第29卷 第4期
2005年4月

高 能 物 理 与 核 物 理
HIGH ENERGY PHYSICS AND NUCLEAR PHYSICS

Vol.29, No.4
Apr., 2005

应用数据挖掘的束流状态描述建模*

谢东[1)] 李为民 宣科 何多慧
(中国科学技术大学国家同步辐射实验室 合肥 230029)

摘要 聚类分析是描述数据群体特征的有效方法. 然而,随着挖掘数据库规模的增大,评分函数极值搜索是计算复杂度很高的问题. 文中提出了一种新的算法,并将其运用到束流状态描述建模中. 试验结果表明,该算法快速有效. 同时,重复性是电子储存环的重要指标,聚类模型可作为机器研究和决策的依据,具有指导意义.

关键词 数据挖掘 描述建模 聚类 搜索 束流状态

1 引言

电子储存环经过长期稳定运行后,产生了大量的数据. 数据库知识发现(knowledge discovery in database,KDD)能够帮助发现有价值的信息,描述数据特征和关系. 数据挖掘是数据库知识发现的核心技术,应用数据挖掘束流状态描述建模具有重要意义,其结果有助于机器研究和决策支持.

然而,电子储存环经过长年累月稳定运行后,积累下来的数据量是相当惊人的. 采用一般的聚类技术使得分析起来十分困难. 本文提出一种着眼于寻找聚类边界的聚类算法,数据经过一系列预处理后,类比质心系统利用牛顿力学的柯尼希定理优化搜索过程,寻找聚类划分并计算聚类中心.

2 数据挖掘技术介绍

随着计算机时代的到来和各行各业信息化、数字化的发展,产生了大量的数据. 传统的数据库技术和统计方法已不能满足人们对数据进行更高层次分析和利用的要求,导致了"数据爆炸但知识贫乏"的现象,迫切需要新的技术和自动化工具来帮助人们将海量数据转化为有用的信息和知识,数据挖掘技术应运而生.

数据挖掘(DM,data mining)又称为数据开采,就是从海量的数据中提取隐含在其中人们事先未知的,但又是潜在有用的信息和知识,并将其表示成最终能被人理解的模式的高级过程. 数据挖掘产生于80年代末期,是目前国际国内的研究热点. 数据挖掘的显著特点就是它强大的数据处理能力,能从大量的数据中发现有用的规律、规则、联系、模式等知识. 包括聚类分析、分类分析、回归分析、孤立点检测、时间序列相似性搜索、文本信息检索和关联度分析等.

本文的数据挖掘任务采用的技术是聚类. 聚类分析源于许多研究领域,包括数据挖掘,统计学,生物学,以及机器学习. 聚类是将数据对象分组成多个类或簇,在同一簇中的对象之间具有较高的相似度,而不同簇中的对象差别较大. 评分函数是聚类间相似性的度量. 聚类技术寻找聚类的划分和聚类中心,是搜索评分函数极值(通常是局部极小值)的过程.

3 数据预处理[1]

存在不完整的、含噪声的和不一致的数据是大

2004-07-14收稿
* 国家计委重大科研项目"国家同步辐射实验室二期工程"资助
1) E-mail:xiedong@mail.ustc.edu.cn

型数据库的共同特点,同时数据库的结构往往是为事务处理目的而设计,对于特定数据挖掘任务的数据预处理总是必须的. 数据预处理占数据挖掘的大部分工作量. 本文的数据预处理主要包括 3 个过程:数据清洗、数据转换和离散化规约. 数据清洗是识别和除去异常;数据转换是对数据进行规格化操作,就是将数据限定在特定范围;离散化规约通过将属性域划分为区间,得到数据集的规约表示,它小得多,但仍接近于保持数据的完整性,在处理后的数据集上挖掘更有效,并产生相同的分析结果.

3.1　数据清洗

数据清洗可以改进数据的质量,从而有助于提高其后的挖掘过程的精度和性能. 本次数据挖掘任务数据源是 HLS 运行数据库中数据测量子系统从 2004/01/01 到 2004/02/29 的数据. 在数据库中,存在两个或多个相同记录将影响数据的分布,数据清洗检测并消除冗余. 采用的方法是对重复数据取平均值. 另一个方面,在实际运行过程中,束流测量服务器程序可能会错误地发送束流流强数据,可能是负的,也可能过大,这些都无意义. 处理方法是检测并清除这些错误数据.

3.2　数据转换

从 2004/01/01 到 2004/02/29,合肥电子储存环主要运行通用光源模式. 制定数据转换步骤如下:

1)限制束流寿命在[1h,25h]范围内. 统计分析发现:寿命低于 1 个小时的数据约占 0.48%,是由于束流瞬间损失或者计算寿命程序产生故障. 高于 25h 的数据约占 0.02%,是低束流运行状态或者其他原因造成. 数据转换删除这个范围之外的数据.

2)限制流强范围为[20mA,400mA]. 统计发现:在机器研究或者实验站调试阶段有低束流运行状态,其流强(< 20mA),约占 21.0%;高流强(>400mA)占低于 0.1%. 超出此范围的数据不感兴趣,在此删除. 数据清洗在关系数据库中通过关系运算完成.

3.3　束流流强离散化处理的数据规约

束流状态是机器研究人员非常关心的问题,如何表示机器状态? 束流状态的特征是什么? 束流测量系统每隔 2s 测量一次,对应数据库中一条记录. 图 1 是 2004/05/30 束流流强历史数据库查询界面,该图以当天时间(hour)为横坐标,束流流强和束流寿命为纵坐标. 我们知道,束流寿命是电子储存环的重要参数,束流寿命被量子效应、气体散射和托歇克(Touschek)效应等因素限制,在合肥电子储存环中,导致束流损失的最主要原因是托歇克效应. 托歇克寿命[2]:

$$\tau_{\text{Touschek}}^{-1} = \frac{\sqrt{\pi} r_e^2 cN}{\delta p_x (\Delta p_{\text{ref}})^2 h V_b} C(\varepsilon), \qquad (1)$$

N 为束流的粒子总数.如果机器状态稳定,在同样束流流强和同样条件下,束流寿命基本相同,即同样流强而束流寿命不同就对应不同束流状态. 于是,束流状态可以反映在以束流流强为横坐标,束流寿命为纵坐标的图中,每次运行对应图上一个序列. 如图 2,散点图中的数据显示了束流寿命从几个小时到十几个小时不等,束流状态明显地有两种聚类划分.

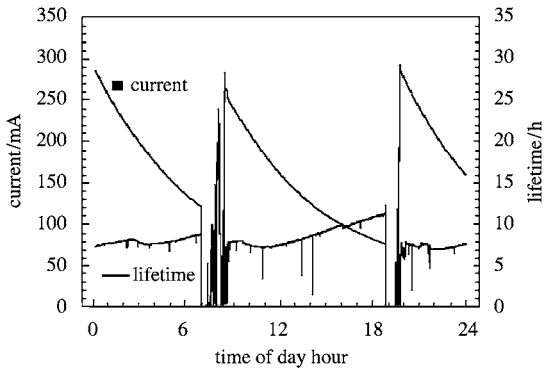

图 1　2004 年 5 月 30 日束流流强、束流寿命随时间的变化

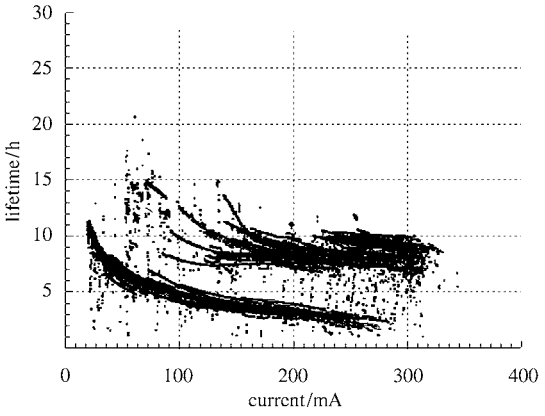

图 2　2004 年第 1、2 月份束流寿命随束流流强的变化

本文束流状态分析方法为:将采集到的数据按束流流强离散化处理,流强窗口宽度 ΔI = 1mA. 在

该模型下,模型参数(束流寿命 τ)反映了某特定束流强下机器状态特征. 数据挖掘能够回答以下问题:特定束流流强情形束流寿命如何聚类? 聚类中心是什么? 数据挖掘能指导特定束流流强情形机器状态聚类划分,因此可以作为机器研究的根据.

3.4 束流寿命散化处理的数据规约

经过以上预处理后,由于束流寿命从几个小时到十几个小时不等,对于特定束流流强,在流强窗口1mA内每次机器运行往往对应很多测量值,得到的数据集仍然将非常大. 在海量数据上进行复杂的数据分析和挖掘将需要很长时间,采用离散化数据规约技术可以用来减少给定连续属性值的个数,节省挖掘时间. 采用的方法是:将束流寿命划分为宽度为0.1h的等区间,区间中每一个数据的束流寿命被其中平均值替换. 数据按束流寿命离散化处理后,尽管细节丢失了,但数据更有意义,更容易解释,同时消除了噪声. 与在大的、未处理的数据上挖掘相比,所需I/O(input/output)更少,并且更有效.

另外,对于某次运行特定束流流强区间中的数据个数为

$$\begin{cases} I = I_0 \mathrm{e}^{-t/\tau} \\ \Delta N = \lambda \times \Delta t \end{cases} \Rightarrow \Delta N = \lambda \times \frac{\tau}{I} \times \Delta I, \quad (2)$$

λ为数据采集速率.可见,不同束流流强、不同束流寿命的束流曲线在数据库中的密度不一样,即流强越小、寿命越大的情形,在同样束流流强窗口1mA中对应的点数越大. 这样直接用于数据挖掘是不合理甚至是不对的. 为保证对于特定束流流强每次运行贡献同样多个数据用于数据挖掘,需要将数据重新调整,于是每个(束流流强 I,束流寿命 τ)窗格中数据个数变换公式为

$$\Delta N' = \Delta N \times \frac{I}{\tau}. \quad (3)$$

4 评分函数[3]

聚类划分就是把数据集划分成 k 个不相交的子集,使每一个子集中的点尽可能同质. 数据集 $D = \{x(1), \cdots, x(n)\}$ 被划分为 N 个区间,第 i 个区间束流寿命为 τ_i,数据个数为 N_i,寻找 K 个聚类 $C = \{C_1, \cdots, C_k\}$,使每一个数据点被分配到一个惟一的聚类 C_k. 定义空间数据点的聚类中心

$$\tau_k = \frac{1}{n_k} \sum_{x \in C_k} (N_x \times \tau_x), \quad (4)$$

其中 n_k 是第 k 个聚类中的点数,在聚类 C_k 中束流寿命为 τ_x 的区间数据点个数是 N_x.

同质性是用评分函数实现的,评分函数的属性对从数据中发现的聚类的类型有非常大的影响,不同的评分函数可能对应不同的聚类结构. 定义

$$wc(C) = \sum_{k=1}^{K} \sum_{x \in C_k} \frac{N_x}{2} (\tau_x - \tau_k)^2, \quad (5)$$

$wc(C)$保证聚类是紧凑的. 评分函数 $S = wc(C)$,计算 $wc(C)$需要遍历数据一次,其复杂度 $O(n)$.

5 评分函数搜索背景

划分聚类是最小化评分函数的过程,是计算复杂度很高的问题. 只要对把各个点分配到聚类 C 的可能分配方案所组成的空间进行搜索就可能了,搜索的目标是使评分函数最小化(或最大化,视选取的评分函数定). 可以认为这种搜索问题本质上是组合优化的一种形式,因为我们是对把 n 个对象放入 K 个类的分配方案进行搜索的最大化(或最小化)选取的评分函数. 可能分配方案(聚类数据的不同方法)的数量可以近似为 K^n. 直接的穷举搜索肯定是不可行的,除非要处理的数据集微乎其微. 通常不存在直接的方法找到最小化评分函数的特定聚类 C.

有各种算法来搜索最优的(或至少是好的)划分,基于局部搜索的递归改善算法在聚类分析中特别流行. 其一般思想是:从随机选取的聚类开始;然后从新分配点使评分函数最大程度的增长(或降低);然后再重新计算更新后的聚类的中心;再次重新分配点,如此继续直到评分函数没有变化或聚类成员没有变化. 著名的 K 均值算法就是运用这一原理的典型算法. 通常,这个算法以局部最优结束,搜索复杂度 $O(KnI)$,其中 I 是迭代次数.

6 寻找聚类边界的搜索算法

K 均值算法无法知道收敛到的聚类与最佳的可能聚类相比的好坏程度,难以判断局部最小值相对于全局最小值的质量,如图3,实际中还必须检查并区分最小值、最大值和鞍点. 而穷举算法复杂度 K^n,难以实现. 本文提出一种着眼于寻找聚类边界的算法,来搜索最佳可能聚类. 聚类是用边界来划分的,假定边界可以通过连接集合中的某些有代表性的点用连续的欧氏空间的面来表示. 对于 k 个聚

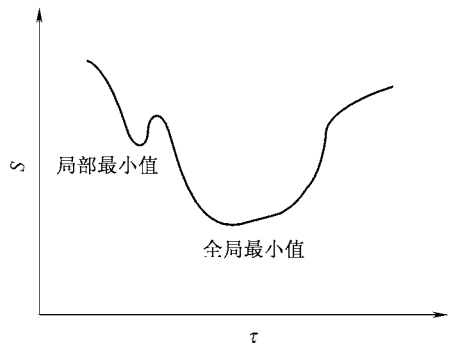

图 3　评分函数的一个例子

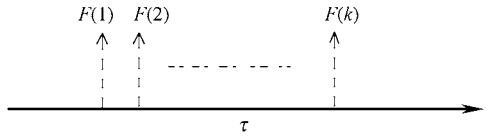

图 4　数据空间被 k 个边界划分

类,这样的界面数目是($k-1$),算法遍历所有点寻找最佳聚类的空间边界.

在数据挖掘实践中我们要搜索的模型有不少一部分是一个纬度的,一纬的情况是相当值得重视的.同时本文的纬的大小为 1,以一纬为例,算法如下:

1)首先将数据在数据空间排序. 用一分为二的方法,将数据空间用($k-1$)个边界 F_k 划分,代表 n 个数据中的某($k-1$)个数据. 如图 4;

2)改变 F_1,用数据空间下一个数据代替当前值;

3)改变 F_2,用数据空间下一个数据代替当前值;

……

4)改变 F_k,用数据空间下一个数据代替当前值;

5)计算当前划分的各个簇的中心和划分的评分函数,并与前一次比较,保留评分函数较小的划分;

6)直到 F_k 位于数据空间边界;

……

7)直到 F_2 位于数据空间边界;

8)直到 F_1 位于数据空间边界;

9)输出划分边界和聚类中心.

算法中每个聚类边界 F_i 遍历数据集 1 次,步骤 5)计算聚类中心和评分函数的复杂度是 $O(n)$,故算法复杂度为 n^k.

类比牛顿力学质点系统,可降低复杂度. 给定数据空间中束流寿命 τ 窗格中有 N_τ 个数据. 数据集类比质点系统,每个数据点对应质量为单位 1,速度大小为 τ 的质点. 数据空间被划分为 k 个簇,对应于 k 个子质点系统. 利用柯尼希定理,对于步骤 5):

$$S = wc(C) = \sum_{k=1}^{K} E_{\text{Relavivek}} = E - \sum_{k=1}^{K} E_{\text{Centerk}} = E - \sum_{k=1}^{K} \frac{1}{2 \times \sum_{x \in C_k} N_x} \left(\sum_{x \in C_k} N_x \times \tau_x \right)^2 = E - \sum_{k=1}^{K-1} \frac{p_k^2}{2 \times m_k} - \frac{\left(p - \sum_{k=1}^{K-1} p_k \right)^2}{2 \times \left(m - \sum_{k=1}^{K-1} m_k \right)}, \tag{6}$$

式中评分函数 S,系统总能量 E,第 k 个子系统质量 m_k,动量 p_k,在其质心坐标系中能量 $E_{\text{Relativek}}$,子系统质心能量 E_{Centerk}. 改进的算法是在搜索开始的时候遍历整个数据集,计算系统总质量 m、系统总动量 p 和系统总能量 E,其复杂度 $O(n)$. 在步骤 5)中计算各子系统质量、子系统动量和各子系统在其质心坐标系中能量之和即评分函数. 则在第 5)步不必遍历所有数据,根据上一次划分的计算结果,利用式(6)就可推算出当前划分下各子系统质量、子系统动量和各子系统在其质心坐标系中能量之和即评分函数. 算法复杂度为 $n^{K-1}+n$.

实践中分类数 k 大于 2 的情形似乎更具普遍性,遗憾的是该算法复杂度随 k 指数增长,在分类数较大的时候搜索似乎有些困难. 当 k 较小,不失为一种好方法,较 K 均值算法能找到全局最优解.

7　结果讨论

应用以上算法,对 HLS 束流状态描述建模. 数据库 Sql Server 2000 运行环境为 hp ProLiant DL360 generation 3,Windows 2000 Server. 数据库中记录的个数为 6815478,抽取从 2004/01/01 到 2004/02/29 运行数据,共有 597619 条. 期间由于机器调试和节假日休息,实际运行 42 天,运行次数 129. 束流流强区间数 319,束流寿命离散化区间最大数 144,离散化处理后有 15086 条. 数据挖掘程序运行环境为 Sun Enerprise 250,Solaris 8. 输入分类数 $K=2$,评分函数搜索程序运行总时间 133s,聚类分析结果如图 5 和图 6.

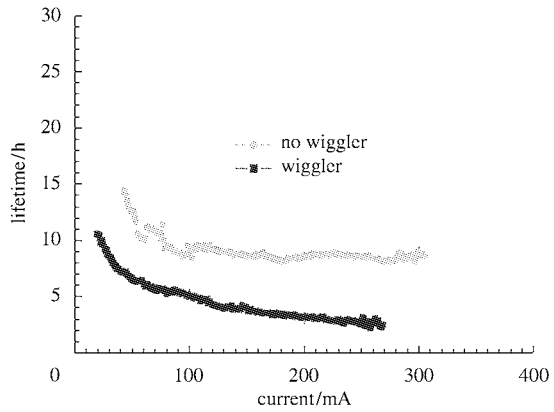

图5 束流状态聚类中心曲线

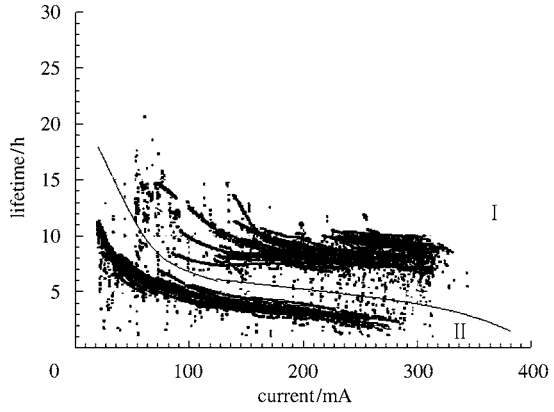

图6 束流状态聚类划分

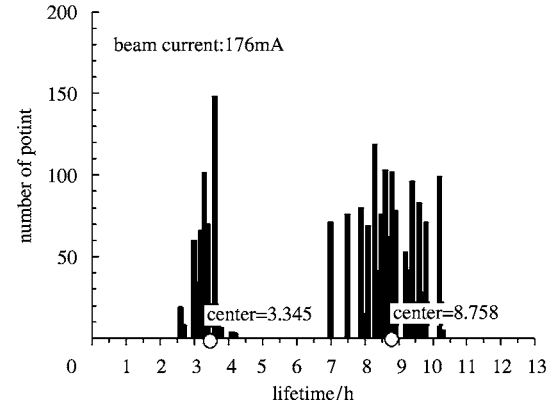

图7 数据密度与束流寿命关系

分析指出,从2004/01/01和2004/02/29电子储存环运行于通用光源模式,束流运行状态受超导扭摆磁铁影响,聚类中心曲线是两种束流状态最可能运行路径,对机器研究和故障诊断提供了依据.

以束流流强176mA为例:图7是数据点在各束流寿命区间分布直方图. 图8是搜索过程中评分函数随聚类划分边界变化曲线,聚类划分边界沿束流寿命从低到高遍历数据集. 束流寿命边界在4.2h处评分函数有最小值,对应聚类中心分别为:3.345h和8.758h,如图7.

图9显示在束流流强区间[256mA,257mA),搜索到评分函数的有两个极值点,分别对应聚类划分

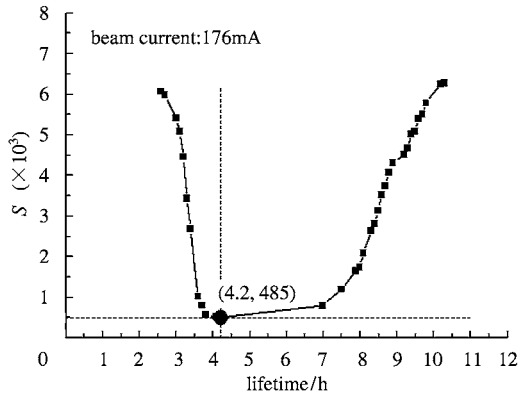

图8 评分函数与束流寿命关系

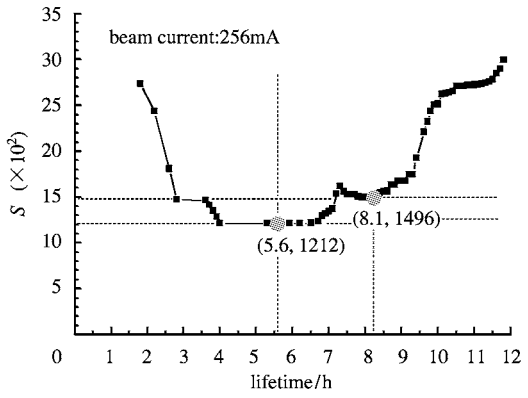

图9 束流寿命-评分函数曲线

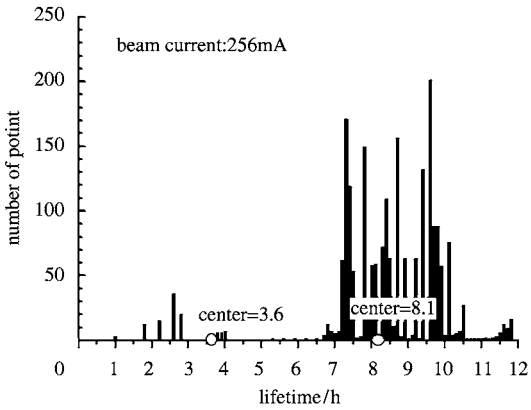

图10 束流流强256mA的束流状态聚合中心

边界 5.6h 和 8.1h,选取评分函数最小的点用于全局最优划分. 对应的聚类中心如图 10.

8 结论

本文建立了着眼于寻找聚类边界的数据挖掘聚类方法,并类比质点系统,利用柯尼希定理优化搜索算法,能找到全局最优划分,且复杂度较低. 试验表明该方法的可行性和有效性. 应用数据挖掘的束流状态描述建模是机器研究的依据,该方法具有较强的借鉴意义,可用于类似数据的聚类描述建模.

参考文献(References)

1 HAN Jia-Wei, Micheline Kambr. Data Mining: Concepts and Techniques. Beijing: China Machine Press, 2001. 70—94(in Chinese)
(韩家炜, Micheline Kambr. 数据挖掘: 概念与技术. 北京: 机械工业出版社. 2001. 70—94)

2 JIN Yu-Ming. Electronic Storage Ring Physics. Hefei: University of Science and Technology of China Press, 2001. 85—102(in Chinese)
(金玉明. 电子储存环物理. 合肥: 中国科学技术大学出版社. 2001: 85—102)

3 HAN Jia-Wei, Heikki Mannila, Padhraic Smyth. Principles of Data Mining. Beijing: China Machine Press, 2003: 180—300(in Chinese)
(韩家炜, Heikki Mannila, Padhraic Smyth. 数据挖掘原理. 北京: 机械工业出版社. 2003: 135—206)

Application of Data Mining in Beam Status Descriptive Modeling*

XIE Dong[1)]　LI Wei-Min　XUAN Ke　HE Duo-Hui

(National Synchrotron Radiation Laboratory, University of Science & Technology of China, Hefei 230029, China)

Abstract　Cluster analysis is an effective method for describing colony character. Hower, as the size of mined databases increased, it is a complex problem in searching for the extremum of score function. In this paper, we propose a novel technique and apply it in beam current descriptive modeling. Our experiments show that the method is quick and efficient. As reproducibility is an important specific of a storage ring, the clustering model is helpful for machine study and decision-making.

Key words　data mining, descriptive modelling, clustering, search method, beam status

Received 14 July 2004

* Supported by NSRL Phase Ⅱ Project

1) E-mail: xiedong@mail.ustc.edu.cn