# 中科院 SAMP 系统仪器运行时间分析与优化①

华 维[1,2], 郭锐锋[2], 尹震宇[2], 彭阿珍[1,2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)

**摘 要**: 中科院 SAMP(大型仪器设备共享管理)系统实现了仪器设备信息化管理的方式, 极大的提高了仪器设备的管理和使用效率. 然而, 当前的仪器设备共享管理系统在某些方面依然存在资源配置不科学的现象, 如仪器设备管理系统无法监控仪器设备开机之后空机运行、仪器异常时间开机运行等. 这种情况迫切的需要我们对仪器设备的使用记录进行分析, 找出其中的异常运行数据并对其进行优化. 本文将采用贝叶斯分类和 logistic 回归的数据统计方法来对仪器设备共享管理系统仪器运行时间数据进行分析, 并对仪器运行时间数据进行预测分类. 将预测分类结果和实际运行时间结果进行对比, 找出异常运行时间结果并分析产生异常结果的原因, 再结合实际的仪器设备运行情况对仪器设备的管理进行优化, 从而达到提高仪器设备使用效率的目的.

**关键词**: 仪器设备; 运行时间; 贝叶斯分类; logistic 回归; 分类

## Analyzing and Optimizing Instrument's Running Time in Chinese Academy of Sciences SAMP System

HUA Wei[1,2], GUO Rui-Feng[2], YIN Zhen-Yu[2], PENG A-Zhen[1,2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

**Abstract**: Information management has been used in the SAMP system of Chinese Academy of Sciences, which greatly improves efficiency of the use and management of the equipment. However, the SAMP system cannot scientifically allocate resources in some ways, such as the system couldn't monitor the equipment while it is just running for doing nothing or is starting up at abnormal time. By this, the user record of the instrument and equipment must be analyzed immediately in order to find out the abnormal data and optimize the use efficiency of the equipment. In this paper, these data will be analyzed and forecast by the method of Bayes classification and Logistic Regression. Then the forecast data with be compared with the real results to find out the abnormal data and the reason how the abnormal data generate. Finally, the management of the instrument and equipment will be optimized base on the practice situation. In this way, the use efficiency of the instrument will be improved.

**Key words**: instrument and equipment; running time; Bayes classification; logistic regression; classification

在仪器设备不断的发展趋势下, 仪器设备的数量和型号也在不断增长[1], 而仪器设备的增长加重了仪器设备的管理维护工作. 传统的设备管理办法是设备采购完成以后, 将设备的基本情况和相关信息登记存档, 然后将档案存档[2], 以后档案基本处于无人维护状态, 如设备变迁、维修情况、设备当前运行状态等信息无法实时更新[2]. 信息化技术的发展加速了传统仪器设备管理的升级, 越来越多的高校、科研院所和公司都设计了自己的仪器设备管理系统, 取代了传统的人工管理模式, 利用计算机、通信技术与其他的办公设备相结合的方式对仪器设备信息进行收集、加工、存储、更新、维护等工作. 用户能随时通过仪器设备

① 基金项目:"数控系统功能安全技术研究"国家科技重大专项(2014ZX04009031)
收稿时间:2015-09-18;收到修改稿时间:2015-10-26

管理系统来查询设备当前运行状态和历史记录, 提高了工作效率和质量, 使管理人员从手工统计工作中解脱出来[2].

中科院 SAMP(大型仪器设备共享管理)系统实现了仪器设备信息化管理的方式, 极大的提高了仪器设备的管理和使用效率. 然而, 当前的仪器设备共享管理系统在某些方面依然存在资源配置不科学的现象, 如仪器设备管理系统无法监控仪器设备开机之后空机运行状况、仪器异常时间开机运行等情况, 降低了仪器设备的使用效率, 增加了仪器设备的运维成本.

针对仪器设备管理系统服务器数据空置浪费的情况, 本文提出用贝叶斯分类和多维 logistic 回归[3]的数据统计方法对仪器设备的运行时间进行统计, 分析出服务器中的异常数据, 并找出异常数据产生的原因, 并将异常数据报告给仪器设备管理人员, 管理人员再结合实际的仪器设备管理系统的实际运行情况进行优化调整, 从而达到提高仪器设备使用率的目的.

## 1 实验原理

### 1.1 朴素贝叶斯分类原理

朴素贝叶斯分类法[4]工作过程如下:

(1) 设 D 是训练元组和相关联的类标号集合. 每个元组用一个 n 维属性向量 $X=\{x_1,x_2,\cdots,x_n\}$ 来表示, 描述有 n 个属性 $A_1,A_2,\cdots A_n$ 对元组的 n 个测量.

(2) 假定有 m 个类 $C_1,C_2,\cdots,C_m$. 给定元组 X, 分类法将预测 X 属于具有最高后验概率(条件 X 下)的类. 也就是说, 朴素贝叶斯分类法预测 X 属于类 Ci, 当且仅当

$P(C_i \mid X) > P(C_j \mid X)$其中$1 \le j \le m, j \ne i$

这样, 最大化 $P(C_i \mid X)$. 其 $P(C_i \mid X)$ 最大的类 Ci 称为最大后验假设. 根据贝叶斯定理

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \tag{1}$$

(3) 由于 P(X)对于所有类为常数, 只需要公式(1)中 $P(X \mid C_i)P(C_i)$ 最大即可. 并将该数据分类到对应的后验概率较大的类别中去即可.

### 1.2 logistic 回归分类原理

logistic 回归分类原理如下: 当需要进行分类的数据是二值逻辑时, 即分类结果 $y \in \{0,1\}$ 时, 有如下所示的 Logistic 回归判别函数:

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \tag{2}$$

其中参数 θ 是要进行拟合的参数, x 是特征数据值, $h_\theta(x)$ 为拟合预测结果. 在利用 logistic 回归法进行拟合时, 定义误差函数为:

$$E = \sum_{i=1}^{n}(y^{(i)} - h_\theta(x^{(i)})) \tag{3}$$

利用梯度下降法[5]最小化误差, 从而求解参数 θ, 梯度下降算法如下:

Begin initialize a, 阈值 θ, □(·), k←0

Do k←k+1

a←a- □(k)∇ J(a)

Until |□(k)∇ J(a)|<θ

Return a

End

利用上述 logistic 回归算法对实验数据进行拟合, 并将拟合结果分类到与其结果相近的分类中.

## 2 实验步骤

实验步骤主要包括实验数据提取、数据清洗、数据拟合和数据预测等四部分.

### 2.1 数据提取

首先, 分析中国科学院仪器设备共享管理平台系统服务器存储的中国科学院不同科研院所 10 年的仪器设备运维数据, 由于中国科学院仪器设备共享管理平台系统运行时间长, 监控设备数量多, 仪器设备类型多样, 故有大量的不同数据类型存储在服务器数据库中. 在本实验中, 选取了生物物理所的低速大容量冷冻离心机 Hitachi CR7-1 和病理分析系统 Lecia 作为分析样本. 在仪器设备管理系统中导出样本数据, 并提取出影响仪器设备运行时间的相关因素.

### 2.2 数据清洗

按照 2.1 节提取出合理的样本数据, 然后样本数据进行清洗操作. 数据清洗步骤为:

(1) 将数据按照 2015 年年份划分成两部分, 其中 2015 年之前的数据作为函数拟合样本数据, 2015 年之后的数据作为预测样本数据;

(2) 删除仪器运行时间小于最小运行时间的样本数据;

(3) 合并按照分段存储的仪器运行时间数据;

(4) 对于固定时间运行的仪器, 对其数据进行优化操作.

(5) 对于非固定时间运行的仪器, 将仪器运行时间划分成不同的区间进行统计.

### 2.3 数据拟合

根据不同的仪器设备, 采用不同的数据拟合方法. 具体实现方法如下所示.

(1) 低速大容量冷冻离心机 Hitachi CR7-1 使用时间较短, 使用时间长度不固定, 需要对连续性数据进行分段统计, 本仪器的数据特征适合采用朴素贝叶斯统计的方法对数据进行统计分析. 其中, 仪器开始使用时间、仪器结束使用时间和仪器使用时间长度可以作为朴素贝叶斯中的属性值, 仪器使用时间是否正常作为分类的类别.

(2) 病理分析系统 Lecia 使用时间长, 使用时间长度固定, 仪器的使用时间点基本固定在整点, 这种数据特征适合将原始数据进行离散化处理, 由于分类只有两种情况, 故可以采用 logistic 回归分类原理, 将仪器设备的使用时间类型进行分类.

### 2.4 数据预测

利用上述步骤中 2015 年之前的数据拟合出来的方程对 2015 年之后的数据进行预测, 并计算出预测结果和实际结果之前的差别, 分析误差产生原因并对函数进行优化, 直到误差在给定要求之内.

## 3 设计与实现

### 3.1 实验设计

本实验选取生物物理所的低速大容量冷冻离心机 Hitachi CR7-1 和病理分析系统 Lecia 作为样本数据进行分析, 选取该仪器设备作为样本的主要原因为: ①样本数据丰富, 可以得到更好的拟合函数; ②低速大容量冷冻离心机 Hitachi CR7-1 使用时间长度不固定, 使用时间较短, 使用该仪器拟合得到的算法适合其他类似的仪器设备; ③病理分析系统 Lecia 使用时间长, 使用时间长度固定, 可以作为类似仪器的代表. 实验具体流程如图 1 所示.

如图 1 所示, 实验主要针对两种不同类型的实验仪器运行时间进行统计分析. 具体实验为:

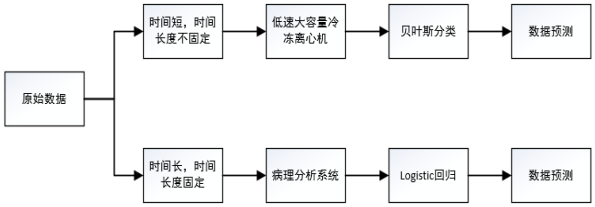


图 1 实验流程图

#### 3.1.1 低速大容量冷冻离心机 Hitachi CR7-1 实验

为了对速大容量冷冻离心机 Hitachi CR7-1 的运行时间数据进行概率统计分析, 本文设计了如图 2 所示的实验. 步骤如下:

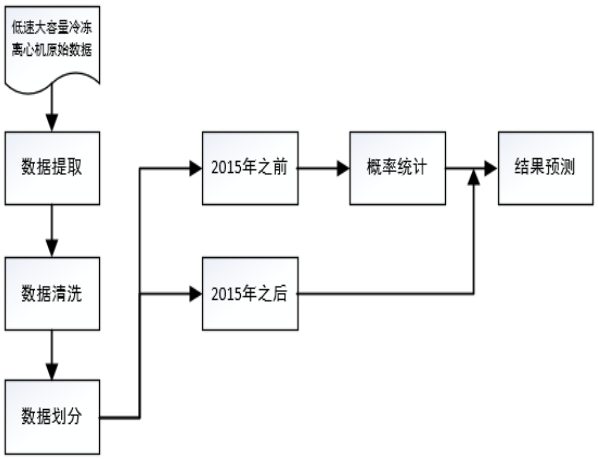


图 2 低速大容量冷冻离心机实验流程

①首先从中科院仪器设备共享管理系统中获取低速大容量冷冻离心机 Hitachi CR7-1 的使用日志记录作为实验的原始数据. 原始数据中包括仪器委托单编号、检测日期、检测开始时间、检测结束时间、检测时间、工作内容、使用人、日志状态和数据类型等数据; ②从原始数据中提取仪器委托单编号、检测日期、检测开始时间、检测结束时间、检测时间和数据类型等信息, 删除其中检测时间小于仪器运行最少时间和最大时间的数据, 并将仪器运行时间划分成不同的区间, 完成数据的清洗工作; ③数据清洗完成之后, 将仪器设备运行时间数据按照检测日期划分成 2015 年之前和 2015 之后两部分, 其中 2015 年之前的数据用于统计仪器运行的先验概率, 2015 年之后的数据用于进行结果预测分析; ④分别对仪器的检测开始时间、检测结束时间和检测时间进行统计, 计算在不同使用时间的先验概率; ⑤将 2015 年之后的数据带入到贝叶斯公式中, 计算不同的数据所属的类型, 从而实现仪

器运行数据使用时间预测功能. ⑥对比预测类型和实际类型, 计算该算法的准确率.

3.1.2 病理分析系统 Lecia 实验

病理分析系统 Lecia 的特点是使用时间长, 时间长度固定, 本文针对这种特殊类型的仪器设计了如图 3 所示的实验. 步骤如下:

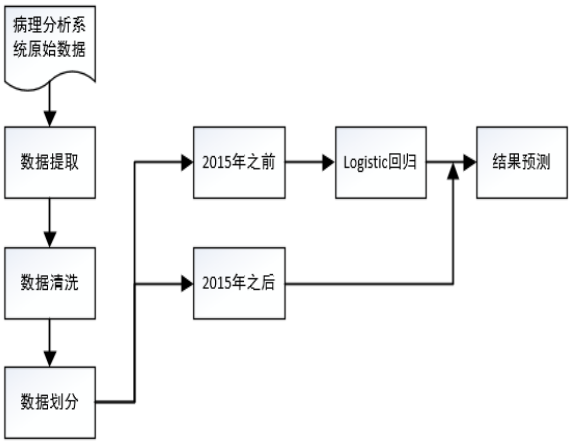


图 3 病理分析系统实验流程

①首先从中科院仪器设备共享管理系统中获取病理分析系统 Lecia 的使用日志记录作为原始实验数据. 原始实验数据中包括仪器委托单编号、检测日期、检测开始时间、检测结束时间、检测时间、工作内容、使用人、日志状态和数据类型等数据; ②从原始数据中提取仪器委托单编号、检测日期、检测开始时间、检测结束时间、检测时间和数据类型等信息, 删除其中检测时间小于仪器运行最少时间和最大时间的数据, 并将仪器运行时间进行整理, 将仪器的使用时间误差小于 0.1 的误差范围时间合并到相连的数据中去, 另外, 对于分段记录的仪器运行时间数据要进行合并操作, 使多条记录数据合并成为一条有效的记录; ③数据整理完成之后, 将仪器设备运行时间数据按照检测日期划分成 2015 年之前和 2015 之后两部分, 并对 2015 年之后的数据进行 logistic 回归分析. 其中, 仪器的检测开始时间、检测结束时间和检测时间作为 logistic 函数拟合因子; ④利用 logistic 回归原理对数据进行拟合, 其中, 设定函数在累加误差范围小于值 1 或者函数迭代次数大于 10000 时, 函数迭代终止, 得到函数拟合参数; ⑤将 2015 年之后的数据带入上述统计结果中进行预测, 并比较预测结果和实际结果, 并计算预测结果的准确率.

## 3.2 结果分析

3.2.1 低速大容量冷冻离心机 Hitachi CR7-1 实验结果

利用上述算法分析低速大容量冷冻离心机, 将原始样本数据进行统计得到如图 4 所示的相对次数图.

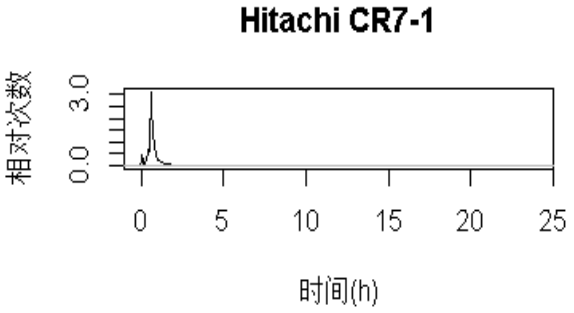


图 4 Hitachi CR7-1 相对次数原始图

从图 4 中可以得到仪器使用时间长度基本处在 0.69 小时附近, 另外从图中可以得到, 在时间长度为 0 小时附近出现了小尖峰, 出现这种情况可能是测试数据造成的, 需要进行去噪处理. 另外, 以及 2.5 小时之后有部分数据, 可以通过函数迭代进行优化. 在对上述进行处理之后可以得到如图 5 所示的数据图.

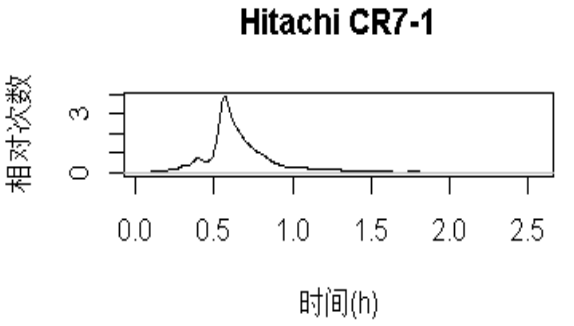


图 5 Hitachi CR7-1 相对次数去噪图

将上述得到的数据带入算法中进行概率统计并用计算结果对 2015 年的数据进行预测. 实验中原始数据为 5988 条, 其中 2015 年之前的数据为 5236 条, 2015 年之后的数据为 752 条, 数据清洗之后参与算法迭代的数据为 4646 条. 预测正确条数 629 条, 其中正确分类 465 条, 错误分类 164 条.

3.2.2 病理分析系统 Lecia 实验结果

利用上述算法分析病理分析系统, 将原始样本数

据进行统计得到如图 6 所示的时间频次图.

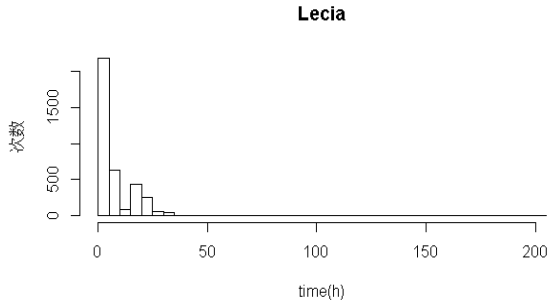


图 6　Lecia 时间频次原始图

从图 6 中可以看到, 数据主要出现在仪器使用时间小于 50 小时之内, 通过统计计算得到, 大部分数据的仪器运行时间都小于 35 小时, 故可以对原始数据进行去噪操作, 完成去噪之后得到如图 7 所示的时间频次图.

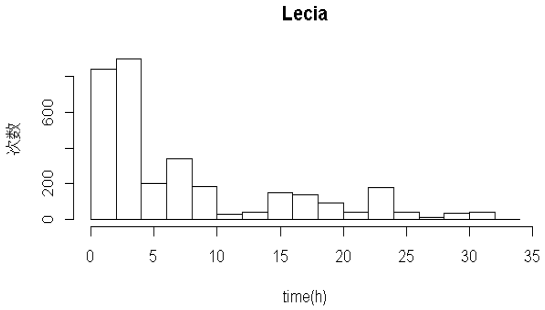


图 7　Lecia 时间频次去噪图

将上述得到的数据带入算法中, 利用 logistic 回归对数据进行拟合, 并用得到的 logistic 方程对 2015 年的数据进行预测, 其中原始数据为 4463 条, 经过清洗之后为 3631 条, 参与算法运算的 2015 年之前的数据为 3268 条, 进行预测分析的 2015 年之后的数据为 363 条, 其中 250 条分类为正常数据, 112 条分类为错误数据. 其中正常数据预测正确为 241 条, 错误数据预测为错误 83 条.

本文针对算法的评价主要采用分类的准确率进行计算, 计算公式如下:

$$\text{准确率} = \frac{\text{条数}\{\text{预测分类} = \text{实际分类}\}}{\text{总的条数}} \quad (4)$$

利用公式(4)分别计算低速大容量冷冻离心机 Hitachi CR7-1 和病理分析系统 Lecia 的分类结果, 得到如表 1 所示的结果:

表 1　仪器设备实验时间预测准确率

| 准确率 | 正常数据 | 错误数据 | 总数 |
|---|---|---|---|
| Hitachi | 94.1% | 63.6% | 83.6% |
| Lecia | 96.4% | 74.1% | 89.3% |

从准确率结果中, 可以得到这两种分析方法的分类准确率都达到了 80%以上, 基本能够满足实际仪器运行时间数据的分析要求. 另外, 对于本实验的算法还可以应用到系统中其他类型相同的仪器设备的数据统计分析中, 实现对系统中所有仪器的整体分析.

## 4　结语

本文通过分析中科院仪器设备共享管理系统中不同类型的仪器设备使用时间数据, 提出了用朴素贝叶斯和 logistic 回归分类的数据统计方法对仪器设备使用时间进行分析. 实验中选取了中科院生物物理所两个典型的实验设备低速大容量冷冻离心机 Hitachi CR7-1 和病理分析系统 Lecia 的数据, 并对两个不同类型的设备采用不同的分类方法进行拟合. 实验表明, 该算法能够有效的分辨出仪器设备运行时间数据的是否正常, 实际可用性高, 对分析该系统的仪器设备使用时间有重要意义, 同时仪器设备管理人员可以利用该统计分析结果对实际的仪器设备管理进行优化, 从而达到提高仪器设备使用效率的目的.

## 参考文献

1 李秀坚.基于 B/S 的高校仪器设备管理系统的设计与实现[硕士学位论文].厦门:厦门大学.2014.

2 杨晓雁.仪器设备管理系统的设计与实现[硕士学位论文].西安:西安电子科技大学.2010.

3 Matsui H. Variable and boundary selection for function data via multiclass logistic regression modeling. Computational Statistics and Data Analysis. 2014, 78: 176−177.

4 Han JW, Kamber M. 数据挖掘概念与技术.北京:机械工业出版社,2007:200−205.

5 Duda RO, Hart PE, Stork DG. 模式分类.北京:机械工业出版社,2009:183−185.