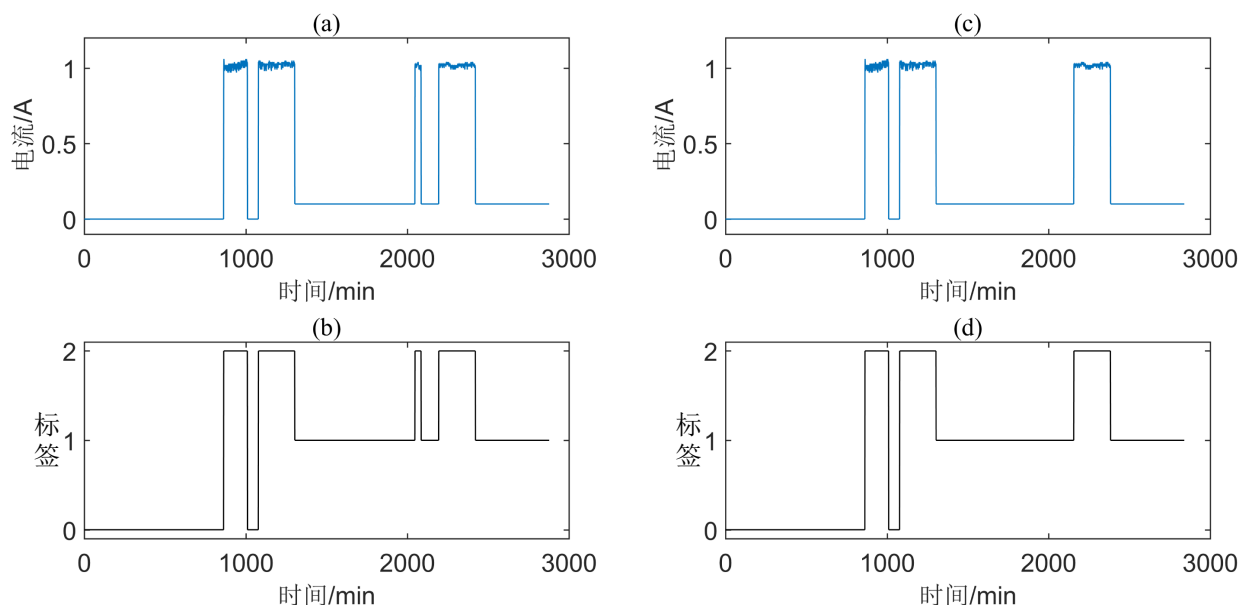


图 3.21 两种算法对原始电流数据处理、聚类的结果

图 3.21 展示了从网络分析仪采集的电流数据经两种算法处理后的结果。图 3.21(a)为原始电流数据，图 3.21(b)为使用 K-means++ 算法聚类结果，图 3.21(c)、(d)为本章算法分箱处理还原的电流数据与聚类结果，两种方法都可以得到较高质量的电流分类标准。

当电流区分明显且波动较小时，K-means++ 算法和本章算法的聚类效果都不错。当仪器电流波动较大，就体现出了差异，比起仅针对单个点的聚类，本章算法经过分箱处理后再聚类的效果要好的多。两种方法构建的电流分类模型实质是其中的电流特征，如表 3.2 所示。

表 3.2 电流特征

仪器编号	标签	电流阈值法（电流值）	本章算法（均值, 标准差）
1	0	0.00	(0.00, 0.00)
	1	1.25	(0.97, 0.00)
	2	2.42	(1.46, 0.47)
2	0	0.02	(0.02, 0.00)
	1	0.19	(0.18, 0.01)
	2	9.54	(5.01, 4.55)
3	0	0.47	(0.46, 0.01)
	1	1.32	(1.33, 0.17)
	2	3.07	(2.77, 0.94)
4	0	0.00	(0.00, 0.00)
	1	0.31	(0.30, 0.02)
	0	0.17	(0.17, 0.01)
5	1	0.18	-
	0	0.00	(0.00, 0.00)
	1	0.10	(0.10, 0.00)
6	2	0.98	(0.99, 0.01)

3.5.3 电流分类对比及分析

本章算法生成电流分类模型时由于对样本质量有一定要求，会剔除一部分数据，而电流分类应用时由于要对所有的电流数据贴上标签，因此被剔除部分电流数据也需要按照连续的时序也形成多个电流子序列进行处理，为了减小分箱带来的误差，此时将分箱参数设置为 $T = 30$ 。

电流阈值法获取电流特征后，状态数为 2 的仪器以其平均值作为阈值，而状态数为 3 的仪器的取值规则为：将电流特征值按照从小到大的顺序进行排列，取每相邻两个值的平均为阈值，假设状态 $state = [0, 1, 2]$ ，对应的电流特征值 $Value = [a, b, c]$ ，那么所求阈值 $t = [(a+b)/2, (b+c)/2]$ ，其类别判定规则如下所示：

$$state_i = \begin{cases} value_i \leq (a+b)/2, i \in 0 \\ (a+b)/2 < value_i \leq (b+c)/2, i \in 1 \\ value_i > (b+c)/2, i \in 2 \end{cases} \quad (3.18)$$

其中， $state_i$ 、 $value_i$ 分别为第 i 分钟电流数据对应的状态以及电流值。

以天为单位，取各仪器共一周的电流数据进行测试，由于各状态的电流数据量级差别不大，因此测试结果用准确率进行评价。基于篇幅考虑，接下来选取其中 1 天的测试结果进行展示说明。

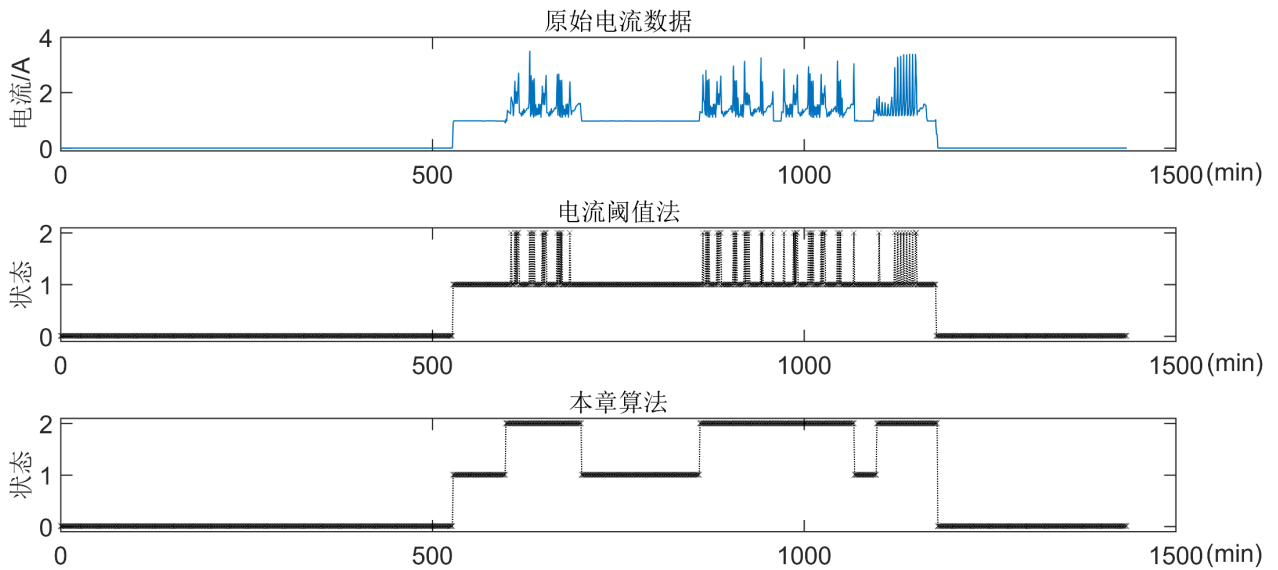


图 3.22 1 号仪器原始电流数据及分类结果

图 3.22 展示了 1 号仪器的原始电流数据以及两种算法对这些数据的分类结果。其中，对于状态 0 的电流两者都可以正确分出，而对于状态 1、2 电流的区分上两者表现出了差异，电流阈值法将大量原本属于状态 2 的电流分到状态 1 中了。这是因为电流阈值法中相关的阈值过大，无法区分各状态重叠部分的电流。而本章算法是对一段电流的均值和标准差进行判别的，所以效果更好。

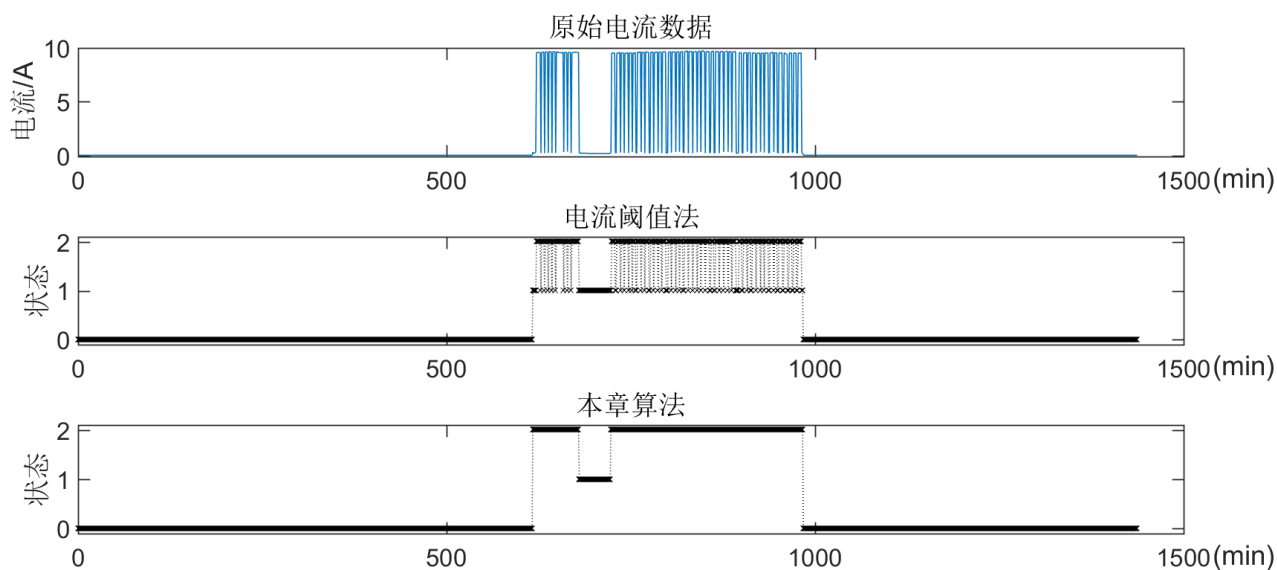


图 3.23 2 号仪器原始电流数据及分类结果

图 3.23 展示了 2 号仪器的原始电流数据以及两种算法对这些数据的分类结果。同样的对于状态 0 的电流两者的性能基本一致，也是在区分状态 1、2 的电流上表现出了差异，差异原因同仪器 1。不过本章算法由于分箱的原因，剩余那些长度不足 30 的电流序列默认在一个箱中的前提导致状态 1 的部分电流数据与状态 2 的电流数据被分至同一箱中，进而导致了一些误判。

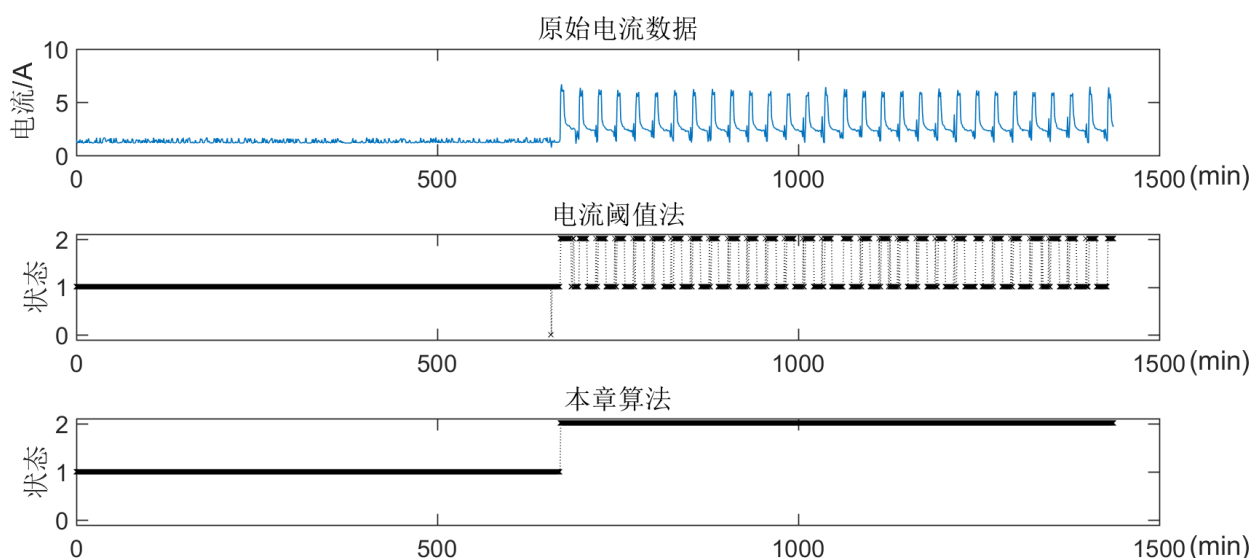


图 3.24 3 号仪器原始电流数据及两种算法分类结果

图 3.24 展示了 3 号仪器的原始电流数据以及两种算法对这些数据的分类结果。同样的，主要也是在区分状态 1、2 的电流上表现出了差异。最根本的原因还是因为该仪器工作时部分电流与待机时的电流值很接近，对电流阈值法来说这两种状态下的电流没有什么区别，即使人工调整到最佳阈值以后还是会出现误判。而本章算法从一段电流的特征进行考量，可以很好区分出不同状态下的电流。

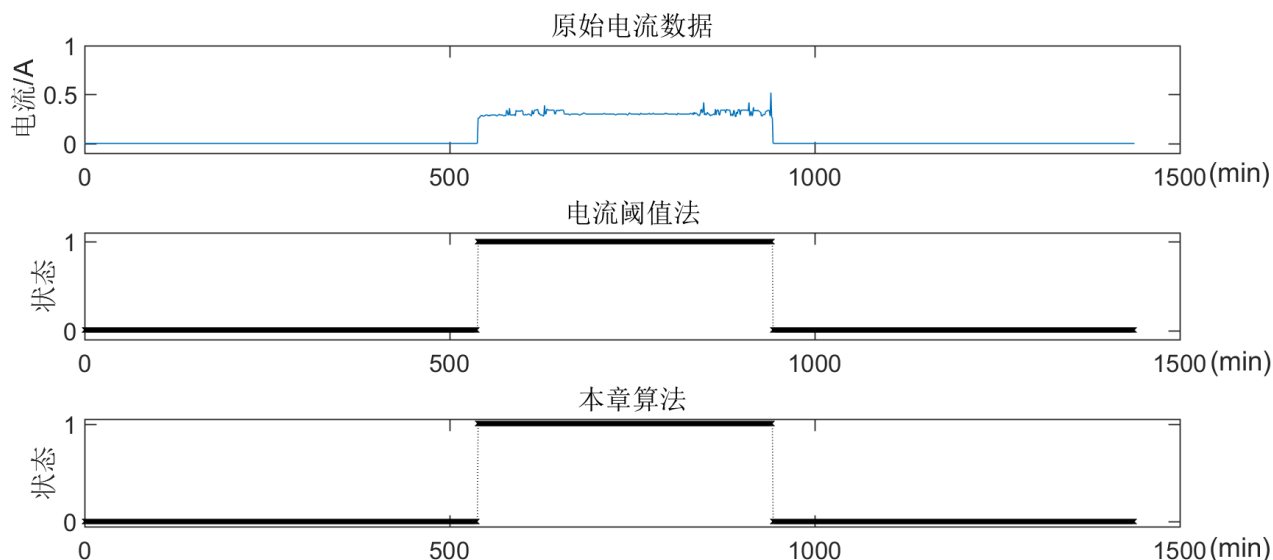


图 3.25 4 号仪器原始电流数据及两种算法分类结果

图 3.25 展示了 4 号仪器的原始电流数据以及两种算法对这些数据的分类结果。这台仪器的电流比较简单，各状态的电流区分度明显，因此两种算法的表现基本一致，都实现了电流的准确分类。

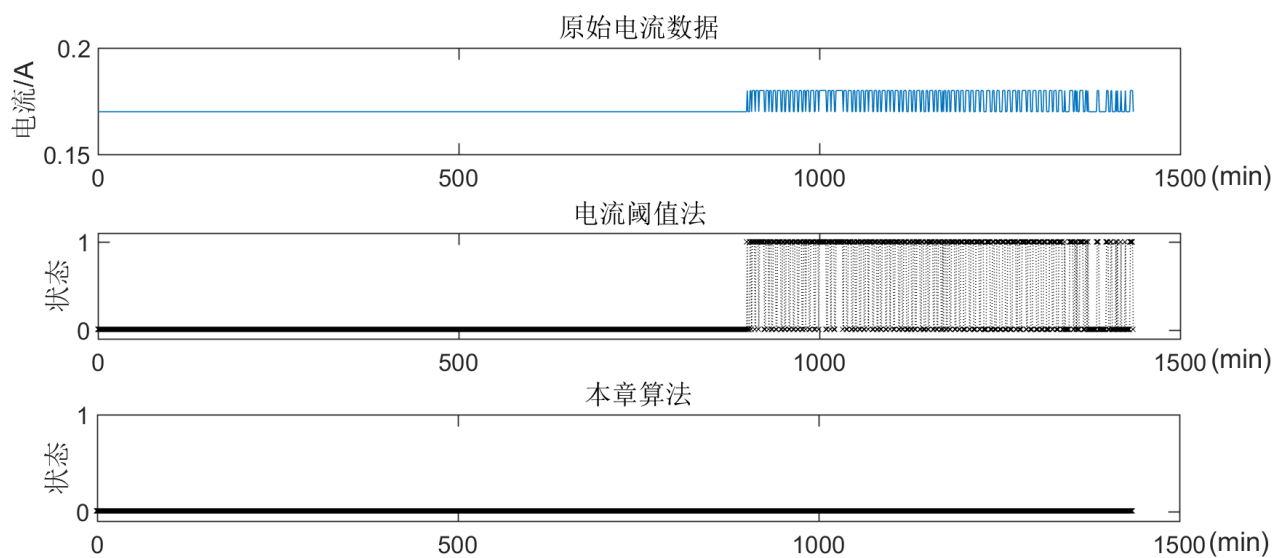


图 3.26 5 号仪器原始电流数据及两种算法分类结果

图 3.26 展示了 5 号仪器的原始电流数据以及两种算法对这些数据的分类结果。其中电流阈值法误判的原因与前几台仪器一样，都是因为不同状态下的电流值存在重叠而导致的误判，而本章算法由于分箱时范围划分的精度为 0.01，两个类别的电流数据被分至同一个箱中了，导致实际效果很差。

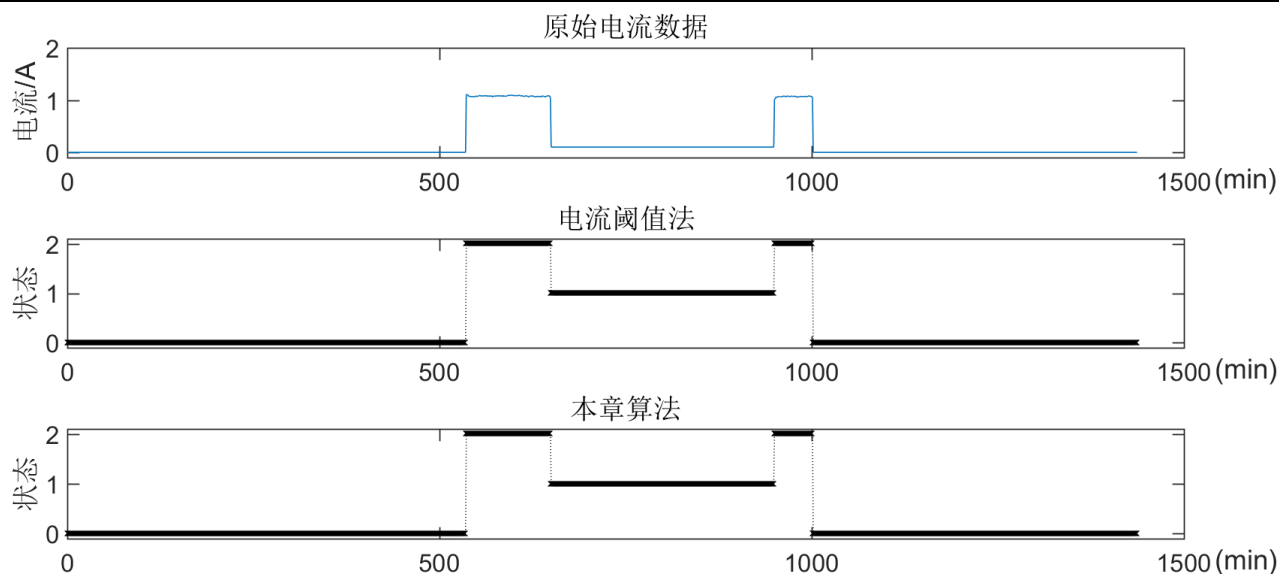


图 3.27 6 号仪器原始电流数据及两种算法分类结果

图 3.27 展示了 6 号仪器的原始电流数据以及两种算法对这些数据的分类结果。可以看到在这台仪器上两者都表现得很好，与实际相符合。

可以看到，对于各仪器电流变化较为平缓的状态，两种算法均可以表现出较好的效果，但是一旦出现状态间电流重叠的现象，电流阈值法就无法进行处理从而导致较大误差出现。本章算法更多的从电流波形上的特点进行分析，避免了电流阈值法判断标准过于单一的缺点。但是，本章算法在分箱处理时，会存在将不同状态电流分至同一个箱中的可能，这就会造成部分电流分类出错的情况出现。不过总体而言，在这 6 台仪器上本章算法的表现效果更好，除了第 5 台以外的其它仪器的分类都较为准确，而电流阈值法仅在第 4、6 台仪器上得到了较好的分类效果，在 1、2、3、5 台仪器上的效果都较差。

表 3.3 平均准确率

序号	电流阈值法	本章算法
1	0.798	0.974
2	0.893	0.988
3	0.785	1
4	1	1
5	0.612	0.236
6	1	1

表 3.3 为两种方法电流分类的平均准确率。可以看到，除了第 5 台仪器以外，本章方法在各仪器上的准确率都要大于等于电流阈值法，这表明根据该电流分类结果得到的仪器运行情况与实际情况是相当接近的。

3.5.4 机时计算测试

仪器机时通过统计电流的分类结果得出，统计不同运行状态下的电流数据个数就可以得到对应的分钟数了。以天为单位，使用本章算法计算上述仪器共 5 日的机时，同样以上述 6 台为典型来具体展示测试结果。

表 3.4 其中一日各仪器机时结果

编号	人工统计/min			电流阈值法/min			本章算法/min		
	关机	待机	工作	关机	待机	工作	关机	待机	工作
1	782	587	71	782	636	22	782	579	79
2	992	60	388	992	196	252	992	51	397
3	0	922	518	0	1255	185	0	921	519
4	1036	0	404	1036	0	404	1036	0	404
5	0	525	915	0	802	638	1440	0	0
6	559	693	188	559	693	188	559	693	188

其中一日机时计算结果的具体情况如表 3.4 所示，除了 5 号仪器外，其余仪器的误差分钟数都较低。为了更加直观体现算法性能，采用机时的相对误差作为指标进行评价，其计算公式如下所示：

$$p = \frac{1}{k} \sum_{i=0}^k \frac{|c_i - r_i|}{r_i} \times 100\% \quad (3.19)$$

其中， k 为仪器实际状态数， c_i 为算法算出状态为 i 的电流个数， r_i 为人工统计状态为 i 的电流个数，若仪器当天某个状态未出现，那么不加入计算。

表 3.5 机时平均相对误差

序号	电流阈值法	本章算法
1	56.1%	6.3%
2	72.0%	6.1%
3	45.4%	0.1%
4	0.0%	0.0%
5	50.5%	-
6	0.0%	0.0%

各仪器机时的平均相对误差如表 3.5 所示。可以看到，与电流分类表现出的性能一致，本章算法在除第 5 台仪器以外，其余仪器的机时平均相对误差均小于 6.5%，基本与实际相符。电流阈值法仅在第 4、6 号仪器上表现出较好的效果，即使后续人工介入调整阈值后，电流阈值法在 2、3、5 号仪器上的效果依旧很差，无法满足现实需要。

表 3.6 部分测试结果

仪器编号	7	8	9	10	11	12
相对误差	0.0%	5.5%	8.9%	0.0%	0.0%	4.5%
仪器编号	13	14	15	16	17	18
相对误差	4.1%	0.0%	0.5%	1.3%	7.6%	3.4%
仪器编号	19	20	21	22	23	24
相对误差	3.0%	7.1%	0.0%	4.7%	0.0%	0.0%

最后，对所选其它仪器进行测试，对机时计算情况进行统计后，发现共有 86 台仪器机时计算效果较好，与实际相符，部分仪器测试结果如表 3.6 所示。

3.6 实际应用分析及错误处理

本章算法在系统中部署后，8848 仪器中共有 1311 台仪器由于数据不足尚未生成模型，仅有一个状态的仪器共 1850 台，其余 5687 台仪器正常生成模型，具体数据如表 3.7 所示。

表 3.7 模型生成情况

模型状态	数量
数据不足	1131
一个状态	1850
两个状态	4445
三个状态	1422

通过对比仪器的电流波形以及模型分析的状态发现，其中两个状态中模型效果符合实际的仪器共 4052 台，与实际不符的仪器共 393 台，三个状态中模型符合实际的仪器共 940 台，与实际不符的仪器共 482 台。对于那些与实际不符的仪器需要进行处理，接下来以模型为三个状态的仪器为例对处理流程进行说明。具体的：

(1) 首先，由于模型生成时由于部分状态电流的缺失或仪器电流幅度变化太大导致模型状态数与实际状态数不匹配的仪器，与使用者沟通过后，重新选取仪器包含全状态的电流进行聚类，得到新的聚类模型。经过这一步之后，其中 196 台仪器的模型已经可以正确反映仪器的使用情况，准确仪器数量占比从 66.1%提升至 79.9%。

(2) 剩余 286 台仪器的模型状态数与实际状态数相符，但电流分类仍存在较大的误差，主要原因分为两类：第一类是由于仪器工作时电流波形存在多种形态，一部分工作电流与其它状态电流更相似，如部分液相色谱、荧光定量 PCR 等仪器；第二类是由于聚类时某状态电流数据量不足，使得分箱过程中两个状态的电流组成了一段电流子序列，聚类后电流特征不佳，如原子吸收分光光度计这一类运行时间较短，电流为脉冲形态的仪器。对于第一类仪器，可以采取人工调整特征值的方式：选取各状态下多个时段的电流数据进行特征求解，选取求解

结果中两个不同状态间最接近的两个电流特征分别替换其对应状态原有的分类特征。由于此类仪器中是工作状态下电流有多种形态，因此选取多个工作电流特征中与待机电流特征最接近那一个作为最终的工作电流特征即可。对于第二类仪器，可以选取包含全部状态的电流数据进行重新聚类或通过缩小分箱参数 T 使其小于其各状态的最小运行时长的方式对模型进行调整从而达到预期的分类效果。经过这一步后，其中 68 台仪器的问题被解决，这一步之后准确仪器数量占比从 79.9% 提升至 84.7%。

(3) 最后，对于剩余的 218 台仪器无论重新聚类或是调整模型特征都无法得到预期的效果，主要原因是仪器的电流十分复杂，导致分箱时就产生了大量误差，故对于此类仪器将采用有监督的方法进行处理。

最终，1422 台仪器中共 1204 台仪器的模型效果基本达到预期，占比 84.7%，本章算法无法处理的仪器占比 15.3%，相关数据如表 3.8 所示。

表 3.8 处理结果

处理阶段	数量	仪器占比
无需处理	940	66.1%
手动输入聚类数	196	13.8%
人工调整特征	68	4.8%
本章算法无法处理	218	15.3%

处理流程如图 3.28 所示。

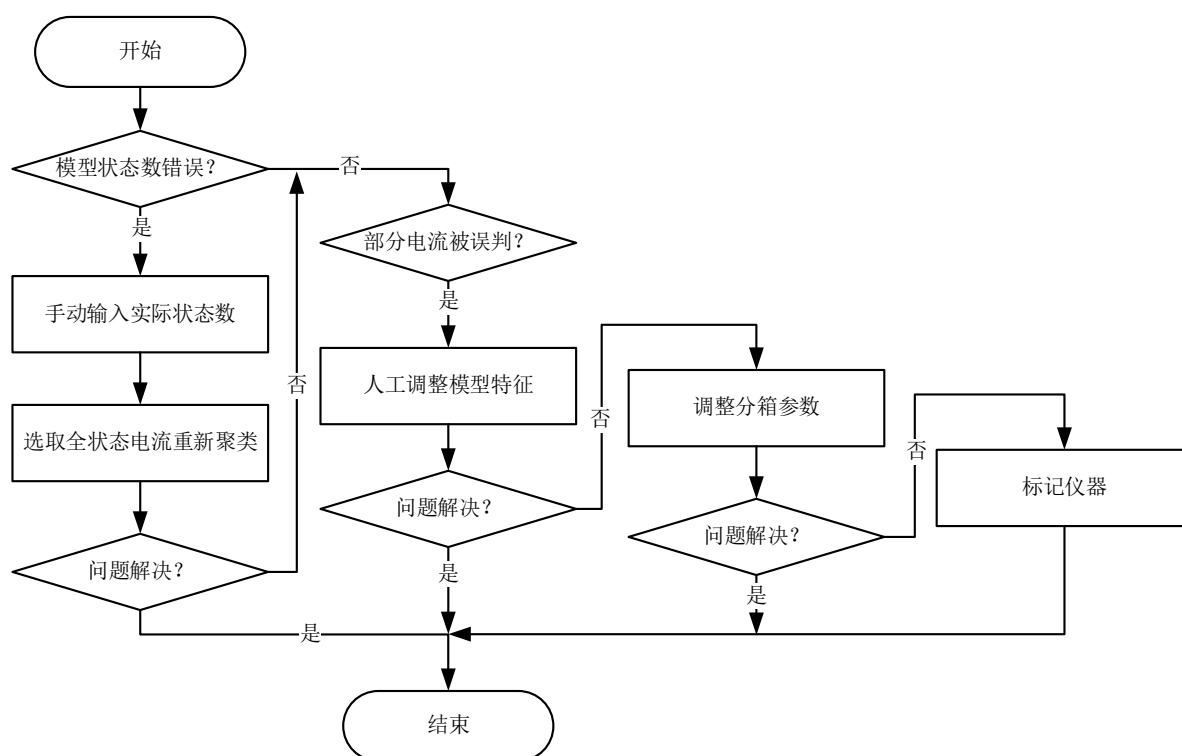


图 3.28 实际应用中错误处理流程

经过上述步骤处理之后，最后发现共有 345 台仪器由于电流较为复杂不适用于本章算法，因此将考虑采用有监督的方法，通过收集仪器各状态下的电流数据进行标定建库，然后针对每一台仪器的电流特点选择适合特征，最后逐一构建电流分类模型进行解决。

3.7 本章小结

机时计算的前提在于电流分类。在当前海量仪器场景下，常用的电流阈值法需要人工为各仪器一一设置阈值不太现实，即使通过聚类自动求解阈值的方案使得电流阈值法具备了一定的可行性，在一些仪器上表现较好，但仍然还有许多仪器不适用或者需要人工二次处理。因此，在对大量仪器的电流数据进行分析后，本章提出了一种基于分箱与改进 K-means 的机时计算方法。首先，设计了一种电流分箱算法对电流数据进行处理，基于通用性考虑选取均值和标准差作为仪器电流的特征。其次，分析了 K-means 算法的缺点，为了使算法可以由系统自动执行，从 k 值、初始聚类中心、聚类结果的选取进行了改进。然后，介绍了基于分箱与改进 K-means 的机时计算流程。选取了 105 台仪器来验证算法的有效性，并以其中 6 台典型仪器为例，从电流分类及机时计算结果上详细展示了本章算法的执行效果及优势。最后，对系统应用的实际情况进行了分析，并给出了针对出错模型的处理流程以及最后效果。该方法操作简单，对使用者友好，各仪器的电流特征可以通过聚类获取，无需人工标定数据、训练模型，达到了减少人工工作量和成本，提高效率的目的。

第 4 章 基于动态规划与支持向量机的机时计算

4.1 引言

基于分箱与改进 K-means 的机时计算方法已经可以实现对许多种仪器机时的自动计算了,但是仍然存在一些仪器的机时计算结果的误差较大。这些仪器的电流各有特点,再设计通用的聚类算法的难度较高,因此将引入有监督分类算法进行处理。按照业务的设计,每日输入电流数据往往是在多个仪器运行状态下产生的,此时希望通过电流分类去实现机时计算还要解决电流序列的切分问题。传统方式直接采用固定窗口切分的方式比较粗糙,存在两个问题:需要人工寻找各仪器合适的切分窗口,以及固定窗口的切分误差无法避免。也就是说使用此方法切分后的电流序列中往往会包含多个状态的电流数据,这会对机时计算的结果带来较大误差。因此,基于限制分割误差思想,本章提出了一种基于动态规划的信号切分算法,并与基于支持向量机的电流分类模型共同组成了一种基于动态规划与支持向量机(DP-SVM)的机时计算方法,总体思路如图 4.1 所示。

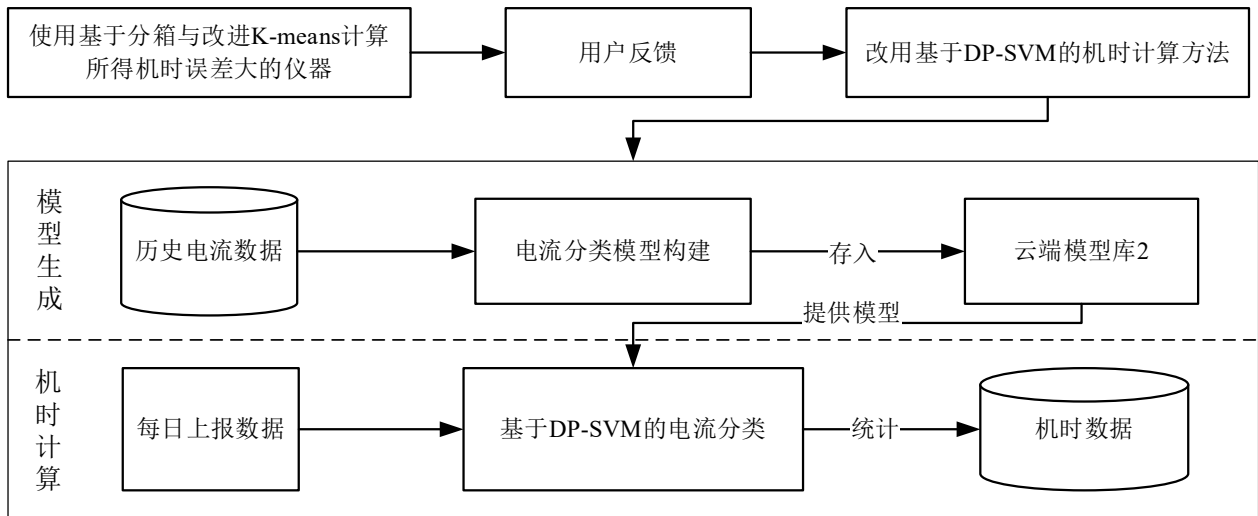


图 4.1 总体思路

对于目标仪器,首先需要从历史电流数据中收集其各个运行状态下的电流数据,然后训练出该仪器的 SVM 电流分类模型并存至云端。后续触发机时计算时,再调用该仪器云端的 SVM 模型配合 DP 算法实现电流分类,最后将分类结果统计后得到机时。

4.2 基于 SVM 的电流分类模型

4.2.1 支持向量机理论

SVM 算法由前苏联学者 Vapnik 提出,常被人们用于解决模式识别、故障诊断、样本分

类、回归预测等问题^[56-59]。该算法从统计学中的 VC 维理论和最小风险理论出发，通过在这两者间寻找平衡点来提升模型的泛化能力^[60]。

SVM 线性分类的过程实际上是寻找一个平面，使得不同类样本分散在该面的两侧。这个平面可以有无数个，但为了增强模型的泛化能力，通常会取距离两类样本的间隔最大的那一个面。此处引入线性可分的样本集 $S = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\}$ 对最佳分类面的求解进行说明，其中样本 $X_i = (x_1, x_2)^T$ ， y_i 表示所属类别，用 -1 和 1 来进行区分。由于样本是 2 维的，因此分类面就是一条能够分开两类的直线，如图 4.2 所示。

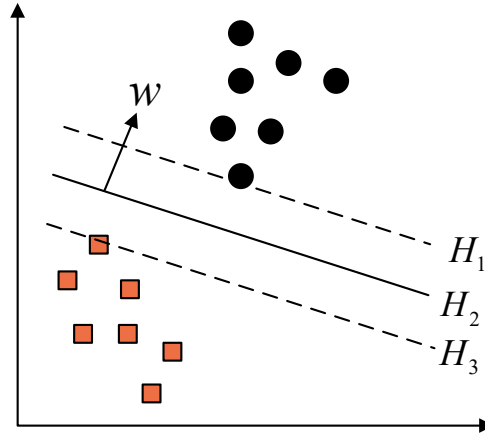


图 4.2 二分类示意图

图 4.2 中 H_1 、 H_2 、 H_3 代表不同的分类直线，表达式如下所示：

$$w^T X + b = 0 \quad (4.1)$$

式 (4.1) 中， $w^T = (w_1, w_2)$ 为直线的法向量， b 为截距。

此时，任意样本到直线的距离 d 如下所示：

$$d = \frac{|w_1 x_1 + w_2 x_2 + b|}{\sqrt{w_1^2 + w_2^2}} \quad (4.2)$$

这样一来，对于任意 X_i 只需要判断 $w_1 x_{i1} + w_2 x_{i2} + b$ 的符号与 y_i 是否一致就可以判断分类是否正确了。推广到 m 维后，任意点至分类面的距离 d 如下所示：

$$d = \frac{|w^T X + b|}{\|w\|} \quad (4.3)$$

式 (4.3) 中， $w^T = (w_1, w_2, \dots, w_m)$ 为 m 维空间中的超平面的法向量， $\|w\|$ 为法向量的 L_2 范数。

所述最优分类直线与两类之间的最大距离如下所示：

$$\gamma = \frac{\min_{y_i=1} (w^T X_i + b) - \max_{y_i=-1} (w^T X_i + b)}{2\|w\|} \quad (4.4)$$

引入函数间隔 ξ 使得正类样本的标准写作 $w^T X + b' \geq +\xi$ ，负类样本的标准写作 $w^T X + b' \leq -\xi$ 。为了方便起见， ξ 取值为 1，此时寻找最优直线等价于求解样本空间里的最大化间隔，如下所示：

$$\max_{w,b} \frac{2}{\|w\|}, s.t. y_i(w^T X_i + b) \geq 1, i=1,2,\dots,n \quad (4.5)$$

求解式 (4.5) 等价于求解式 (4.6):

$$\min_{w,b} \frac{\|w\|}{2}, s.t. y_i(w^T X_i + b) \geq 1, i=1,2,\dots,n \quad (4.6)$$

由于分类过程中可能会存在噪声的干扰, 因此 SVM 引入了松弛变量 ξ_i , 从而使得样本距平面的距离可以小于 1, 此时式 (4.6) 中约束条件写作:

$$y_i(w^T X_i + b) \geq 1 - \xi_i, i=1,2,\dots,n \quad (4.7)$$

当然, 松弛变量不是越大越好, 否则任意分类面都能符合条件。因此, 引入惩罚因子 C 进行限制。 C 越大则分类时更不能容忍错误的出现, 会使得模型训练的效果较好, 但是容易过拟合。 C 越小说明模型越能容忍错误, 错分的可能性变大, 但会更加合理。引入 C 和 ξ_i 后, 式 (4.6) 可以写作:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i, s.t. y_i(w^T X_i + b) \geq 1 - \xi_i, i=1,\dots,n \quad (4.8)$$

对上式进行求解就可以得到最佳分类直线 (超平面) 和分类决策函数表达式。

由于现实生活中许多问题是线性不可分的, 对于这类问题往往通过升维的方式进行求解。但高维空间的往往意味着高计算复杂度, 故 SVM 引入核函数使得在低维空间就可以求出结果, 节省了计算时间。常见核函数如表 4.1 所示。

表 4.1 常见核函数

名称	形式
多项式核函数	$k(x, y) = (ax^T y + c)^d$
径向基核函数 (Radial Basis Function, RBF)	$k(x, y) = \exp\left(-\frac{\ x - y\ ^2}{2\sigma^2}\right)$
线性核函数	$k(x, y) = (x^T y)$

二分类能够解决的问题毕竟有限, 当出现多分类的需要时, SVM 通常采用两种方式进行解决。一种方式是直接将多个分类面的参数求解放到一次求解中, 这种方式思想简单, 但计算复杂度高, 仅适合中小型问题。另一种方式是用二分类的思想进行推广, 将多分类问题抽象成一对一分类 (One-Versus-Rest, OVR) 或一对多分类 (One-Versus-One, OVO) [61] 解决。

4.2.2 核函数及参数选择

在训练 SVM 模型时, 不同种类仪器的电流分类模型在核函数及参数的选择上会有所不同。几种常见核函数的应用场景各有特点, 比如: 非线性问题可以使用多项式核解决, 但是需要设置较多的参数, 且当多项式阶数较高时会存在计算复杂度高的问题, 对于没有操作经验的人不够友好; 线性核最简单, 在许多特征较少且线性可分的场景下表现优秀; 径向基核函数较为通用, 无论对于大样本还是小样本都有比较好的性能, 是当前所有核函数中应用中最广泛的。基于当前场景下各仪器电流的差异, 此处优先考虑选取径向基作为各仪器电流分类模型的核函数。

RBF 核函数主要参数为 c 和 g (gamma), 模型训练时需要进行设置。其中, c 代表前文所述的惩罚因子, 该值越大表示模型越无法容忍误差, 对于分类面的要求更加苛刻, 容易造成过拟合; g 代表核函数的宽度, g 值越大支持向量越少, 反之则支持向量越多, 该值会对模型的训练速度造成影响。不同仪器对这两个参数选择会有不同, 在应用时往往根据经验进行选取, 考虑到后续模型训练功能可能会对仪器使用者开放, 故此处采用网格搜索算法来确定 SVM 的参数。网格搜索算法^[60]是常用的参数优化算法, 根据设定步长将设置的参数网格化, 然后遍历网格中所有节点来确定最优参数组合, 具体流程如下:

- 1、设定网格的搜索空间, 包括 c 、 g 的范围以及变化的步长。
- 2、将搜索空间中所有节点代入 SVM 模型。
- 3、采用 K 折交叉检验测试 SVM 的分类性能。
- 4、选择性能最高的 SVM 所使用参数搭配, 并保存 SVM 模型。

4.3 基于 DP 的信号切分算法

4.3.1 动态规划理论

动态规划算法在上个世纪 50 年代提出, 常被用于多阶段决策问题的优化^[62]。多阶段决策问题是指一个问题被拆分成多个子问题然后进行多阶段求解, 每个阶段都要解决一个子问题。其中, 求解当前阶段子问题的结果与前面所有阶段的结果有关, 当前阶段的结果又会影响到后续阶段的结果。当所有子问题求解完毕后, 就形成了一条求解路径, 一个多阶段问题中的求解路径是有多种组合的, 因此动态规划需要从中找出一条最优路径。基于这一特点, 该算法在工业生产、经济分配、工程技术等领域得到了大量应用^[63, 64]。

4.3.2 基于 DP 的信号切分算法设计

现在以一个经典剪绳子问题来引出本节信号切分算法。假设有一根长为 n 的绳子, 现要将绳子剪为任意段, 在绳子长度为整数的情况下求出剪切后的最大乘积。该问题可以这样分析: 若 $n = 1$, 那么无法被剪, 最大乘积为 1; 若 $n = 2$, 那么可以剪为两段长度为 1 的短绳, 或者不剪, 那么此时最大乘积为 2; 直到长度为 n 时, 可以假设最后一段绳子长度为 m , 那么

问题就转变成了求一根长度为 $n - m$ 的绳子的最大切分乘积问题了。设长度为 n 的绳子剪后的最大乘积为 $f(n)$ ，可以得到状态转移方程如下所示：

$$f(n) = \max(m \times f(n - m), n), m \in N^+, f(1) = 1 \quad (4.9)$$

根据式（4.9）的状态转移方程就可以求出最后的结果。

上述绳子可以转换为信号，求解最大乘积的过程可以转换为求解一个信号中各个类型的子信号切分最高准确度的过程。求解该问题的具体思路为：假设现存一种评价机制，且信号已经按照一定的方式切分成了 k 个子信号。此时对每个子信号的纯度进行评价并以一定的方式将评价结果量化为分数，然后再求出各子信号得分的平均值作为当前切分方式的得分，那么求最优解的过程就是在多种切分方式中找到得分最大的那一种。根据上述思想，长度为 n 的信号切分后的最大得分均值设为 $f(n)$ ，可以写出状态转移方程如下：

$$f(n) = \max_{l,k} \frac{f(n-l) \times k + score_l}{k+1} \quad (4.10)$$

其中， l 为最后一个子信号的长度， k 为该信号前 $(n-l)$ 部分被切分的子信号个数， $score_l$ 代表最后一个子信号的得分。

在动态规划的求解过程中为了记录最佳切分位置，可以定义 $g(n)$ 为以第 n 个数据为结尾的最佳子信号长度，此时各子信号长度的计算方式如下式所示：

$$g(n) = l_1, g(n-l_1) = l_2, \dots, g(n - \sum_{i=1}^{k-1} l_i) = l_k, s.t. \sum_{i=1}^k l_i = n \quad (4.11)$$

可以看到，以第 n 个数据为结尾的子信号的起始位置为 $n - l_1 + 1$ ，由于信号不重叠，那么下一个子信号的结尾位置为 $n - l_1$ ，起始位置为 $n - l_1 - l_2 + 1$ ，依次类推即可求出所有的切分点进而得出各子信号。

在当前场景下还存在一个约束条件：在搜索过程中，子信号的长度需要有限制，不能太长也不能太短。若子信号太短，那么一个很小的异常或是仅取了局部很小一段信号反映不出原本的状态就会影响评分的准确性；若子信号太长，那么其中存在的多个子信号反映纯度的特征可能会被平均掉，也会出现错误的结果。因此，加上约束条件后，最大得分均值的状态转移方程如下所示：

$$\begin{aligned} f(n) &= \max_{l,k} \frac{f(n-l) \times k + score_l}{k+1} \\ s.t. & l_{\min} \leq l \leq \min(l_{\max}, n) \end{aligned} \quad (4.12)$$

其中， $l_{\min}$ 为子信号的最小长度， $l_{\max}$ 为子信号的最大长度。此时，状态转移方程的初值可以设置为 $f(l_{\min})$ ，第一个搜索位置为 $l_{\min} + 1$ 。

所述算法的流程如图 4.3 所示。

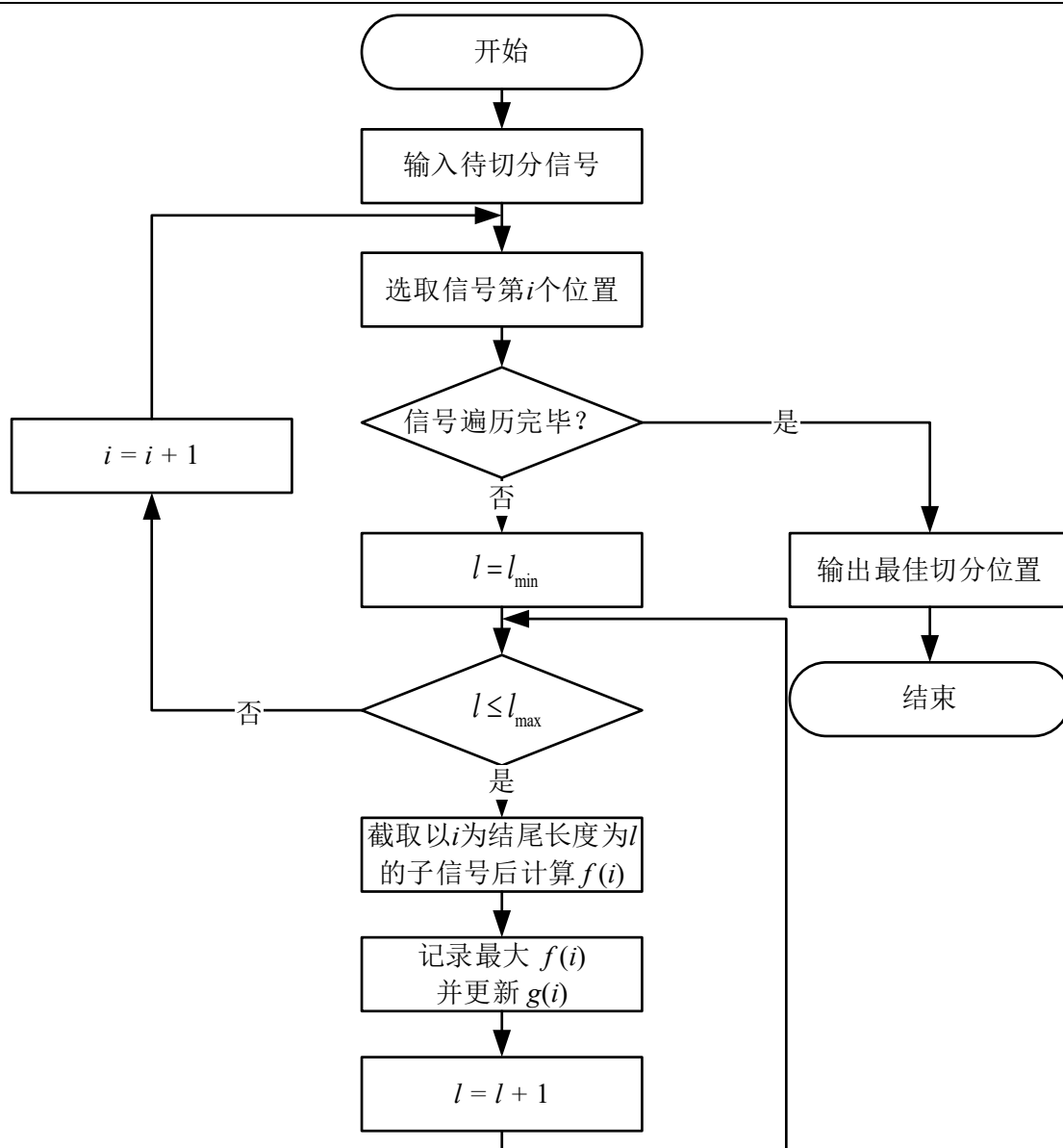


图 4.3 求解信号最佳切分方式流程

4.3.3 基于 DP 的信号切分算法测试

4.3.3.1 测试准备

为了验证算法的效果，此处选择两种模拟信号进行仿真测试。具体信息如下：

1、模拟信号 1 模拟较为简单的仪器：当信号的幅值为 1 时表示仪器处于工作状态，当信号的幅值为 0 时表示仪器处于关机状态。关机状态的时长在 10 至 30 分钟内随机生成，工作状态的时长 8 至 35 分钟内随机生成。

2、模拟信号 2 模拟待机态和工作态电流存在重叠仪器：仪器关机时电流始终为 0，而仪器工作时电流是多个幅值为 1 的方波组成，两个方波间会存在电流为 0 的几分钟间隔。

信号得分量化工具采用 sklearn 工具包中的 predict_proba 函数，该函数需要与分类模型配

合,可以给出样本属于某一类的概率。 predict_proba 基本原理为: 计算分类样本至最佳分类面的距离, 然后按照将该距离转换为对应概率。假设共有 k 类, 那么该函数就是一个 k 维的向量, 向量第 j 个位置的值代表样本 i 是第 j 类的概率。当某一类的概率要远远大于其它类的概率时, 就说明当前样本的质量越好, 得分也就越高。当某个样本与所有训练样本都有较大的差异, 那么该样本属于每个类的概率都相同, 此时得分就为 0。样本 i 的分类概率输出如式 (4.13) 所示, 信号得分量化公式如式 (4.14) 所示。

$$\text{predict_proba}_i = [p_1, p_2, \dots, p_k], \sum_{j=1}^k p_j = 1 \quad (4.13)$$

$$\text{score}_i = k * \max(p_1, p_2, \dots, p_k) - 1 \quad (4.14)$$

其中, score_i 即为第 i 个样本的得分。

图 4.4 展示的是测试信号及在测试信号上使用滑动窗口截取出的各子信号得分变化曲线。测试信号共两种状态: 第一种状态信号值为 1, 第二种状态信号值为 -1。滑动窗口大小为 20, 每次求解截取出的子信号得分后向后移动 1 个单位, 直到剩余信号长度不足 20。信号评分工具使用的 SVM 模型采用 RBF 核, 为了更好展示信号得分的变化曲线, 此处设置参数 $c = 1$, $g = 1$ 。取不同状态下信号的均值作为特征构建数量平衡的训练样本集, 且训练样本中不存在混合状态的信号。

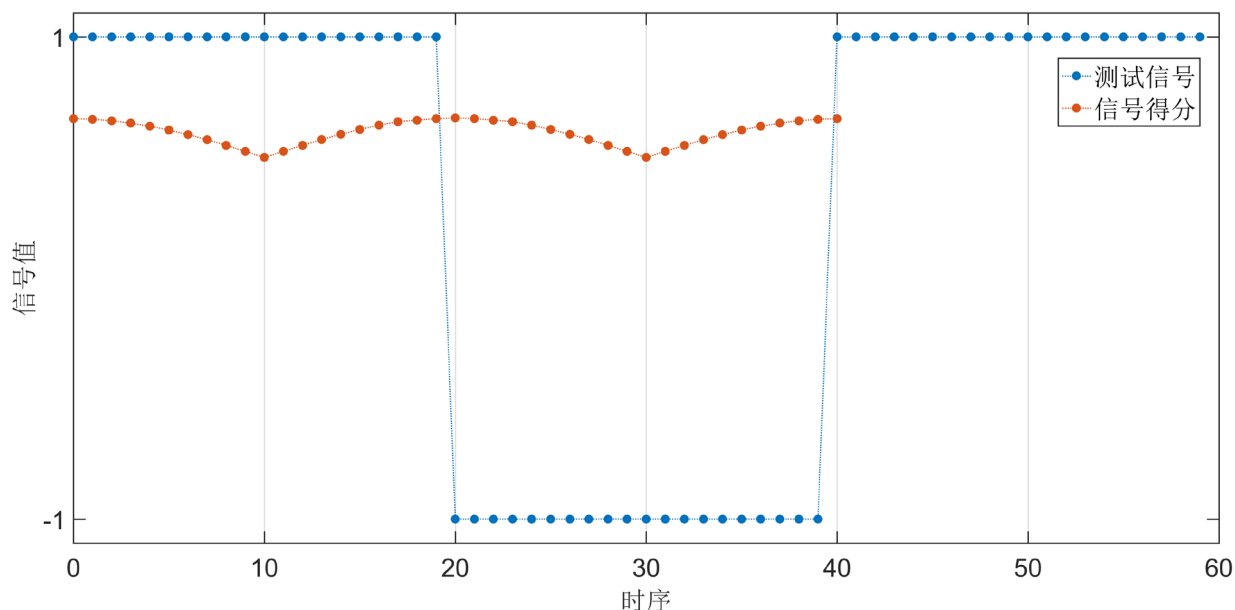


图 4.4 测试信号及信号得分曲线

其中, 各子信号起始位置横坐标对应的得分就是该子信号的得分。为可以看到, 最开始阶段滑动窗口中截取的信号完全属于第一种状态, 此时信号得分最高。随着窗口滑动, 第二种状态的信号慢慢开始出现, 此时得分越来越低, 到达横坐标 10 时窗口中的信号一半属于第一种状态, 一半属于第二种状态, 此时得分最低。然后窗口中第二种状态的信号慢慢占据大部分, 到横坐标为 20 时, 窗口中只包含第二种状态的信号, 此时信号得分恢复为最大值。后

续得分变化情况也是符合此规律的，可以看到该得分函数可以反映出样本的质量。

测试中 SVM 核函数选用 RBF，核心参数 c 和 g 分别设置为 1 和 10，选取信号各状态的均值作为特征进行训练。DP 算法的搜索范围设置为[5, 10]，采用传统直接使用固定窗口进行切分再分类的方法（下文简称固定窗口法）进行对比。固定窗口法的窗口值 w 分别为 5、6、7、8、9、10，与 DP 算法使用相同模型。每个信号共测试 5 轮，使用准确率对算法的效果进行评价。

4.3.3.2 测试结果

测试结果如下所示。

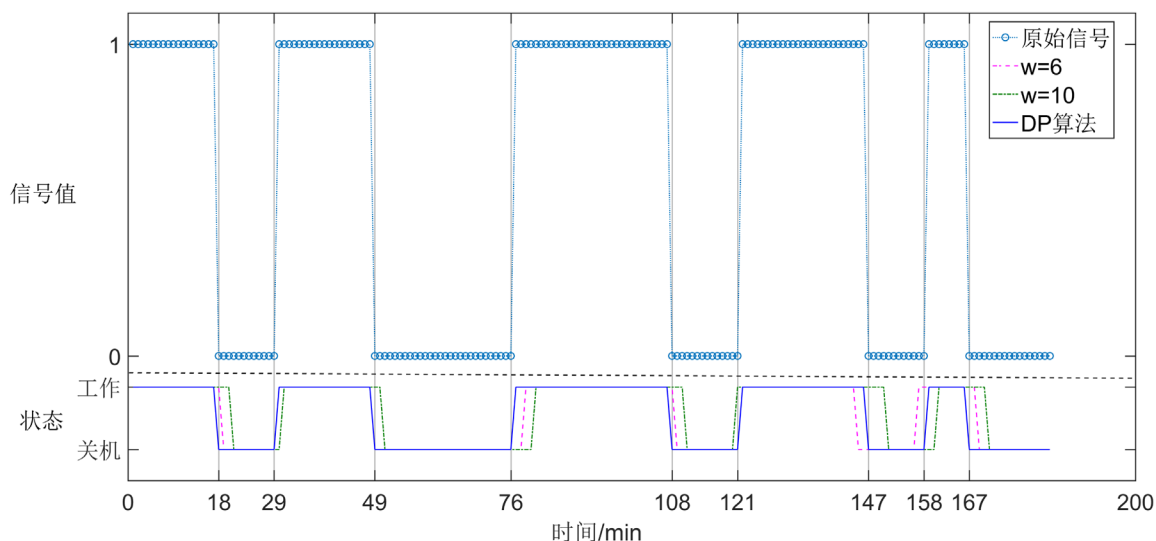


图 4.5 值为 6、10 的固定窗口法与 DP 算法的效果对比

图 4.5 展示了窗口值为 6、10 的固定窗口法与 DP 算法对信号 1 的状态分类结果，可以看到 DP 算法可以准确切分信号的各个状态，而固定窗口则效果较差，在多个状态变化的位置都产生了误差。

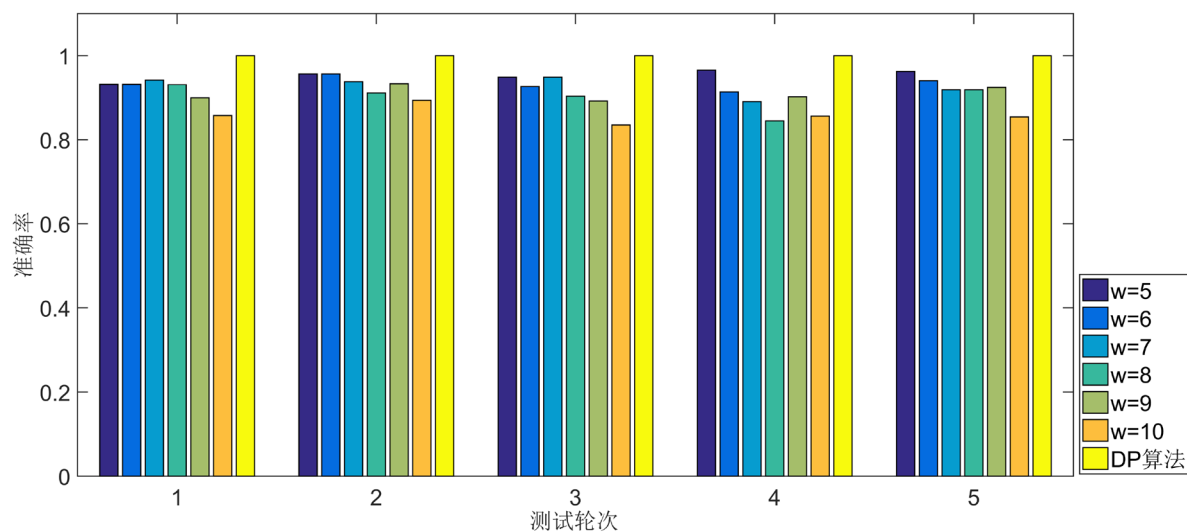


图 4.6 5 轮测试中各算法准确率对比

图 4.6 所示为测试 5 次中 DP 算法与不同参数的固定窗口法准确率对比。可以看到 DP 算法的准确率始终为 1，固定窗口法不论取多大的窗口都始终存在误差。并且，同一窗口值在每次的测试中的表现结果都存在着差异，不存在最佳的窗口。

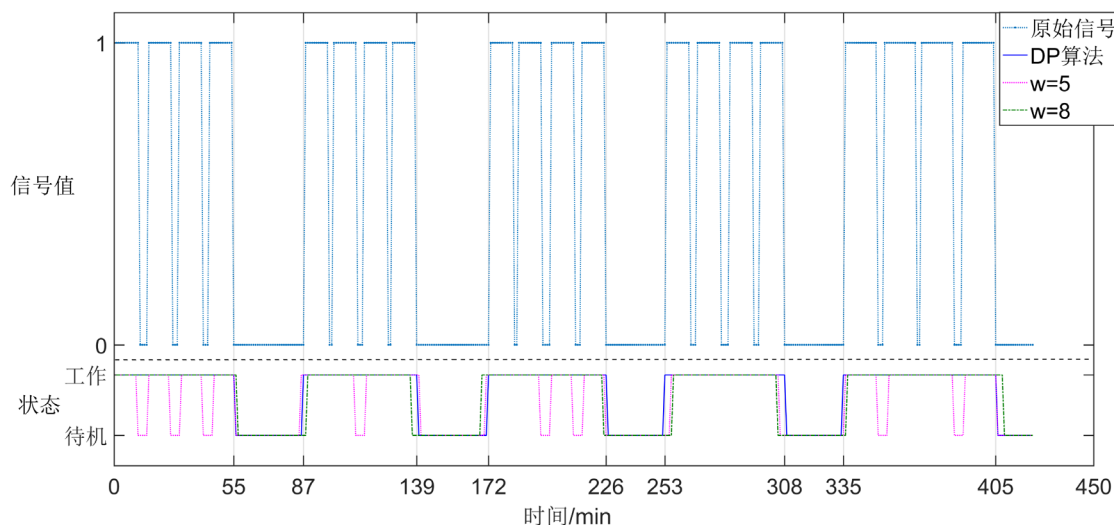


图 4.7 值为 5、8 的固定窗口法与 DP 算法的效果对比

图 4.7 展示了值为 5、8 的窗口与 DP 算法对信号 2 的状态分类结果，可以看到当窗口值为 5 时误判较多，一共出现了 8 处较为明显的误判，窗口值变大时虽然减少了误判的出现，但信号切分的边界也变得不准确。而 DP 算法仅在信号状态变化的边界处出现微小的误差。

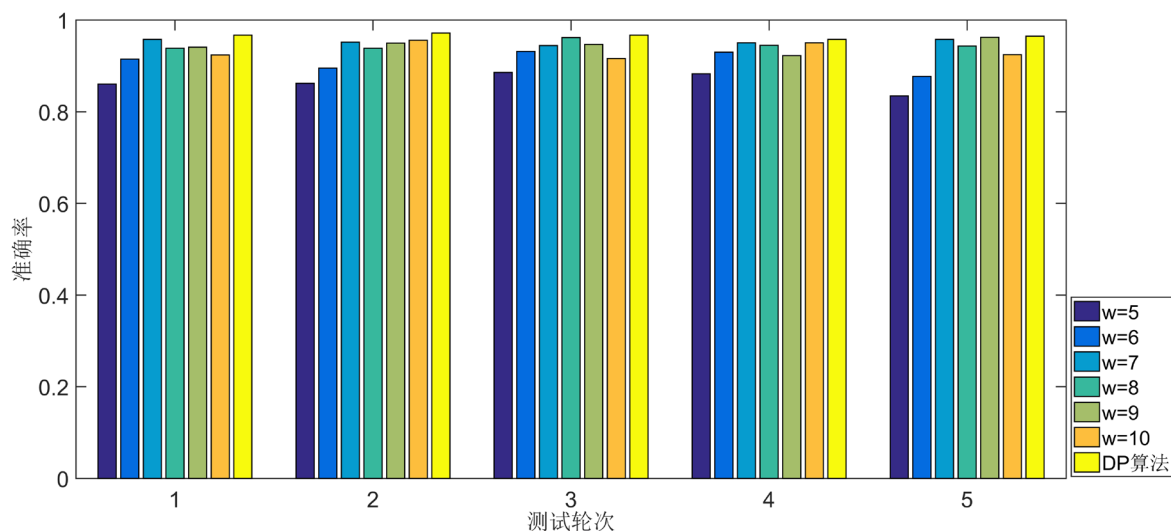


图 4.8 5 轮测试各算法准确率对比

图 4.8 所示为测试 5 轮 DP 算法与不同大小的固定窗口准确率对比。可以看到 DP 算法的准确率最高，且在多轮测试中表现稳定。

总体而言，固定窗口法的准确率受到窗口大小和各状态出现时间的影响，不同信号适用的窗口值是不同的。比如在信号 1 上窗口值越小分类准确率越高，而在信号 2 上窗口值太小或者太大分类的准确率都会降低，实际应用中就需要有一定经验的人工进行设置。而在 DP 算法中，不用特意去寻找最佳窗口值也可以得到较高的分类准确率，后续的电流分类也可以采取该方法。

4.4 基于 DP-SVM 的机时计算流程

根据设计思路，基于 DP-SVM 的机时计算方法详细执行流程如图 4.9 所示。

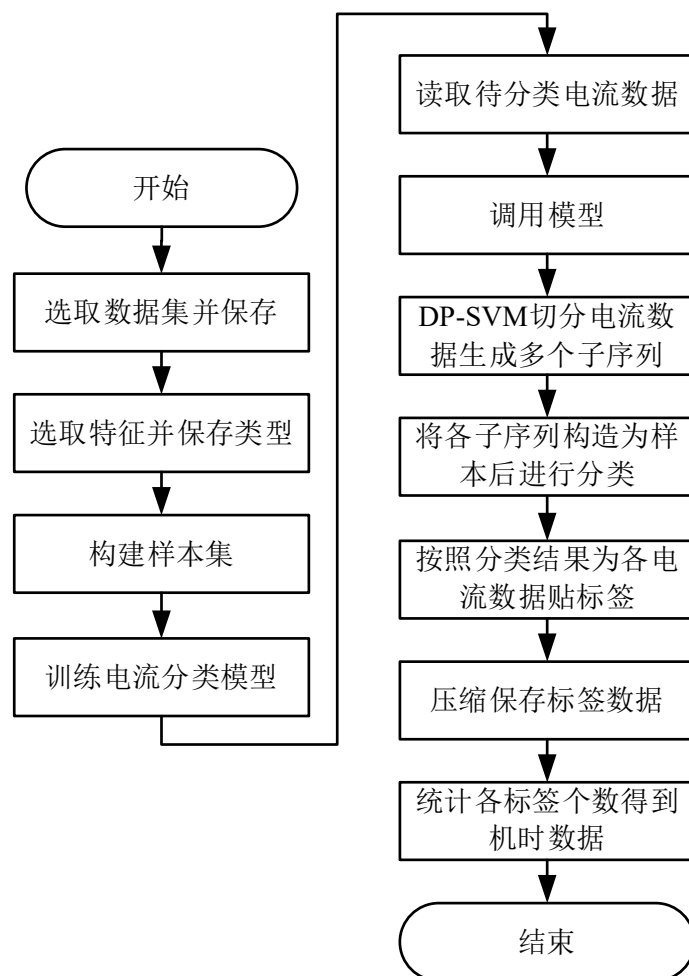


图 4.9 DP-SVM 机时计算流程

具体执行步骤为：

1、获取各状态下的电流数据。数据库中存储了各仪器大量电流数据，只要根据使用者反馈就可找对各状态下的电流数据进行选取，然后将对应时间段存入数据库进行记录。

2、构建样本集。获取到各状态的电流数据后，首先需要提取特征构建训练、测试所用的样本集。由于仪器种类不同，需要的特征个数、类型各不相同，按照实际情况选取选好后需要进行记录。提取特征时若使用如 *z-score* 标准化方式处理，那么涉及到的参数也需要保存。同样的，样本的标签按照关机、待机、工作记为 0、1、2。

3、训练模型。模型的参数通过网格搜索算法与 K 折交叉检验进行搜索，选择性能最好的那一组参数作为最终参数，模型训练完成后以仪器 IMEI 命名并保存至云端。

4、实际应用。当触发每日机时计算业务后，首先读取前一日的电流数据，然后根据 IMEI 读取模型以及特征类型等参数。然后，通过动态规划算法对电流数据进行切分，形成电流子序列后按照训练模型时使用的特征将各电流子序列构建为样本。再使用模型对各样本进行分

类，按照分类结果为所有电流数据打上标签，各电流数据标签与所属样本相同。最后，将状态标签进行压缩存储，统计其中 0、1、2 标签的个数，得到最终的机时计算结果。

4.5 实验测试

本章涉及到的动态规划算法采用 Python 编写，SVM 算法以及相关的工具采用 Python 机器学习 sklearn 工具包实现，开发工具为 PyCharm Community Edition 2020.2.3。

4.5.1 实验准备

为了验证 DP-SVM 的实际效果，选取 10 台仪器作为样本进行测试，以其中 3 台典型的仪器为例进行详细展示分析，相关的信息如表 4.2 所示。

表 4.2 测试仪器相关信息

序号	IMEI	仪器名称	训练样本	状态类别
1	869858032096688	荧光光谱仪	120	3
2	864823041521121	液相色谱仪	100	2
3	864823041538547	气相色谱-质谱联用仪	100	2

其中，2、3 号是采用第 3 章算法会存在较大机时误差的仪器，因为其使用特性，历史数据中只包含待机和在工作两种状态。1 号是使用第三章算法可以处理的仪器，共包含关机、待机、工作三种状态。各台仪器不同状态下的电流特点如图 4.10、图 4.11 及图 4.12 所示，同样的，为了方便展示，绘制时横轴不展示具体的时间，以各电流数据出现时序作为坐标，单位为分钟。

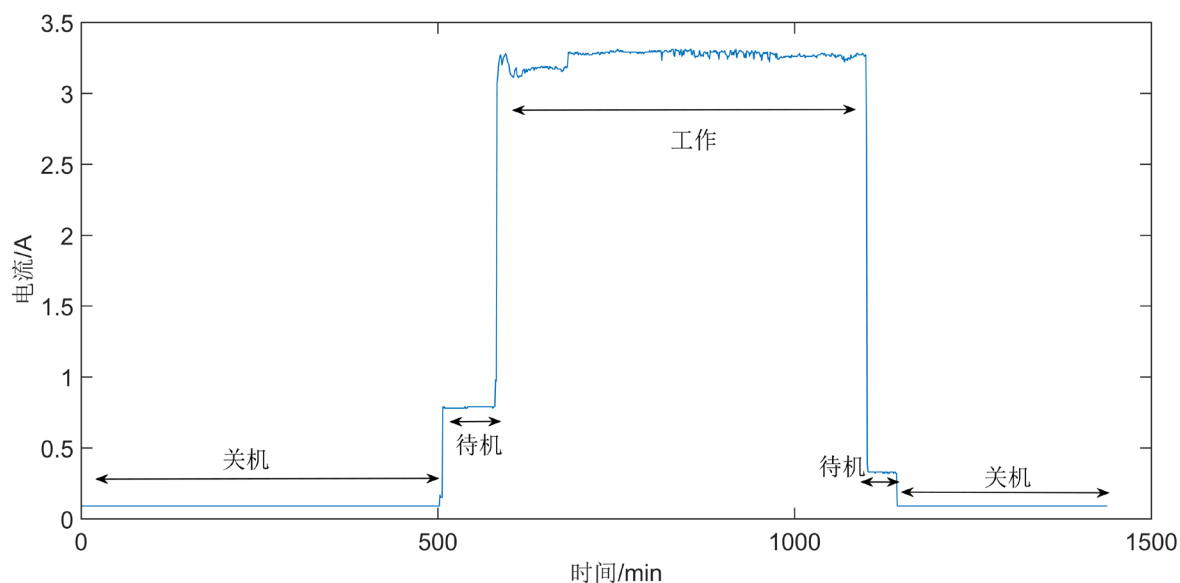


图 4.10 荧光光谱仪电流波形

图 4.10 展示的是荧光光谱仪的电流波形图，可以看到该仪器各个状态下的电流差别很大，

关机电流在 0.09A 至 0.1A 之间浮动。待机时电流比较平稳，工作时电流超过 3A 且存在一定的波动。为了使模型的性能更好，此处选取各状态电流的均值和标准差作为样本特征。

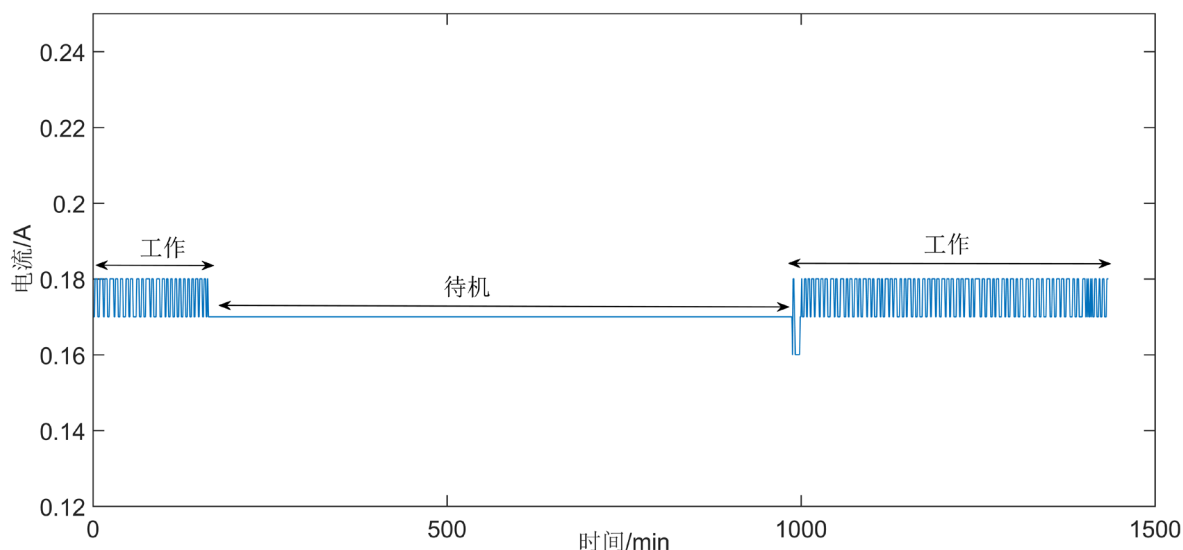


图 4.11 液相色谱仪电流波形

图 4.11 展示的是液相色谱仪的电流波形图，可以看到该仪器两个状态下的电流值差别不大，此时仪器待机电流值为 0.17A，工作电流值在 0.16、0.17、0.18A 之间上下波动。0.01A 的电流值分辨率太小，很容易受到噪声、仪器使用环境的干扰，故不使用均值作为特征。不过，工作电流和待机电流波动变化明显，可以选用标准差作为样本特征。

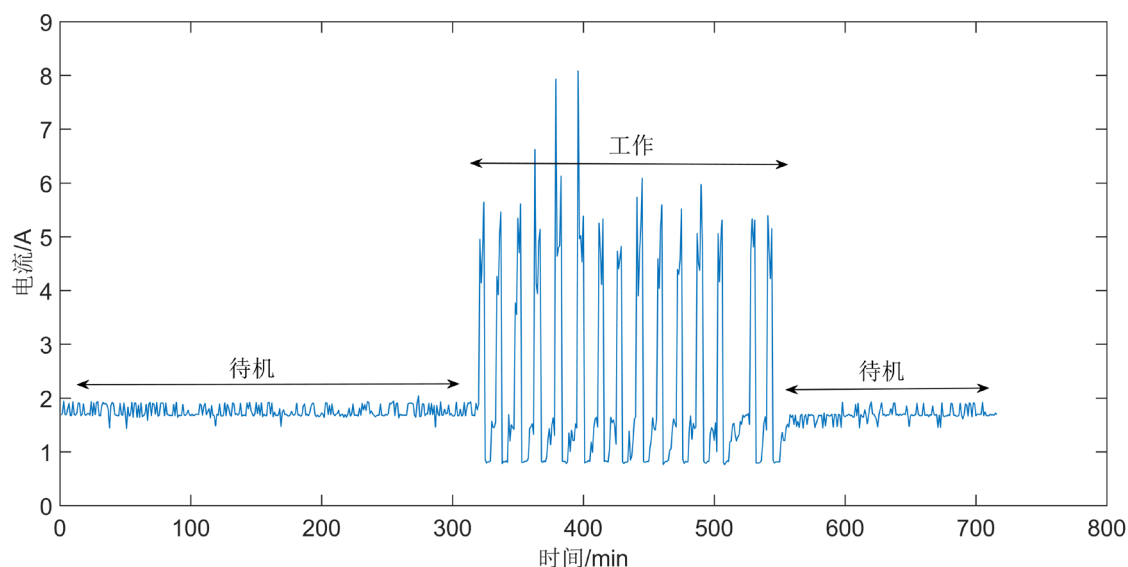


图 4.12 气相色谱-质谱联用仪电流波形

图 4.12 展示的是气相色谱-质谱联用仪的电流波形，可以看到该仪器两个状态下的电流值差别较大，除了均值以外，标准差也可以作为特征选取。

各仪器训练测试所用的样本数据均来源于数据库中的历史电流数据，由于有些时候一个状态的持续时间较长，可以进行切分处理，这样不仅可以对整体样本数量进行扩充，还可以更好的模拟实际仪器使用的各种情况。各仪器相关特征部分原始数据如表 4.3 所示。

表 4.3 特征原始数据

序号	均值	标准差	状态
1	0.100	0.00	关机
	0.332	0.00	待机
	0.791	0.00	待机
	3.174	0.12	工作
	1.471	0.04	工作
2	0.170	0.00	待机
	0.180	2.64e-16	待机
	0.176	4.80e-3	工作
	0.165	4.89e-5	工作
3	1.734	0.33	待机
	1.541	0.12	待机
	2.722	1.56	工作
	2.339	1.53	工作

这些仪器使用的 SVM 模型参数 c 和 g 的范围分别设置为 $(0, 10)$ 与 $(0, 100)$ ，步长分别设置为 0.1 和 1 进行搜索，搜索的过程中 K 折交叉检验的折数设置为 5。基于训练集中各样本的长度考虑，动态规划窗口搜索范围设置为 $[15, 60]$ 。此外，动态规划的搜索需要一个初始值，此处将前 15 分钟的电流子序列的得分作为初值。信号评价量化工具仍选取 `predict_proba` 函数，DP-SVM 测试所用数据为仪器端后续采集到的多日数据，不与训练时所用样本的数据重复。

4.5.2 测试结果与分析

由于分类任务较为简单，SVM 训练过程不再赘述，DP-SVM 的测试结果如下所示。

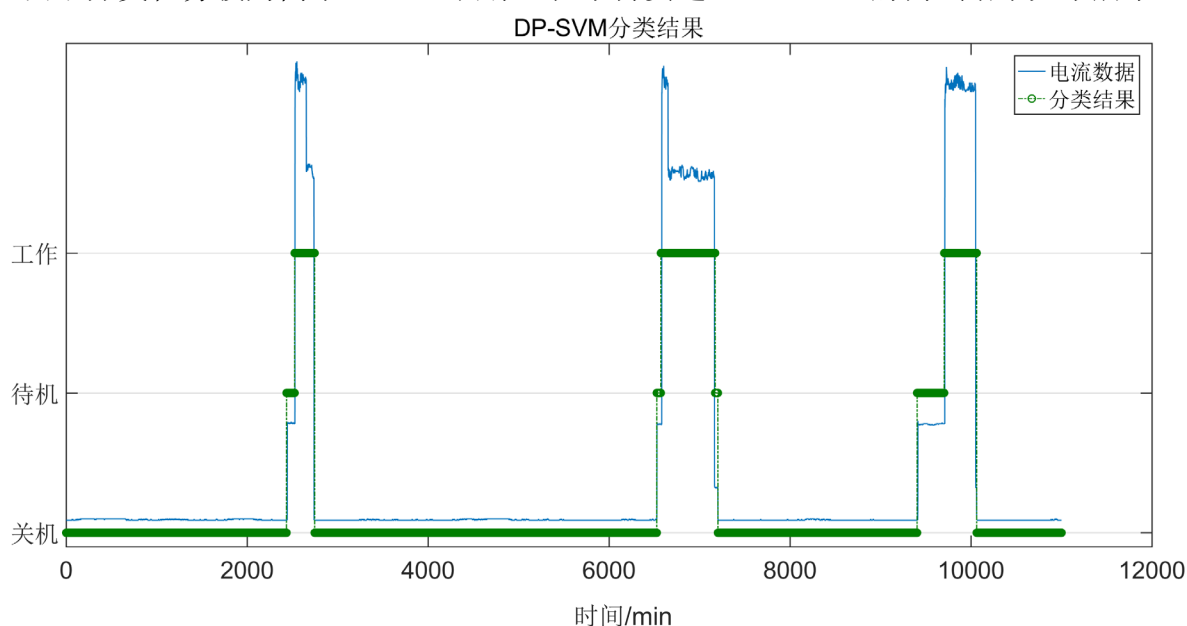


图 4.13 荧光光谱仪电流分类结果

荧光光谱仪电流分类结果如图 4.13 所示。可以看到，DP-SVM 有效地区分出了关机、待机、工作这三种状态下的电流，与实际情况相符。

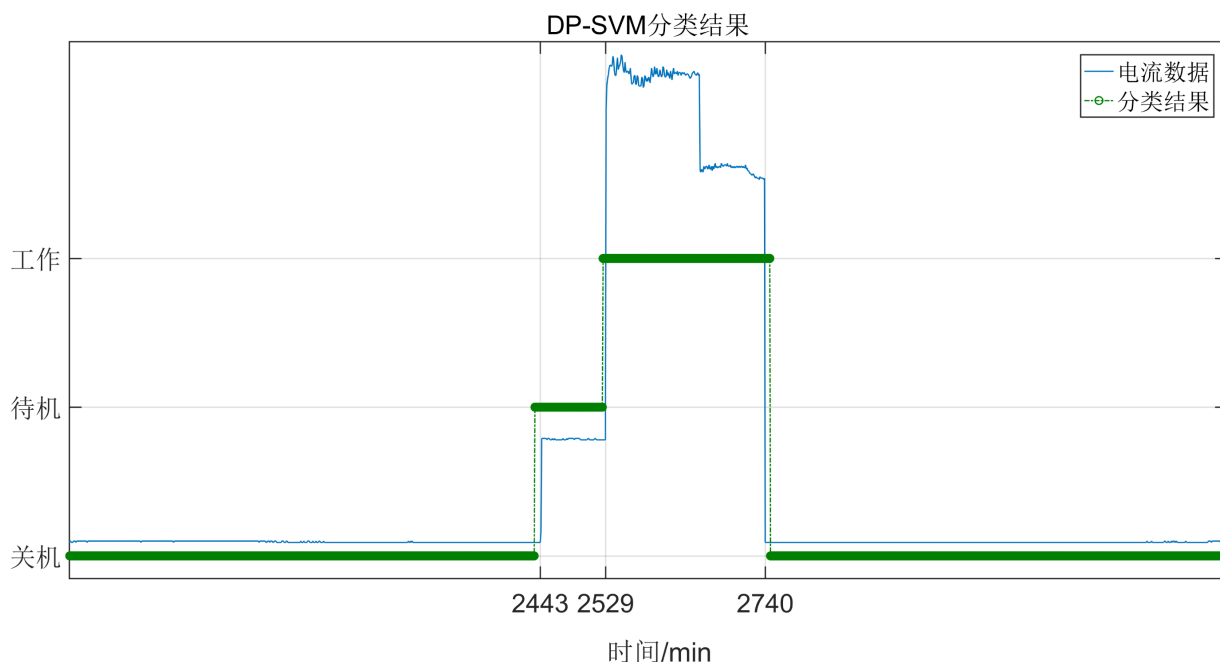


图 4.14 荧光光谱仪电流分类的局部细节

电流分类结果的局部细节如图 4.14 所示。可以看到，荧光光谱仪在 2443 分钟后由关机状态变为待机状态，在 2529 分钟后由待机状态转换为工作状态，在 2740 后由工作状态转换为关机状态。DP-SVM 的电流分类结果基本和实际是一致的，存在略微的误差的原因是子序列的最小值存在限制，在切分时为了更好的结果会不可避免的将两个状态的电流数据放在一起。

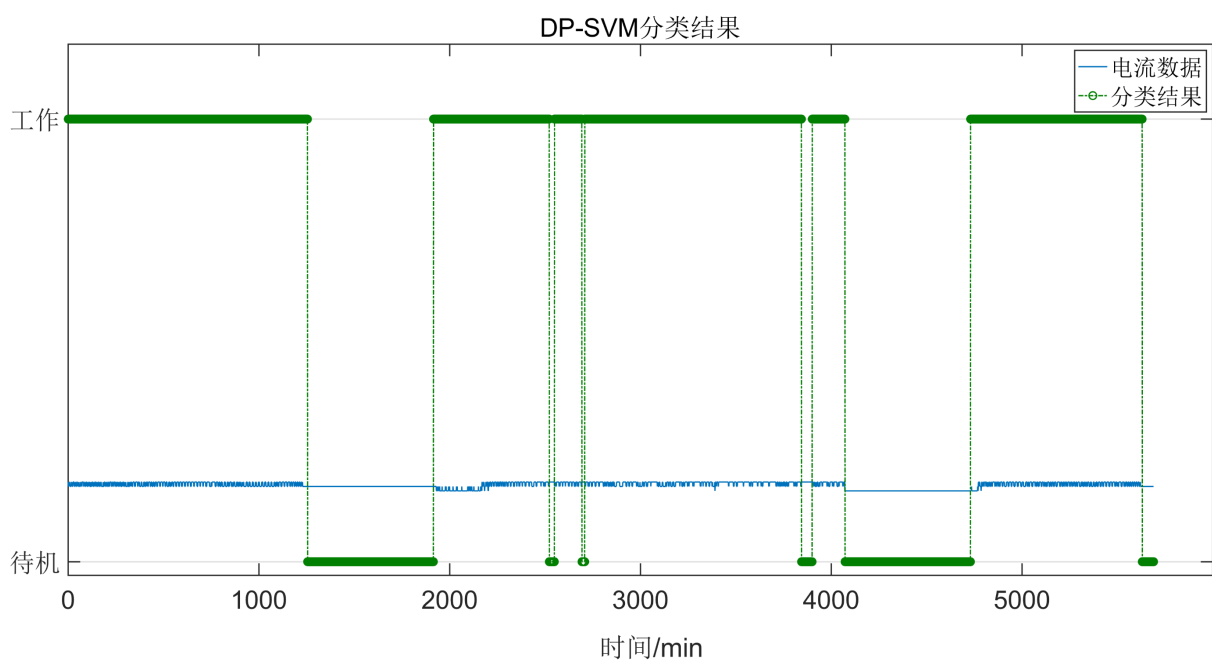


图 4.15 液相色谱仪电流分类结果

液相色谱仪电流分类结果如图 4.15 所示。可以看到，DP-SVM 的电流分类效果较好，总

体符合实际情况，在第 2000 至第 4000 分钟内存在一些分类错误。这类错误产生的原因在于采样频率太低，导致该仪器有时工作电流没有变化，其特征与待机电流的特征更相似，所以造成误判。

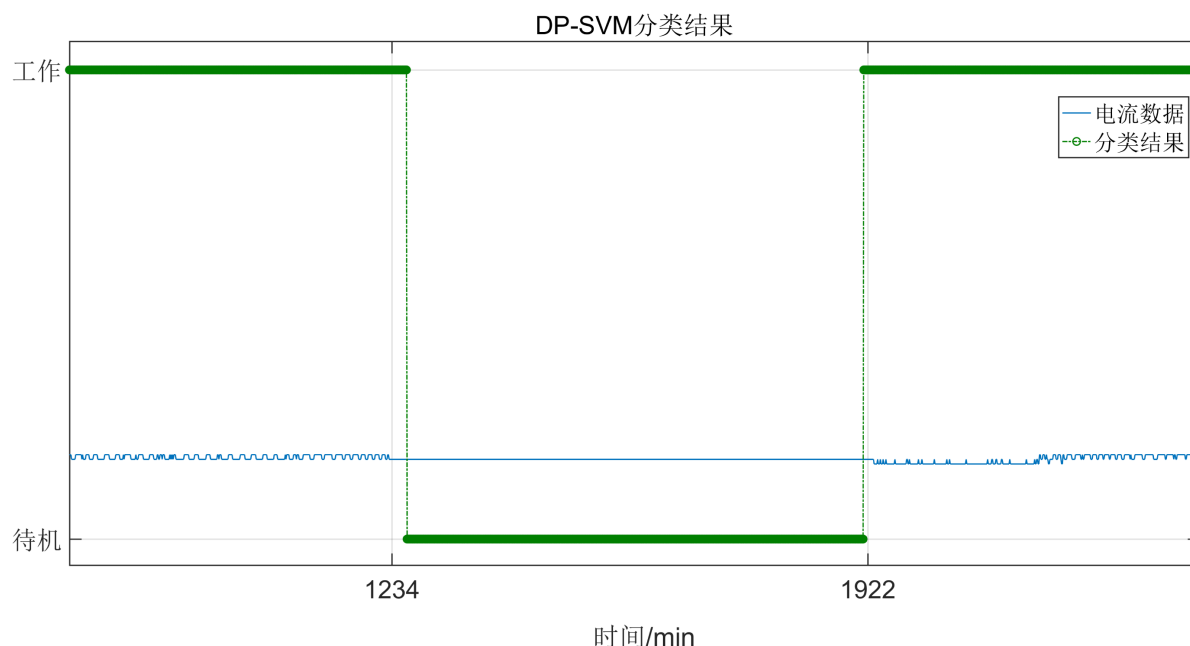


图 4.16 液相色谱仪电流分类局部细节

电流分类结果的局部细节如 4.16 所示。可以看到，液相色谱仪在 1234 分钟后由工作状态变为待机状态，在 1922 分钟后由待机状态转换为工作状态，在这两个位置附近出现误差的原因与分类无关，而因为算法在最小窗口的限制下为了得到更好分类结果，从而不可避免地将不同状态的电流数据切分到同一个电流子序列中。

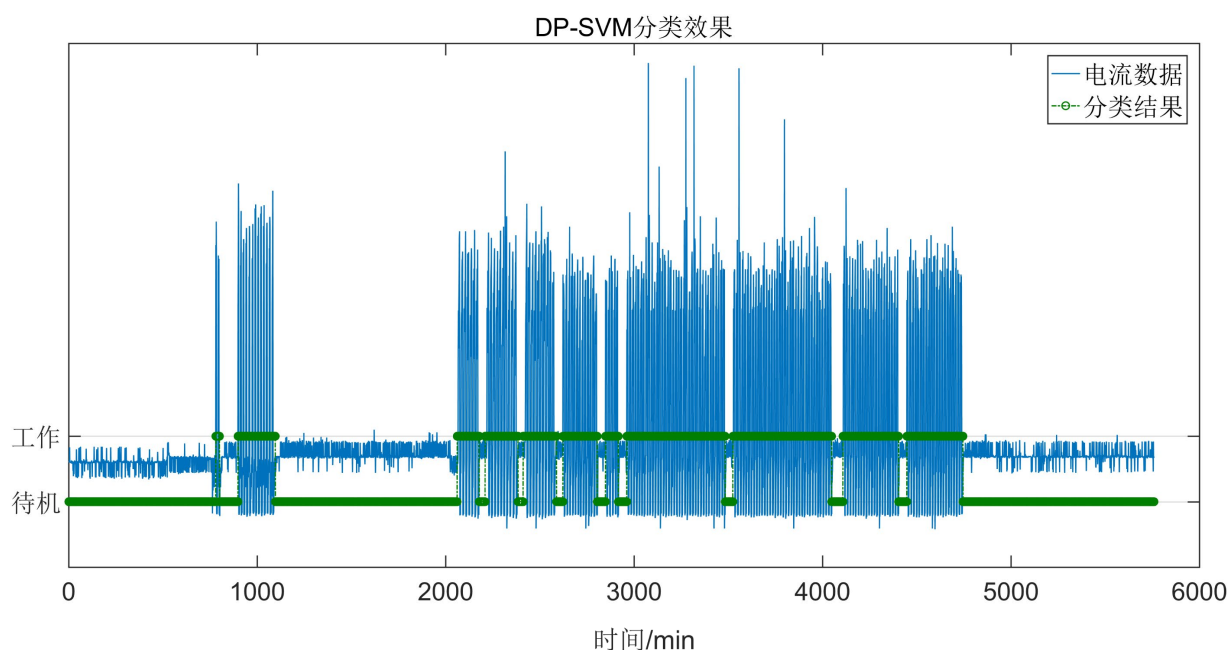


图 4.17 气相色谱-质谱联用仪电流分类结果

气相色谱-质谱联用仪电流分类测试结果如图 4.17 所示。可以看到，电流分类结果总体符

合实际情况。

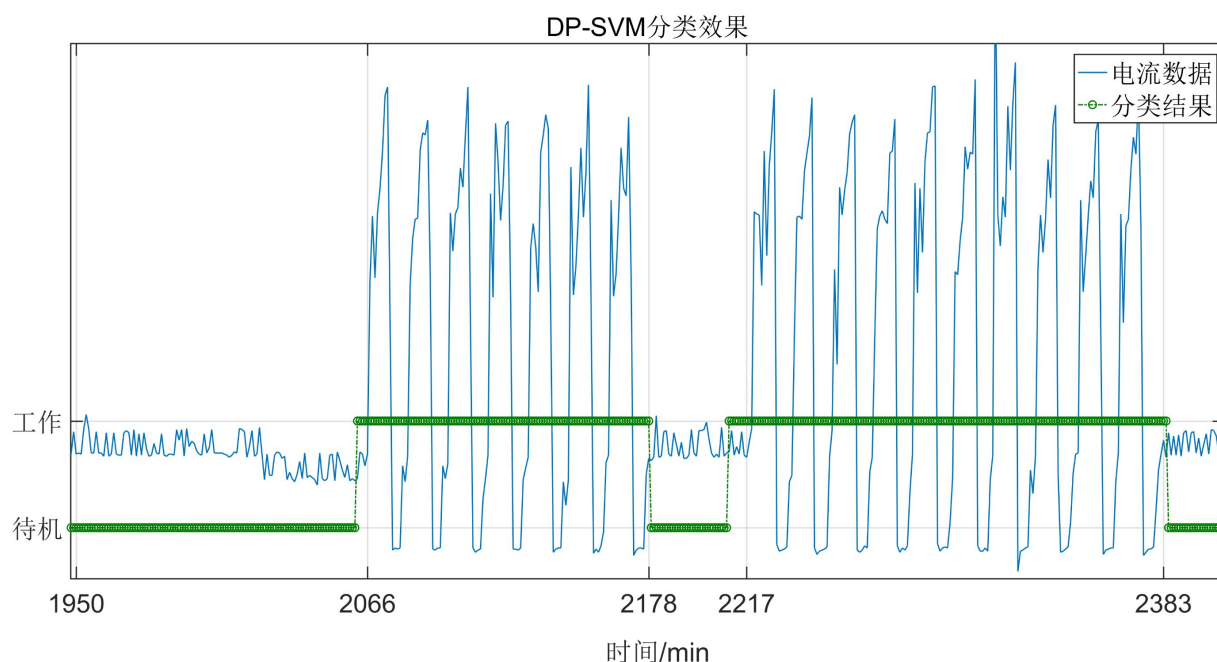


图 4.18 气相色谱-质谱联用仪电流分类局部细节

电流分类结果的局部细节如图 4.18 所示。可以看到，气相色谱-质谱联用仪在 2066 分钟后由待机状态变为工作状态，在 2178 分钟后转变为待机状态，在 2217 分钟后转变为工作状态，在 2383 分钟后再次转变为待机状态。DP-SVM 算法整体效果较好，但在 2066 分钟与 2217 分钟附近还存在几分钟的误差。

表 4.4 各仪器电流分类准确率及算法耗时详情

序号	分类准确率	耗费时间/s
1	0.995	42.197
2	0.973	24.938
3	0.982	28.315

分类准确率及算法耗时如表 4.4 所示。主要产生的误差有两种，一种是由于有时电流特征不明显导致分类错误，另一种是算法在状态变化边界的切分误差。不过总体来说，DP-SVM 算法在这 3 台仪器的电流分类上都取得了不错的效果。

4.5.3 电流分类对比及分析

为了体现本章算法优势，此处与直接使用固定窗口切分识别的方法（固定窗口法）进行对比。其中，固定窗口法使用与本章算法相同的电流分类模型，窗口大小选用常用的 15、30、60 进行测试。一共测试 5 次，以每天上报的电流数据为单位，结果采用平均准确率作为评价指标。

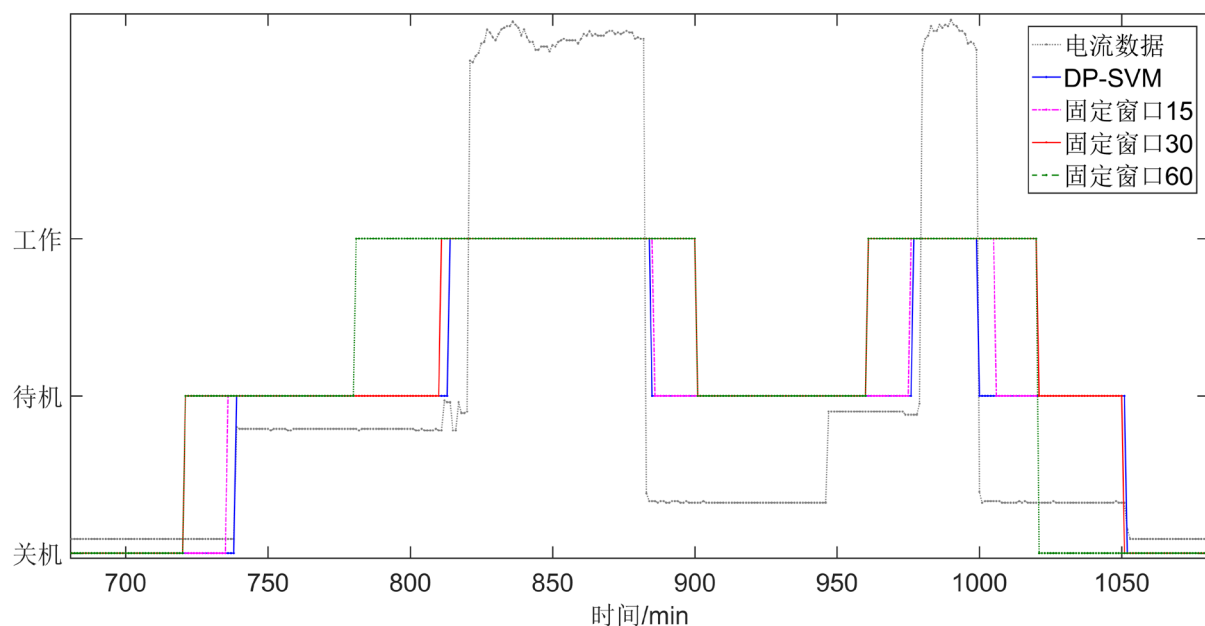


图 4.19 各算法分类部分结果

由于各算法对于仪器关机状态识别基本一致，为了清晰起见，图 4.19 主要展示了其中一次测试各算法对仪器 1 的电流数据分类结果中主要存在差异的部分。可以看到，DP-SVM 的效果最好，仅在 800 至 850 分钟之间存在一处较大的误差。固定窗口中，值为 15 的窗口误差最小，这是因为该仪器的各状态下电流比较平稳，即使是局部特征也能反映出整体的状态。而窗口值较大时，不恰当的切分是误差产生的主要原因。

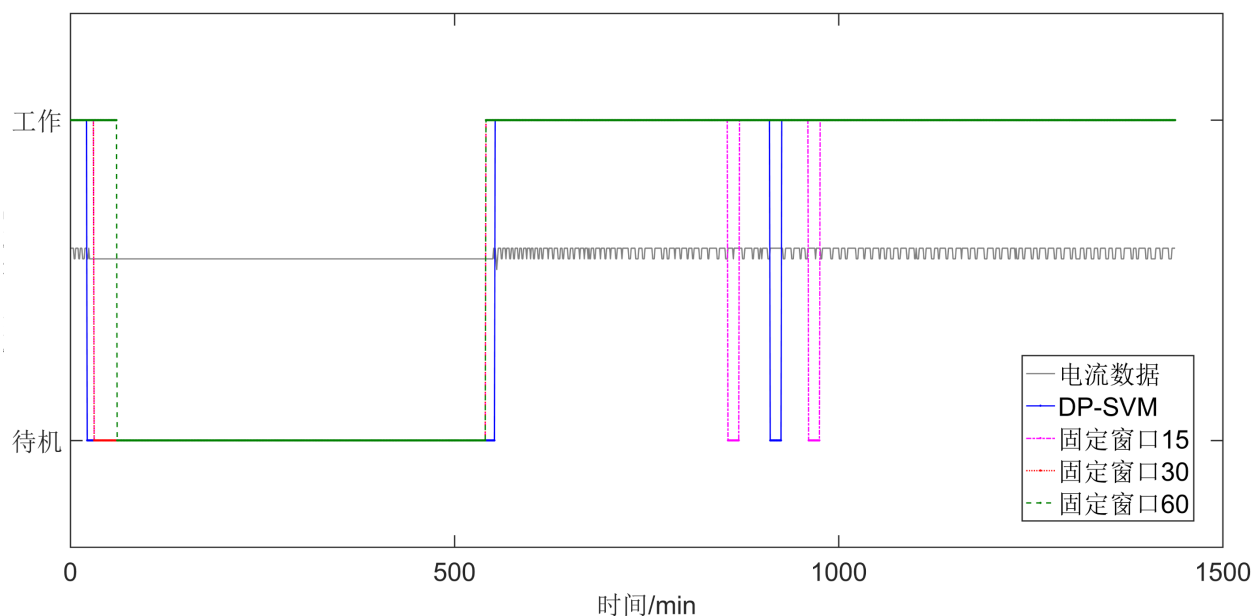


图 4.20 各算法分类结果

图 4.20 展示了其中一次测试各算法对仪器 2 电流数据分类结果。可以看到，DP-SVM 对于仪器状态变化处更敏感，但在 911 分钟处存在一处错误，这是因为仪器工作 15 分钟电流没有变化，此时使用特征与待机一致。固定窗口 15 和 60 的误差较大，误差主要来源分别是分类错误和切分误差。固定窗口 30 的效果相对更好，误差来源于仪器状态变化处的存在切分误

差。

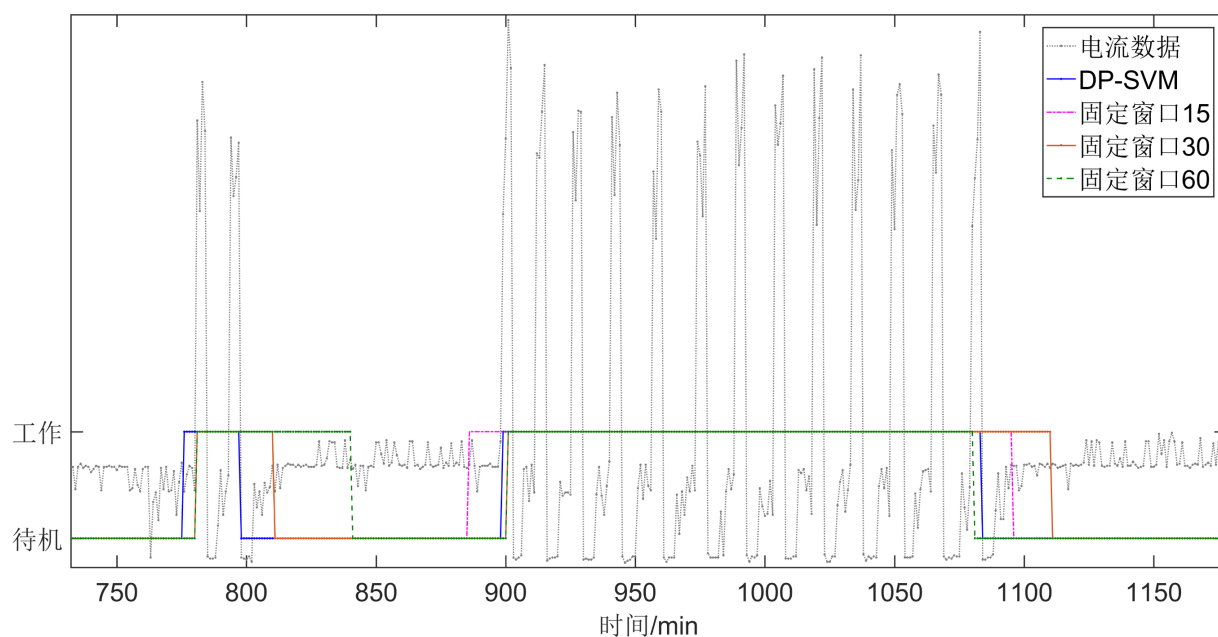


图 4.21 各算法分类部分结果

图 4.21 展示了其中一次测试各算法对仪器 3 电流分类结果中的主要差异之处。可以看到，DP-SVM 的效果最好，仅在 750 至 800 分钟之间存在一处较大的误差。固定窗口误差较大的原因如前文分析的，一是窗口值太小时导致切分电流数据太少，特征受各数据的干扰较大；二是窗口值太大时在仪器状态变化处会不可避免的切入多种数据，此时无论识别结果如何都会带来误差。

表 4.5 不同方法平均准确率对比

序号	窗口 15	窗口 30	窗口 60	DP-SVM
1	0.979	0.960	0.942	0.996
2	0.951	0.956	0.930	0.975
3	0.966	0.944	0.941	0.983

表 4.5 展示了各算法在 5 轮测试中的平均准确率，可以看到 DP-SVM 的效果要优于固定窗口。虽然固定窗口有时准确率也较高，但对于工作仅几个小时的仪器来说，误差几十分钟带来的影响也是不可忽视的。

4.5.4 机时计算测试

以天为单位，取各仪器使用一周的机时计算结果与人工统计结果进行对比来说明本章方法在机时计算上的效果，同样采用相对误差进行评价。同样以上述 3 台典型仪器来展示具体测试结果。

表 4.6 算法计时与人工计时

序号	方法	机时/min		
		关机	待机	工作
1	人工统计	766	90	584
	固定窗口 15	765	75	600
	固定窗口 30	750	60	630
	固定窗口 60	780	0	660
	DP-SVM	765	93	582
2	人工统计	0	166	1274
	固定窗口 15	0	180	1260
	固定窗口 30	0	150	1290
	固定窗口 60	0	120	1320
	DP-SVM	0	172	1268
3	人工统计	0	1106	334
	固定窗口 15	0	1065	375
	固定窗口 30	0	1020	420
	固定窗口 60	0	1080	360
	DP-SVM	0	1095	345

表 4.6 为其中一日各算法与人工统计的机时，可以看到，本章算法记录机时与人工统计差距不大，各机时误差基本不会超过 10 分钟。而固定窗口的误差均更大，最大误差达到了 90 分钟。

表 4.7 机时平均相对误差

序号	平均误差			
	固定窗口 15	固定窗口 30	固定窗口 60	DP-SVM
1	4.6%	8.3%	22.1%	1.2%
2	4.2%	6.1%	11.2%	2.4%
3	5.1%	13.6%	12.7%	1.6%

测试所得机时平均相对误差如表 4.7 所示，可以看到该算法记录机时可以达到较高的准确度。通常来说，电流分类准确率可以反映出机时计算的准确率，电流分类准确率越高则机时计算结果越准确。但当电流分类准确率不是那么高时，可能会出现误差抵消的现象，该现象可能会导致电流分类准确率较低时的机时计算结果要比电流分类准确率高时的机时计算结果更好。总体而言，本算法的电流分类效果与机时计算结果比使用固定窗口更准确。

最后，对剩余仪器进行测试，相关测试结果如表 4.8 所示。

表 4.8 相关测试结果

仪器编号	1	2	3	4	5	6	7
机时平均相对误差	2.0%	3.2%	1.1%	2.6%	4.0%	5.1%	3.7%

可以看到，测试结果中各仪器的机时结果均较为准确，平均相对误差基本不超过 5%，可以满足实际应用的需要。最终，系统中第三章算法无法处理的仪器均采用本方法进行了解决，从而实现应用的闭环。

4.6 本章小结

为了使系统能够处理更多种的仪器，本章提出了一种基于 DP-SVM 的机时计算方法，该方法由基于 SVM 的电流分类模型和基于 DP 的信号切分算法共同组成。在所提方法中，各仪器都需要根据自身电流特点选择合适的特征，然后通过网格搜索算法与 K 折交叉检验训练出各自的 SVM 电流分类模型。之后，一种基于 DP 的信号切分算法被提出并用于减少实际电流分类时依靠人工经验设置固定窗口切分带来的机时计算误差，使得机时计算结果更加准确。最后给出了基于 DP-SVM 的机时计算流程，并选取了多台仪器对本方法进行了验证，并以其中 3 台典型仪器为例详细展示了算法的电流分类以及机时计算的效果。在测试中，这 3 台仪器的电流分类准确率均达到了 0.97 以上，并在对比中测得机时平均相对误差小于 3%，优于使用固定窗口的方法。最后，从测试结果来看，本方法可以有效的对基于分箱与改进 K-means 算法做出补充，协助构建系统的闭环。

第 5 章 总结与展望

5.1 总结

本系统是浙江省大仪设备管理服务协作平台项目的研究成果之一，主要负责对浙江省多个高校、机构的大型科研仪器的运行电流数据进行采集以及对机时数据进行计算。本文详细阐述了整个系统的设计、搭建过程，并针对传统机时统计依赖于人工的痛点，在对大量科研仪器电流分析的基础之上提出了两种自动化的机时计算方法，为大型科研仪器管理的智能化提供数据支撑。本文主要的工作内容如下：

(1) 搭建了大型科研仪器机时智能计算系统。具体包括：一种非侵入式的电流采集终端，通过 NB-IoT 将仪器电流数据传输至电信 AEP 平台；针对 AEP 平台数据读取的限制，设计并搭建了本系统的数据存储中心。在该存储中心中，电流数据、仪器状态、仪器机时数据经过 RLE 算法压缩处理，按照一台仪器每日一个文档的存储结构存储，相比未压缩前节省了约 80% 的存储空间。搭建了基于 SpringBoot 与 Python 的后端业务架构，并设计了相关核心业务，最终实现了对各仪器机时的自动化计算与记录，并在 Web 端对相关内容进行了展示。

(2) 传统机时计算依赖于人工，存在较多问题。现有的电流阈值法实现机时自动计算在当前场景下具有一定的可行性，但在许多仪器上还是存在一些问题。为了在满足更多种仪器机时计算需要的同时尽可能降低系统实施成本，提高处理效率，在对大量仪器的电流特点研究后，提出了一种基于分箱与改进 K-means 的机时计算方法。通过分箱算法对电流数据进行预处理，提取了均值和标准差作为电流特征。为了可由系统执行及减少人工介入，从初始聚类中心、 k 值、聚类结果的选择上对 K-means 进行了改进，然后对电流样本聚类得出电流分类模型并保存至云端。后续上报电流数据经处理后，通过调用云端电流分类模型进行分类，统计各类电流数据个数再结合电流采集频率就实现了机时的自动计算。在多种仪器上进行了测试，并以其中典型的 6 台仪器为例详细展示了本章算法的效果及优势，表明了方法的有效性。最后，对系统应用的实际应用情况进行了分析，并给出了针对出错模型的处理流程以及最后效果。

(3) 由于仪器种类众多，对于基于分箱与改进 K-means 无法处理的仪器，提出了一种基于 DP-SVM 的机时计算方法进行补充。该方法首先通过仪器历史电流数据训练出电流分类模型，然后通过一种基于动态规划的信号切分算法减少了实际应用中使用固定窗口带来的切分误差。最后选取了多台仪器进行测试，并以其中较为典型的 3 台仪器进行展示开说明了本章方法不仅可以有效完成机时的自动计算的任务，并且效果要优于人工设置的固定窗口，得到的机时更加符合实际，可以很好地与基于分箱与改进 K-means 的机时计算方法一起形成业务闭环。

5.2 展望

本文设计系统虽然可以满足许多大型科研仪器机时计算的需要，但由于本人水平有限，还存在一些地方需要完善：

(1) 使用 SVM 模型的仪器需要通过人工标定数据以及训练模型，这些操作需要有一定经验的人才能完成，还无法简化给使用者操作。并且支持向量机部分还未加入 Web 端，只能通过后台人工的方式训练模型，不够方便，后续要将这部分功能加入到系统中。

(2) 聚类需要仪器所有的状态都在数据中有体现，若其中某种状态数据的缺失就会导致模型效果不佳，可能需要人工进行处理。后续可以研究自动迭代的机制，在数据不足的情况下可以通过不断迭代使得模型能自动改进。分箱过程可能会造成一定的切分误差，会对结果造成影响，后续需要对分箱参数的自适应性进行研究。

(3) 对于需要训练模型的仪器，还可以继续深入研究，尽量通过无监督或者更易操作的方式处理，尽量减少人工的任务量。

本文提出的系统除了计算、记录仪器机时以外，还可以在以下这些方面更进一步：

(1) 仪器的模型可以考虑在采集终端运行，不仅可以过滤无效数据从而降低流量成本，还可以用于仪器状态的实时判断。

(2) 仪器的状态还可以再进行细分，如许多仪器的工作状态是由多个小状态组合而成的，若能把控这些细节，后续可以延伸出仪器的故障检测、寿命分析、乃至用户使用分析等一些很有价值的功能。

(3) 仪器种类众多，可以研究模型的通用性，并且可以对那些适用相同模型的仪器进行关联，此后接入新仪器时就不需要人工再训练，直接使用相似仪器的模型即可，这也可以很大程度上减轻人工压力。

致谢

行文至此，意味着三年的研究生生涯即将画上一个句号。初来乍到之时的情景依旧历历在目，在本科及研究生这七年时光里，杭电给予我实在太多太多，难以言表，如今即将进入新的征程，特在此万分感谢。

桃李不言，下自成蹊。感谢我的导师陈科明教授，他学识渊博，治学严谨，工作勤勉，在工作中给予了我许多指导，是令人敬佩的榜样。本文从选题到修改再到初稿及终稿完成都离不开陈老师的悉心指导及合理建议。在此，谨向陈老师致以最崇高的敬意。

树高千尺，根在沃土。其次我要感谢我的父母，没有他们的辛勤劳作和无私奉献，也就没有一个农村孩子的今日。天道无常，父亲突然病倒，幸得医治，望日后他们身体健康，一切顺利。

良师益友，朝夕相处。感谢我的同门徐志杰、吴文楷、邢豪蔚、张百军、蔡坤、胡晓双和黄盼盼。感谢这三年的相互扶持，感谢你们提供的宝贵建议，在困难的时候为我提供的帮助。感谢我的女朋友，西湖、南高峰、五云山风景秀丽，欲买桂花同载酒，与君共作少年游。

寥寥数语，难诉衷肠。感谢论文评阅和答辩委员会的老师，感谢你们给予的耐心以及宝贵的意见。

最后，由衷的祝愿所有人，工作顺利，万事如意！

参考文献

- [1] 程豪, 周琼琼. 我国重大科研基础设施调查数据分析[J]. 今日科苑. 2018(5): 7.
- [2] 包郁明, 顾今吾, 李娟, 等. 浅谈新时代背景下大型科研仪器与设施开放共享工作[J]. 农业科技管理. 2020, 39(1): 4.
- [3] 李璐, 杨丽娟, 杨阳. 高校科研仪器设备资源整合与共享模式初探[J]. 产业科技创新. 2020, 2(15): 89-90.
- [4] 彭雪, 薛友林. 高校实验室仪器设备科学管理办法的探讨[J]. 教育教学论坛. 2020(11): 20-21.
- [5] 王浩. 高校大型仪器设备的管理现状及对策[J]. 当代化工研究. 2020(19): 157-158.
- [6] 权颖, 林秋星, 魏麟苏, 等. 高校大型仪器设备开放共享的问题分析及对策建议[J]. 大众标准化. 2021(20): 236-238.
- [7] 人民出版社. 国务院关于国家重大科研基础设施和大型科研仪器向社会开放的意见[M]. 国务院关于国家重大科研基础设施和大型科研仪器向社会开放的意见, 2015.
- [8] Lin Y, Tsai M. A novel feature extraction method for the development of nonintrusive load monitoring system based on BP-ANN[C]. In: Computer Communication Control and Automation (3CA), 2010 International Symposium on.IEEE, 2010. 215-218.
- [9] Liang J, Ng S K, Kendall G, et al. Load Signature Study—Part I: Basic Concept, Structure, and Methodology[J]. IEEE Transactions On Power Delivery. 2010, 25(2): 551-560.
- [10] Lin Y H, Tsai S M. Development of an Improved Time–Frequency Analysis-Based Nonintrusive Load Monitor for Load Demand Identification[J]. Instrumentation & Measurement IEEE Transactions On. 2014, 63(6): 1470-1483.
- [11] 王珏, 王晓玲, 刘荣斌, 等. 一种基于仪器电流数据的大数据分析方法[Z]. 2019.
- [12] 刘克冰. 基于深度学习的电力设备状态识别方法与应用研究[D]. 北京: 华北电力大学, 2018.
- [13] 邢晓莹. 面向大数据应用的室内安全监测智能终端设计[D]. 杭州: 杭州电子科技大学, 2017.
- [14] 燕续峰. 电器运行状态模型的研究及应用[D]. 上海: 上海交通大学, 2020.
- [15] 李玉功. K-means聚类算法的改进及其在彩色图像分割中的应用[D]. 合肥: 安徽大学, 2018.
- [16] Bablani A, Edla D R, Kuppili V, et al. A multi stage EEG data classification using k-means and feed forward neural network[J]. Clinical Epidemiology and Global Health. 2020, 8(3): 718-724.

- [17] Fahad A, Alshatri N, Tari Z, et al. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis[J]. IEEE Transactions On Emerging Topics in Computing. 2014, 2(3): 267-279.
- [18] 宋阳. 基于划分的聚类算法的改进及其应用[D]. 沈阳: 沈阳工业大学, 2019.
- [19] Kaufman L, Rousseeuw P J. Partitioning Around Medoids (Program PAM)[M]. Finding Groups in Data, 2008.
- [20] Ng M K, Li M J, Huang J Z, et al. On the Impact of Dissimilarity Measure in k-Modes Clustering Algorithm[J]. IEEE Transactions On Pattern Analysis & Machine Intelligence. 2007, 29: 503.
- [21] Wong J A H A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society. 1979, 28(1): 100-108.
- [22] 王庆圣. 基于FCM聚类的噪声图像分割算法研究[D]. 兰州: 兰州交通大学, 2021.
- [23] Guha S, Rastogi R, Shim K. ROCK: a robust clustering algorithm for categorical attributes[C]. In: Proceedings 15th International Conference on Data Engineering (Cat. No.99CB36337).1999. 512-521.
- [24] Guha S, Rastogi R, Shim K. CURE : An Efficient Clustering Algorithm for Large Databases[J]. Information Systems. 1998, 26(1): 35-58.
- [25] Deng D. DBSCAN Clustering Algorithm Based on Density[C]. In: 2020 7th International Forum on Electrical Engineering and Automation (IFEEA).2020. 949-953.
- [26] Hotait H, Chiementin X, Mouchaweh M S, et al. Monitoring of Ball Bearing Based on Improved Real-Time OPTICS Clustering[J]. Journal of Signal Processing Systems. 2020, 2: 1-17.
- [27] 张婕. 基于EM聚类分析模型的地中海生地化特征分析[D]. 厦门: 厦门大学, 2019.
- [28] 张万星. 基于高斯混合模型的逆变电源故障诊断研究[D]. 天津: 天津工业大学, 2021.
- [29] 王宏默. 基于SOM神经网络的通风机叶片裂纹故障诊断方法[D]. 北京: 北京工业大学, 2019.
- [30] 杨帅. 基于证据K近邻与网格聚类的WiFi定位算法设计与实现[D]. 广州: 华南理工大学, 2018.
- [31] 许名震. 基于改进CLIQUE算法的支持向量机的研究[D]. 哈尔滨: 哈尔滨工程大学, 2017.
- [32] Quinlan J R. Introduction of decision trees[J]. Machine Learning. 1986, 1(1): 81-106.
- [33] Azmi A U, Hadi A F, Anggraeni D, et al. Naive bayes methods for rainfall prediction classification in Banyuwangi[J]. Journal of Physics Conference Series. 2021, 1872(1): 12028.
- [34] Cortes C, Vapnik V. Support-Vector Networks[J]. Machine Learning. 1995, 20(3): 273-297.

- [35] Lecun Y, Boser B, Denker J, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation. 2014, 1(4): 541-551.
- [36] 李诗雨. 光纤振动传感信号的聚类[D]. 北京: 北京邮电大学, 2019.
- [37] 王文亮. 基于聚类分析的声波信号异常检测方法研究[D]. 北京: 北京化工大学, 2020.
- [38] 王江月. 基于不均衡样本的脉象信号分类方法研究[D]. 哈尔滨: 哈尔滨工业大学, 2018.
- [39] Kevric J, Subasi A. Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system[J]. Biomedical Signal Processing & Control. 2017, 31: 398-406.
- [40] 李威. 基于NB-IoT的环境智能监控系统[D]. 南京: 南京航空航天大学, 2020.
- [41] 李庆鸿. 基于NB-IoT智能灌溉系统研究与设计[D]. 淮南: 安徽理工大学, 2020.
- [42] 张浩雨. 基于NB-IoT的小区安防与环境监测系统设计与实现[D]. 西安: 西安科技大学, 2020.
- [43] 时薇. 基于NB-IoT的智能家居安防报警系统设计与实现[D]. 邯郸: 河北工程大学, 2020.
- [44] 张志伟. 基于IPv6和LwM2M的智能家居设备管理系统设计[D]. 武汉: 华中科技大学, 2019.
- [45] 余金龙. 基于云平台的造纸机远程健康管理系统设计[D]. 西安: 陕西科技大学, 2018.
- [46] 贾俊婷. 云计算综述及云计算在通信行业的应用[D]. 呼和浩特: 内蒙古大学, 2014.
- [47] 刘仪轩. 基于MongoDB的排水设施监管养护系统的应用研究[D]. 上海: 东华大学, 2020.
- [48] 易琳. 传感器网络中精简数据传输算法设计与实现[D]. 北京: 北京理工大学, 2015.
- [49] 林翔. 基于SpringBoot技术的M公司厨电产品质量管理系统的设计与实现[D]. 成都: 电子科技大学, 2021.
- [50] 安建伟. 仪器设备状态智能监测系统设计[D]. 杭州: 杭州电子科技大学, 2021.
- [51] 韩琮师. K-means聚类算法优化及其在电商平台精准营销中的应用研究[D]. 青岛: 山东科技大学, 2020.
- [52] 孙茂斌. 基于动态时间规整的时序数据相似度量方法研究[D]. 重庆: 重庆邮电大学, 2020.
- [53] Arthur D, Vassilvitskii S. k-means: the advantages of careful seeding[C]. In: Society for Industrial and Applied Mathematics, 2007. 1027-1035.
- [54] Rodriguez A, Laio A. Machine learning. Clustering by fast search and find of density peaks[J]. Science. 2014, 344(6191): 1492-1496.
- [55] 李亚岚. 基于改进K-均值算法的电力负荷数据聚类分析方法的研究[D]. 兰州: 兰州

理工大学, 2020.

- [56] Nitish N A, Singh A K. Automatic detection classification and area calculation of brain tumour in MRI using wavelet transform and SVM classifier[J]. International Journal of Intelligent Systems Technologies and Applications. 2020, 19(6): 526.
- [57] Kuila S, Dhanda N, Joardar S. Classification of Heart Arrhythmia in ECG Signals using PCA and SVM[J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE). 2020, 9(10): 193-198.
- [58] Li J, Bai J, Yu B. C-Mn yield forecasting model based on SVM[J]. IOP Conference Series: Earth and Environmental Science. 2021, 781(2): 22024-22025.
- [59] Agalya V, Sumathi S. Heart Signal analysis on Multi-Domain Features Extraction by SVM Classifier in Smart Monitoring System[J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE). 2019, 8(12): 1841-1843.
- [60] 张强. 基于支持向量机的热轧板带轧制力预测[D]. 西安: 西安建筑科技大学, 2021.
- [61] 辛文超. 基于张量分解的音频信号分类研究[D]. 包头: 内蒙古科技大学, 2020.
- [62] 邢浩. 基于动态规划的雷达弱目标检测前跟踪算法研究[D]. 大连: 大连海事大学, 2020.
- [63] Tumen R S, Sezgin M. Segmentation and Recognition of Offline Sketch Scenes Using Dynamic Programming[J]. IEEE Computer Graphics and Applications. 2021, PP(99): 1.
- [64] 石静. 基于自适应动态规划的非线性系统最优控制及其在微电网中的应用[D]. 南京: 南京邮电大学, 2020.

附 录

作者在读期间发表的学术论文及参加的科研项目

参加的科研项目：

- [1] 传感器设备采购、安装运维及数据解析. 浙江省科技信息研究院 (ZJXL-KJXX-201903). 2019.12-2021.12
- [2] 物联网传感器研发、运维及数据解析升级项目. 浙江省科技项目管理服务中心 (ZJXL-KJXX-202101). 2021.10-2022.10

发表的专利：

- [1] 陈科明, 刘剑, 张百军, 吴文楷. 一种基于电流的大仪工时计算方法[P]. 浙江省: CN114280352A, 2022.04.05. (已公开, 2/4, 导师第一发明人)
- [2] 刘剑, 曾佳, 林董源, 陈科明. 一种基于电流数据的仪器状态检测方法[P]. 浙江省: CN112924743A, 2021.06.08. (已公开, 1/4)