第40卷第4期 原 子 能 科 学 技 术 Vol. 40, No. 4
2006年7月 Atomic Energy Science and Technology July 2006

应用数据挖掘的束流流强预测

谢 东,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,安徽 合肥 230029)

摘要:文章简要介绍数据挖掘技术。在详细分析束流流强曲线特点和束流流强曲线与历史曲线的相似性基础上,提出了1种基于数据挖掘中时间序列相似性研究的束流流强预测方法,并将其运用到合肥光源运行数据分析中。试验结果表明,该方法是有效的,能够满足运行要求。

关键词:束流流强;数据挖掘;时间序列;预测
中图分类号:TL503.6 **文献标识码**:A **文章编号**:1000-6931(2006)04-0465-05

Application of Data Mining in Beam Current Forecast

XIE Dong, LI Wei-min, XUAN Ke, HE Duo-hui
(*National Synchrotron Radiation Laboratory, University of Science and Technology of China, Hefei* 230029, *China*)

Abstract: Data mining technique is briefly introduced in the paper. The comparability of history beam current curves was analyzed first, then a method to forecast the beam current was put forward based on time sequence comparability study, and used in Hefei light source operational data analysis. The result indicates it's useful.

Key words: beam current; data mining; time sequence; forecast

为用户提供高品质服务是同步辐射光源的主要任务。束流流强是电子储存环的主要参数之一。将当前束流流强序列与历次观测序列进行对比,能够用以束流流强预报,从而提供决策支持。由于出光口的光通量与时间的变化规律与束流流强变化规律一致,实验站用户可利用该预报更好地安排试验进度。

本文研究的束流流强预测的中心思想是:采用一种方法,规范化束流流强序列。以储存环运行可重复性为依据,利用本次已经测量到的束流流强序列 Q,在一定范围内的历史序列 L_i 中进行相似性查寻,找出与该序列最相似的几个历史序列,将这几个序列所对应的流强加权平均后即得到预测流强。

1 数据挖掘技术[1,2]

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,导致出现了"数据爆炸但知识贫乏"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

收稿日期:2004-10-08;修回日期:2005-04-25
基金项目:国家计委重大科研项目"国家同步辐射实验室二期工程"资助项目
作者简介:谢 东(1976—),男,重庆人,博士研究生,加速器控制专业

有用的信息和知识,数据挖掘技术应允而生。

数据挖掘(DM ,Data Mining)又称为数据开采,即从海量数据中提取隐含在其中、人们事先未知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模式的高级过程。数据挖掘产生于 20 世纪 80 年代末期,是目前国内外的一研究热点。数据挖掘是数据库知识发现(KDD,Knowledge Discovery in Database)的核心技术,它的显著特点是强大的数据处理能力,能从大量数据中发现有用规律、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。

2 时间序列基本概念[3,4]

束流流强数据库属于时序数据库。时序数据库系指由离散时间序列组成的数据库。离散时间序列简称为时间序列,通常是指按等时间间隔顺序排列的测量值集合。时间序列是一维的,它的一系列观测值是按照时间间隔进行的。时间序列数据的应用非常广泛,对应的领域也五花八门,诸如经济、生物医学、生态学、大气、海洋科学、空间物理和核科学。

时间序列相似性搜索(又称为相似性查寻)是在时间序列数据库中发现与给定序列模式相似的序列或查找库中相似的时间序列对。

定义 1:给定阈值 $\varepsilon \geqslant 0$ 和序列间距离公式 D ,如果对于时间序列 X 和时间序列 Y,有 $D(X,Y) \leqslant \varepsilon$,则称为时间序列 X 和 Y 在阈值 ε 内相似, 简称时间序列 X 和 Y 相似。

对时间序列的相似分析,当前有许多定义时间序列间距离的公式,通常采用欧几里得距离作为相似计算的依据。

定义 2:等长序列 X 和 Y 间的欧几里得距离为:

$$D_0(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

其中:n 为序列长度,x_i、y_i 为时间序列 X、Y 的第 i 个值。

3 数据预处理

数据挖掘虽以数据为驱动、根据数据来发现知识,但实际的物理模型能够指导挖掘过程。结合储存环物理模型,采用一些适当的数据预处理技术,有助于进行相似性查寻。数据预处理主要包括 4 个过程:数据规范化处理;数据清洗;数据转换;数据消减。数据规范化处理是寻找数据特征,将序列进行适当变化,这些特征将直接指导数据挖掘任务(在此是预测);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,即将数据限定在特定范围;数据消减是采用分段平均值技术降维,缩小数据的规模。

3.1 规范化处理[5,6]

合肥电子储存环有两种运行模式:普通光源模式和高亮度模式。普通光源模式又可按超导 wiggler 是否充电分为两种情况。于是产生 3 种机器状态,分别对应形状大不相同的束流流强曲线。另外,束流流强变化还由初始流强决定。所以,流强时间序列无法直接用于流强预测,需进行规范化处理,寻找流强时间序列的特征,以区分以上各种不同的机器状态。

通过对大量束流流强数据分析研究可获得束流变化的如下 2 个显著特点:1) 束流以运行轮次为周期表现出明显的周期性,各次运行的流强变化规律是相似的;2) 每次运行的流强与时间有强相关性,表现为特定形状的图像。采用的规范化处理方法是:将流强 I 取自然对数 $\ln I$,将 I_i 的时间序列变换为 $\ln I_i$ 的时间序列。根据电子储存环物理理论可知,束流近似地按指数形式衰减。对于恒定的储存环物理参数,在一定范围内,流强对数 $\ln I$ 大致呈线性形式。实际运行数据已证实了这一特点。

应用束流流强预测的数据挖掘不要求序列在时间轴上完全一致。若历史时间序列初始流强不一样,但具有相同的形状,在时间上虽存在偏移,但仍可认为它们是匹配的。以当前查寻序列 Q 最后时刻为时间零点,对应的流强 I_0 为历次时间序列的基点 t_0,分别寻找历次时间序列流强最接近 I_0 的时刻 t_i,然后将历次时间序列相对平移 Δt_i($\Delta t_i = t_i - t_0$)。在新坐标系中,所有序列皆最靠近点$(0, I_0)$。

3.2 数据清洗

实际数据库中的数据一般是脏的,含噪声的。数据清洗可改进数据质量,从而有助于提高其后的挖掘过程的精度和性能。对于同一数

据,若存在两个或多个相同记录,则将影响数据分布,需进行检测并消除冗余。采用的方法是对重复数据取平均值。另一方面,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,数据可能是负的,也可能过大,这些数据均无意义。处理方法是检测并清除这些错误数据。

3.3　数据转换

本文数据挖掘的主要任务是对加 wiggler 情况下以通用光源模式运行的机器状态进行束流流强预测,数据转换经历以下步骤。

1) 限制束流寿命在 1～25 h 范围内。该状态下束流寿命不应偏离这一范围。统计分析发现:寿命低于 1 h 的数据约占 0.56%,源于束流瞬间损失或计算寿命程序产生故障。高于 25 h 的数据约占 3.7%,是由低束流运行状态或其它原因所造成。数据转换则删除这一范围之外的数据。

2) 限制流强范围为 20～40 mA。统计发现:在机器研究或实验站调试阶段,有低束流运行状态出现,其流强(<20 mA)约占 7.1%;高流强(>400 mA)则低于 0.1%。超出此范围的数据无实用价值,机器研究人员对它们不感兴趣,在此删除。

3.4　数据消减

合肥光源通用光源模式运行周期约为6 h,束流流强数据采集每隔 2 s 为 1 个点。经以上处理后的数据规模很大。每次运行的数据对应 n(n 为查寻序列 Q 的长度)维空间上的 1 个点。在此,采用一定的降维技术来缩小数据规模。

本工作采取序列分段平均值技术(PAA, Piecewise Average Approximation) 作为长序列降维方法。PAA 方法是将序列分成长度相等的区段,然后,用各区段的平均值组成新的序列来近似表示原序列。PAA 平均值序列的维数比原始时间序列维数低,具有简单、直观、高效的优点。区段平均值法将长为 n 的规则序列等分为 $N(N<n)$段,取每段内各数据点的算术平均值组成新的平均值序列,这样,即将序列 L(维数为 n)变换为变换域上的平均值序列 L'(维数为 N)。将长为 n 的序列 L 分为 N 段,为处理方便,设 n 可被 N 整除,否则,末尾序列用不等分处理,那么,n/N 表示每区段内包含的数据个数。然后,计算第 i 个($i=1,2,\cdots,N$)区段对应的平均值,也即序列 L' 的第 i 个数据点 L'_i。PAA 特征提取的时间复杂度为 $O(n)$("O"为数学符号,是算法的时间耗费,它是该算法所求解问题规模 n 的函数),即与 n 同数量级。

$$L'_i = \frac{N}{n}\sum_{j=n(i-1)/N}^{ni/N} L_j \tag{2}$$

其中:$j=1,2,\cdots,n$。

噪声是测量变量中的随机错误或偏差,它的危害很大。PAA 方法减少了数据挖掘过程所处理的值的数据量,同时平滑了数据,还消除了数据噪声。

4　相似性查寻

数据库查寻的目的是找出符合查寻的精确数据,而相似性查寻则是为找出与给定查寻序列最接近的所有数据序列。常用的时间序列相似性搜索方法是:首先将数据从时间域变换到频率域,采用信号分析的离散傅里叶变换技术(DFT,Discrete Fourier Test)或离散小波变换(DWT,Discrete Wavelet Transform),然后通过查寻前几个傅里叶系数来确定相似的序列。本工作采用的方法是,经以上一系列预处理过程后,直接在时域上寻找最匹配的时间序列。

所有的历史数据经特征提取后映射为 N 维空间上的点集,查寻序列 Q 映射为特征空间上的 1 个点,称为查寻点。查寻点和相似阈值 ε 一起确定了查寻区域,对应于 N 维空间上的 1 个球。在该区域内的点对应的时间序列即为匹配的时间序列。

5　加权平均

利用上述的相似性搜索所得到的几个序列的预测时刻的束流流强对数值,用加权平均的方法来求取预测的束流流强对数值。考虑到两个序列之间的相似程度可由它们之间的距离来表示,即距离越小,相似性就越高,在选择权重时以它所对应的序列与所预测的序列之间距离的倒数作为权值,即:

$$L_f = \frac{\sum_{i=1}^{m} L'_i / D(L_i, L_q)}{\sum_{i=1}^{m} 1/D(L_i, L_q)} \tag{3}$$

式中：L_q 表示查寻序列；L_f 表示待预测的值；L_i 表示查寻到的匹配序列；m 表示所查寻到的匹配序列数目；L_i' 表示匹配序列对应待预测时刻的值；D 表示两个序列之间的距离，其倒数是对应的权重。

最后，将预测的值取自然对数，得到预测束流流强。

6 试验

利用时间相似性搜索算法，对 2004 年 4 月束流流强数据进行处理。运行环境为 hp ProLiant DL360 generation 3，Windows 2000 Server。数据库中记录的个数为 535 559，其间，电子储存环实际运行次数为 108，合计运行时间 524 h。

以 2004 年 4 月 30 日第 3 次运行束流流强预测为例。

从数据库查寻出运行 1 h 后的束流流强为 270 mA，其自然对数值为 5.598。图 1 是对应于 2004 年 4 月 30 日第 3 次运行预处理后的时间序列，通过共同点(0，5.598)；图 2 显示预测流强和实际测得流强；图 3 是预测流强的相对偏差。为方便起见，图 3 上时间坐标以 d 为单位。从图 3 可看出，预测值与测量值间的相对偏差为 −0.2%～7.34%，绝对相对偏差的均值为 3.22%，相对偏差与预测时间长度成正比，即预测时间越长，相对偏差越大，8 h(约合 0.33 d)预测相对偏差在 10%以内。

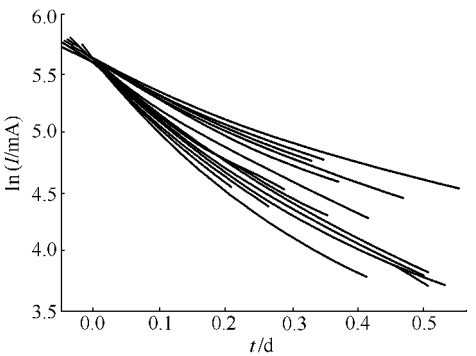

图 1 对 2004 年 4 月 30 日第 3 次运行进行预处理后的时间序列

Fig. 1 Time sequence after post-processing reference data of 3rd operation on 2004-04-30

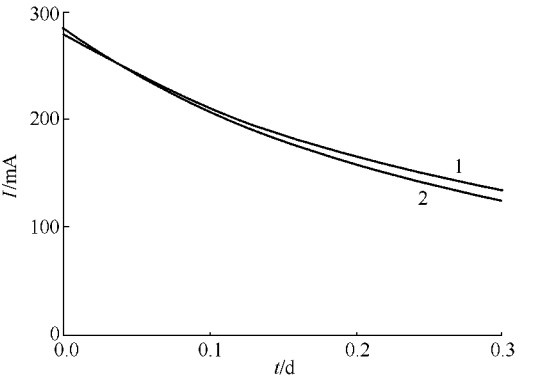

图 2 2004 年 4 月 30 日第 3 次运行束流流强预测值和实际值比较

Fig. 2 Forecasted beam current compared with real value on 2004-04-30

1——预测值；2——测量值

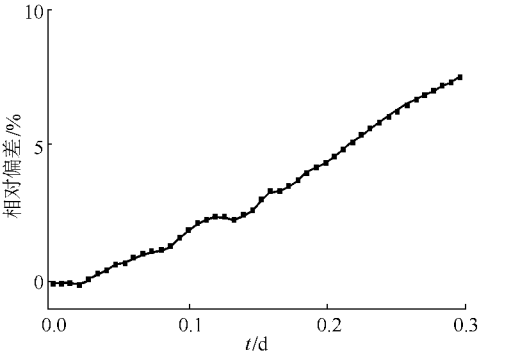

图 3 2004 年 4 月 30 日第 3 次运行束流流强预测相对偏差

Fig. 3 Relative deviation of forecasted beam current from real beam current on 2004-04-30

表 1 列出了对 4 月 28 日至 4 月 30 日的所有预测结果。在这期间，电子储存环运行 7 次，选取其中连续运行时间长度超过 1 h 的 6 次。以初始 1 h 束流流强序列为查寻序列，预测该次运行 1 h 以后的束流流强。

从表 1 可看出，预测的绝对相对偏差的均值在 0.76%～5.91%之间，预测的束流流强曲线与实际的束流流强变化规律接近。

在 2004 年 4 月 29 日第 1 次运行时间段，机器重复性好，预测的误差最小；2004 年 4 月 28 日第 2 次运行的预测误差最大，原因是查寻出的历史序列中有束流丢失。

表 1　试验结果

Table 1　Experiment results

预测序号	时间	运行次数	阈值	匹配序列数	绝对相对偏差的均值[1)]/%
1	2004-04-28	第 1 次	0.245	4	3.43
2	2004-04-28	第 2 次	0.273	4	5.91
3	2004-04-28	第 3 次	0.397	5	2.18
4	2004-04-29	第 1 次	0.118	16	0.76
5	2004-04-30	第 1 次	0.264	5	2.17
6	2004-04-30	第 2 次	0.394	4	2.81
7	2004-04-30	第 3 次	0.447	5	3.22

注:1) 绝对相对偏差的均值$=\frac{1}{N}\sum_{i=1}^{n}|(I_{f,i}-I_{m,i})/I_{m,i}|$,$N$ 为数据总数,$I_{f,i}$、$I_{m,i}$分别为束流流强第 i 个预测值和测量值

时间序列相似性搜索算法假设查寻时间序列与历史时间序列具有相似性,其物理上的依据是储存环运行的可重复性,因此,时间序列相似性搜索更具有普遍性。束流流强对时间也有极强的相关性,且表现为特定的函数形式,可用回归方法予以处理。在这次挖掘任务中,若将这两种方法结合在一起,预测效果可能会更好。此外,查寻序列不能太短。查寻序列越短,对机器状态的反映越不准确。当机器状态不佳时,将出现零点几乃至几十 mA 的束流非正常损失。分析表明,2004 年,束流损失现象普遍,平均每 1 次运行有 2 次$-1\sim-0.1$ mA/s 范围内的束流损失。这种偶然丢失束流将影响预测精度。束流流强是对时间积分类型的变量,基于时间序列相似性搜索的预测方法对变化率类型变量的预测效果更佳。

7　结论

本工作基于对束流流强曲线特点的详细分析,建立起新的坐标系规范化处理束流流强序列,提出了 1 种基于时间序列相似性搜索的束流流强预测方法,并讨论了算法具体实现中的若干问题。实例证明了该方法的有效性。

数据挖掘作为 1 个在海量数据中获取知识的有力工具,适合用于核科学研究。本工作是这方面的初步尝试。今后,将结合电子储存环的物理模型,挖掘更多更深层次的知识。

参考文献:

[1] HAN Jiawei, KAMBER Micheline. 数据挖掘:概念与技术[M]. 北京:机械工业出版社,2001:70-94.

[2] HAND David, MANNILA Heikki, SMYTH Padhraic. 数据挖掘原理[M]. 北京:机械工业出版社,2003:135-206.

[3] 袁贵川,程利,王建全. 利用数据挖掘进行短期电价预测[J]. 电力系统及其自动化学报,2003,15(2):19-23.
YUAN Guichuan, CHENG Li, WANG Jianquan. Electric price forecasting using datamining[J]. Proceedings of the EPSA, 2003,15(2):19-23(in Chinese).

[4] 蔡智,岳丽华,王熙法. 时序模式发现算法研究[J]. 计算机研究与发展,2000,37(9): 1 107-1 113.
CAI Zhi, YUE Lihua, WANG Xifa. Research on an algorithm for time E-series patterns discover[J]. J Comput Res Develop,2000,37(9): 1 107-1 113(in Chinese).

[5] 金玉明. 电子储存环物理(第 2 版)[M]. 合肥:中国科学技术大学出版社,2001:85-102.

[6] SUN Baogen, LU Ping, WANG Junhua, et al. Beam measurement system in NSRL[J]. J Syst Eng Electron, 2000,11(3):9-13.