文章编号:1671-4598(2019)09-0137-04 DOI:10.16526/j.cnki.11—4762/tp.2019.09.030 中图分类号:TP3 文献标识码:A

激光质谱中基于数据挖掘的激光输出功率预测技术研究

刘莲花[1]，杨文喜[2]，张晓卫[1]，但勇军[2]，刘 彬[2]
(1. 粒子输运与富集技术国防重点实验室，天津 300180；
2. 核工业理化工程研究院，天津 300180)

摘要：由于激光质谱系统逻辑结构复杂多样，激光输出功率是激光质谱系统进行的关键条件之一，提前掌握激光输出功率未来状态的发展趋势可为激光质谱系统运行决策提供重要依据，因此进行激光质谱系统激光输出功率的预测技术研究非常必要；采用M5预测模型、线性回归模型、向量机模型对质谱系统的激光输出功率历史数据进行了建模及预测分析，通过比较几个预测模型的预测误差及平均误差，结果表明M5预测模型的预测结果相对最优；通过对激光输出功率历史数据的分析及预测，确定了激光质谱系统激光输出功率的研究预测模型。

关键词：数值预测；预测模型；预测算法

Research on Laser Output Power Prediction Technology Based on Data Mining in Laser Mass Spectrometry

Liu Lianhua[1], Yang Wenxi[2], Zhang Xiaowei[1], Dan Yongjun[2], Liu Bin[2]
(1. Key Laboratory of Science and Technology on Particle Transport and Separation, Tianjin 300180, China;
2. Research Institute of Physical and Chemical Engineering of Nuclear Industry, Tianjin 300180, China)

Abtract: Because the logic structure of laser mass spectrometry system is complex and diverse, laser output power is one of the key conditions for laser mass spectrometry system to carry out. To grasp the development trend of laser output power in advance can provide an important basis for the operation decision of laser mass spectromery system. Therefore, it is necessary to study the prediction technology of laser output power. The system adopted the M5 prediction model, the linear regression model and the vector machine model to modele and predicte by history data. By comparing the prediction errors and the average errors of several prediction models, the prediction results of the M5 prediction model are relatively optimal. Through the analysis and prediction of the output power historical data, the research and prediction model of the output power of laser mass spectrometry system is determined.

Keywords: data prediction; prediction model; prediction algorithms

0 引言

预测是定期更新对未来数据的当前观察，以反映新的或变化中的信息过程。它是基于分析当前和历史数据来决定未来趋势的过程。预测分析是一种统计或数据挖掘解决方案，包含可在结构化和非结构化数据中使用以确定未来结果的算法和技术。可为预测、优化、预报和模拟等许多其他用途而部署，也可为规划流程提供各种信息，并对未来提供关键洞察[1]。

数据挖掘主要应用于描述类及预测类工作。其适用的关键在于两个方面：一方面数据之间确实存在一定关系；另一个方面需要大量数据。通过定性分析，已经确定了参数的关系确实存在，通过数据库技术，为系统参数积累了大量的真实历史数据，因此开展数据挖掘技术研究条件基本满足[2]。

收稿日期:2019-02-26； **修回日期**:2019-03-26。

作者简介:刘莲花(1978-)，女，山东德州人，硕士研究生，高级工程师，主要从事激光控制、智能化科研工作方向的研究。

由于激光质谱系统逻辑结构复杂多样，激光质谱系统包括激光系统、质谱装置、质谱信号测量装置、温湿度仪表、压力仪表等设备，激光质谱系统运行状态受到激光系统、质谱装置、质谱测量装置、激光器特性参数以及环境参数的影响，同时激光质谱系统内的多种设备在运行期间相互影响。激光输出功率对激光质谱系统的运行状态影响比较大，对激光输出功率的合理预测，做到提前掌握激光系统未来状态的发展趋势，为激光质谱系统运行决策提供重要依据，因此，进行激光输出功率预测技术研究对整个激光质谱系统具有很重要的意义。

由于激光系统的物理过程相对较为复杂，目前，还未建立完整的物理仿真模型，因此激光输出功率与其他参数关系的描述还没有。因此采用数据挖掘方法模拟关系模型，体现所有可能的影响因素，进而实现对激光输出功率的准确预测。

1 激光质谱系统结构及激光输出功率预测原理

激光质谱系统由激光系统、质谱装置、质谱信号测量装置及辅助供水系统等设备组成。激光质谱系统具有复杂

的物理逻辑关系和工艺结构，并且相互关联，相互影响，任何环节的变化都会影响质谱系统的运行状态。激光质谱系统的运行状态由多个参数表征，包括激光系统特性参数、质谱信号参数及环境参数等，任何一个参数出现异常都会标志着整个系统状态出现异常，而激光系统是激光质谱系统运行的前提条件，因此对激光系统运行状态的提前掌握对质谱系统运行具有重要意义。激光系统参数包括激光功率、脉冲延时、光束质量、光斑大小和形状等参数，而激光功率是激光系统运行状态的关键参数，因此，试验期间，需要实时预测激光输出功率的未来发展趋势，发现可能影响激光系统运行状态的因素，提前解决潜在问题，为质谱系统的稳定运行提供保障。

激光输出功率预测结构如图 1 所示，激光数据采集系统实时采集激光输出功率、脉冲延时、光斑等实时监测数据，将实时数据存储到历史数据库中，同时将实时数据发送给预测模块。预测系统读取历史数据库中历史数据建立预测模型，根据实时数据对输出功率进行实时预测。

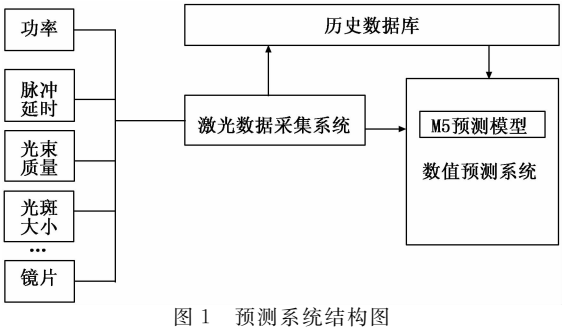

图 1 预测系统结构图

2 数据处理

数据样本是数据预测模型及相关技术的研究的关键因素，也是模型选定和验证的根源，因此数据格式及数据的正确性具有确定性作用。

2.1 时间对齐

由于激光数据采集系统需要实时采集不同独立运行的多个设备的多个数据，数据采集及存储的时间不是完全的相同，因此需要对历史数据进行相应的处理才能作为数据学习样本。将一分钟均分为 12 份，即 5 秒钟为一个时间段，一个时间段内的所有数据进行平均后进行保存，如果在该时间段内没有数值则以上一个时间段内的数值作为该时间段内的数值进行保存，从而实现数据的时间一致性。

2.2 异常数据判断

激光系统在运行过程中，功率输出会受到多种因素的影响，会出现异常数据，同时会自行恢复正常状态，这样的异常数据无法预测，因此在预测过程中需要将异常数据进行剔除。采用两种方式进行数据的优化处理，分别为正态分布的 3σ 原则和参考在网络传输信号中的通信延时的计算方式 RTT。

3σ 原则以数据符合正态分布为参考，每次当有新的捕获数据值时，通过已有数据计算得到的均值 μ 和方差 σ 得到 3σ 的范围，基本涵盖 99.74%的数据分布，当超出 3σ 范围的数据值，则按异常值进行处理。

RTT 计算方式中每一次捕获到的数据值为 RTT，$SRTT$ 是用于计算 RTO 的部分的参数值（性质上类似于均值），$DevRTT$ 同样是用于计算 RTO 部分的均值（类似于方差），最后计算 RTO。

$$RTO = \mu * SRTT + / - \delta * DevRTT \tag{1}$$

通过（1）式计算 RTO 得到一个符合条件的数据范围，再通过比对新捕获的数据值与已有的数据范围之间的关系判断捕获的数据是否为异常值。

3 数据预测建模

预测型数据挖掘大体可分为分类和回归，回归一般包括线性回归和非线性回归，许多非线性回归都可以经过适当的变化转化为线性回归。

采用激光系统历史数据作为样本数据，对几种预测模型进行测试研究，从而确定所选取的模型。

3.1 M5 模型树算法

M5 模型树算法是一种回归树算法。它结合了传统的决策树的理念，并且有一定的概率在叶子结点处生成线性回归函数。模型树的生成和决策树的生成是十分地类似。

M5 模型树算法即为输入空间 X_1、X_2 被分到各个区域上，独立的回归模型能分别产生于这些区域中。在生成模型树时，一个特征首先被放置在根节点，并为每一个可能的数值生成一个树枝；然后根结点的样本集被划分为几个子集，每一个树枝下有一个子集。这个过程被不断重复，直至某一个结点下的所有样本拥有相同的分类时，那一个部分的生成过程方才停止。这个被选择来划分特定的样本集的特征，是通过叫做“划分准则”的统计学特性来决定的。对于普通的决策树来说，划分准则是要尽可能地减少产生的子集中的熵值，即尽可能多地把同一类的样本划分在一个子集中。而 M5 模型树是一个数值预测算法，它的划分准则是基于某一个结点下的所有数值的标准差来决定的。这个标准差被用作该结点的误差度量，而能够减少最多误差值的特征就被选择为该结点的划分。划分过程在某一结点的数值标准差很小时停止，或者在某一个子集中只剩下很少的样本时停止[3]。线性回归模型于划分停止后在每个终止结点上生成。

根据 M5 模型的算法原理，采用激光系统输出功率作为预测目标，根据影响激光输出功率的影响因素生成的模型树如图 2 所示。

采用激光系统某一段时间的历史数据作为样本数据在 M5 算法模型上进行了测试，激光输出功率的历史预测结果与历史真实值对比结果如图 3 所示，此段时间内，历史真实值与历史预测值偏差不大。

3.2 多层感知机模型

多层感知机由多层神经元组成。输入的信号被提交到隐藏层的神经元中。在使用多个隐藏层时，每一层的输出都被作为输入提交到下一层神经元中。按照标准的回归模型，每一个神经元使用一个非线性激励函数：

$$Y_t^* = \phi\left(\sum_{j=1}^{h}\omega_j X_{it} + \omega_0\right) \tag{2}$$

式（2）中，Y_t^* 是这个神经元的输出（即预测的数

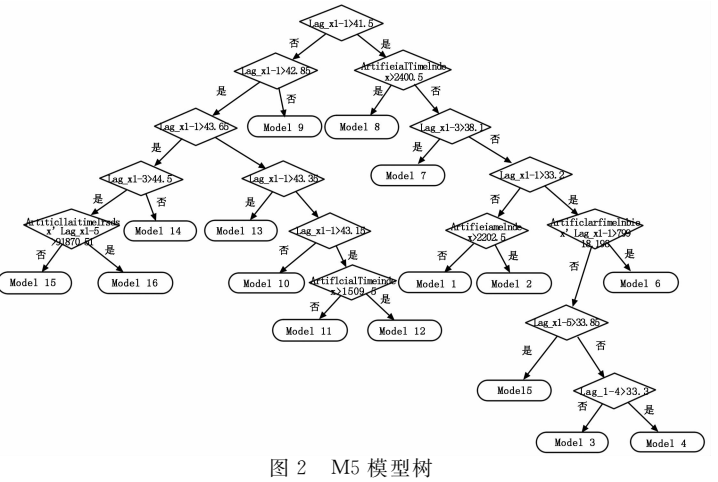

图 2　M5 模型树

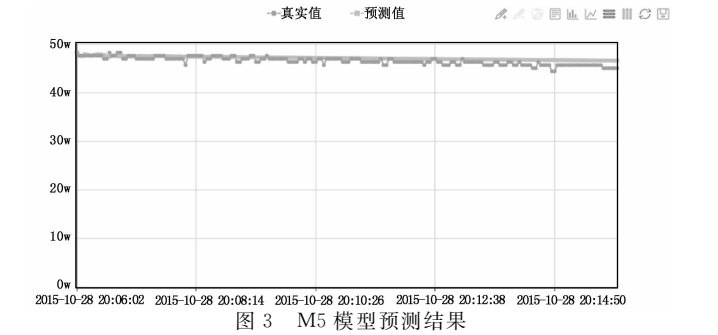

图 3　M5 模型预测结果

值），X_u是神经元的输入（被用来预测的数值），ω_j 是神经元的参数（权重）。

多层感知机模型的理念是：激励函数概略地体现了一个真实神经元的激励作用。最高层隐藏层神经元的输出被提交到输出层。只有一层隐藏层神经元配合以下函数使用[4]：

$$Y_t^* = \varnothing_{output}\left(\sum_{j=1}^{h} v_j \phi_{hidden}\left(\sum_{l=1}^{k}\sum_{i=1}^{p} \omega_{p(l-1)+i} X_{u-l} + \omega_0\right)\right) \quad (3)$$

式（3）中，Y_t^* 是第 t 个数值的预测结果，X_i 是第 i 个样本值，K 是“历史长度”，p 是输入向量的维度，ω_i 是多层感知机的隐藏层的系数，ν_j 是多层感知机输出层的系数，N 是输入值的个数，h 是隐藏层的神经元的个数，Φ（x）是激励函数[5]。在输出层：

$$\varnothing_{output}(x) = c\left(\frac{1}{1+e^{-x}} - 0.5\right) \quad (4)$$

式（4）中，c 是放缩参数使得多层感知机的输出拟合数据。

在隐藏层使用传统的 sigmoid 函数：

$$\varnothing_{hidden}(x) = \frac{1}{1+e^{-x}} \quad (5)$$

而感知机的权重则通常由梯度下降最小化算法来得出。

采用激光系统历史数据作为样本数据，预测某一特征值 X_1 按照如下过程：输入的特征值 X_1 生成一些滞后特征，再使用这些特征构造多层感知机模型，如图 4 所示。

同样采用激光系统某一段时间的历史数据作为样本数据在多层感知机算法模型上进行了测试，激光输出功率的历史预测结果与历史真实值对比结果如图 5 所示，此段时间内，历史真实值与历史预测值偏差比较大。

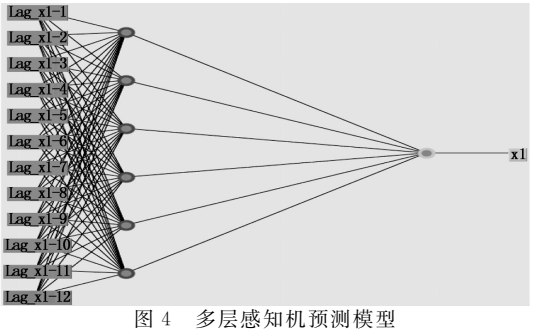

图 4　多层感知机预测模型

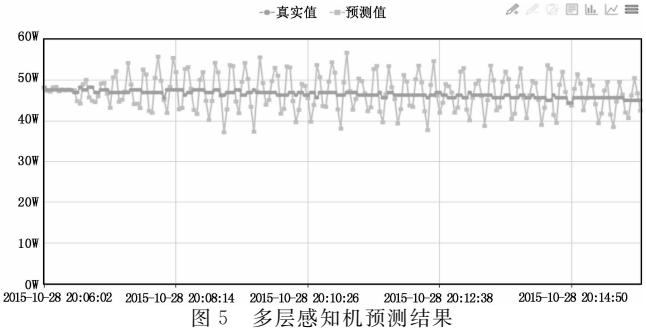

图 5　多层感知机预测结果

3.3　线性回归

在统计学中，线性回归是利用称为线性回归方程的最小二乘函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。在线性回归中，数据使用线性预测函数来建模，并且未知的模型参数也是通过数据来估计[6]。

给定一个随机样本（Y_i，X_{i1}，X_{i2}，…，X_{ip}），$i=1$，…，n，一个线性回归模型假设回归子 Y_i 和回归量 X_{i1}，X_{i2}，…，X_{ip}之间的关系是除了 X 的影响外，还有其他的变数存在。加入一个误差项 ε_i（也是一个随机变量）来捕获除了 X_{i1}，X_{i2}，…，X_{ip}之外任何对 Y_i 的影响[7]。所以一个多变量线性回归模型表示为以下的形式：

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, i = 1, \cdots, n \quad (6)$$

采用激光系统样本数据，预测某一特征值 X_1 按照如下过程：输入的特征值 X_1 生成一些人造的时间戳，并生成一些滞后特征，根据数据生成线性回归模型。

采用激光系统同一段时间的历史数据作为样本数据在线性回归算法模型上进行了测试，激光输出功率的历史预测结果与历史真实值对比结果如图 6 所示，此段时间内，历史真实值与历史预测值偏差非常小。

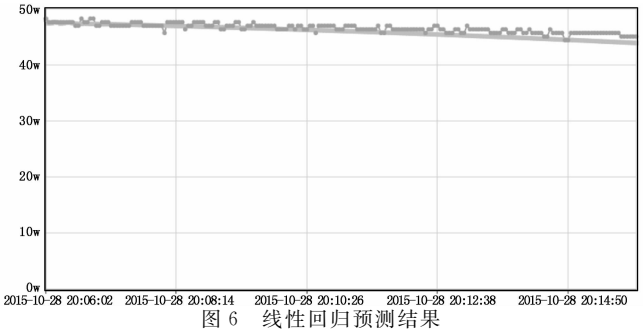

图 6　线性回归预测结果

3.4 支持向量机

支持向量机使用线性模型，通过一些非线性映射输入向量 x 到高纬度特征空间，从而生成非线性分类边界。一个在此新空间生成的线性模型可以代表一个原空间的非线性决策边界。在新的空间里，一个最优的分隔超平面被建立。这一最大间隔超平面给出了决策集之间的最大间隔。靠这个最大间隔超平面最近的训练样本被称为支持向量。所有其他的训练样本都和决定这个二元分类边界无关[8]。

在线性可分的数据中，一个拥有 3 个特征的分隔二元决策集的超平面可以由以下方程表示：

$$y = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 \tag{7}$$

在式（7）中，y 是输出，xi 是特征值，而且有四个需要算法学习的权重 ω_i。这些权重 ω_i 就是决定超平面的参数[9]。这个最大间隔超平面可以被支持向量由以下方程表示：

$$y = b + \sum \alpha_i y_i x(i) \cdot x \tag{8}$$

在式（8）中，y_i 表示训练样本 x（i）所属的类。向量 x 表示一个测试样本，而向量 x（i）是支持向量。在这个方程中，b 和 α_i 是决定超平面的参数。

在线性不可分的数据中，一个高纬度版本的方程简单地如下表示：

$$y = b + \sum \alpha_i y_i K(x(i), x) \tag{9}$$

在式（9）中，函数 K（x（i），x）被定义为核函数。常见的核函数有多项式核函数等。

采用激光系统样本数据，预测某一特征值 X_1 按照如下过程：输入的特征值 X_1 生成一些滞后特征，经计算后获得支持向量机模型及权重值。

通过支持向量机模型，计算激光输出功率的历史数据预测值。

4 结果分析

根据激光质谱实验实际情况，在不同季节以及一天的不同时段实验结果略有不同，因此采用激光质谱系统多次实验的激光系统功率及相关历史数据进行了历史预测测试。通过选择三次试验的 7 个不同时间段数据，选择每个时间段为 30 分钟，进行预测 10 分钟内的数据，将预测的历史数据与真实历史数据进行比较，并计算平均误差。通过对已经建立的支持向量机预测模型、线性回归模型、M5 模型和多层感知机模型分别进行多个时间段数据的训练、预测和平均误差计算，结果如表 1 所示。在 7 个时间段内，支持向量机模型和多层感知机模型给出的预测结果的平均误差都大于线性回归模型和 M5 模型。根据激光系统的功率数据特性，预测误差应小于 1。M5 模型在其中的 4 个时间段内的平均误差小于 1，其中 3 个时间段的误差比较大。经过与激光系统的实际运行状态进行了分析与对比，其中 3 个误差比较大的时间段为系统调节或故障阶段，数据波动较大，预测偏差较大，因此 M5 预测模型的预测结果更接近激光系统输出功率的历史数据。

同时又对 4 种预测模型的预测误差的平均值和方差进行了计算，结果如图 7 所示。其中支持向量机模型的误差的平均值和方差为 40.59 和 92.02，远远超过了线性回归模型、M5 模型和多层感知机模型，线性回归模型和 M5 模型的误差平均值和方差相差不多，与误差结果基本一致。由于 M5 模型在预测精度和稳定性上都为最优选择，因此选择 M5 模型作为激光输出功率的研究预测模型。

表 1 预测模型误差比较

	8:50—9:20	14:32—15:02	10:30—11:00	16:00—16:30	18:41—19:11	20:23—20:53	21:36—22:06
支持向量机	6.503	5.284	6.668	249.157	11.689	1.059	3.824
线性回归	1.129	5.005	0.431	16.317	7.903	1.236	3.918
M5	0.406	4.601	0.431	14.316	6.638	0.487	0.569
多层感知机	4.85	9.172	24.475	17.115	12.365	3.709	9.527

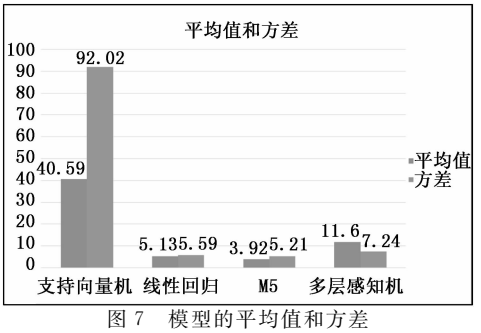

图 7 模型的平均值和方差

5 结论

根据激光系统的部分历史数据分别对 M5 预测模型、线性回归模型、向量机模型进行了建模及预测分析，通过比较几个预测模型在不同时段间的预测误差、平均误差及方差结果，M5 预测模型的预测结果相对最优。模型分析结果表明，M5 预测模型适合进行质谱系统激光输出功率的预测技术研究。

参考文献：

[1] 张君枫. 数据挖掘算法综述 [J]. 电脑学报，2010 (4)：120-121.

[2] 王光宏，蒋 平，等. 数据挖掘综述 [J]. 同济大学学报，2004，32 (4)：246-251.

[3] 刘克准，廖志芳. 数据挖掘中聚类算法综述 [J]. 福建电脑，2008 (8)：5-6.

[4] 王伦文，冯彦卿，张 铃. 动态数据挖掘的构造性学习方法综述 [J]. 小型微型计算机系统，2016 (9)：1953-1958.

[5] 薛 薇，陈欢歌. Clementine 数据挖掘方法及应用 [M]. 北京：电子工业出版社，2012.

[6] 张文彤，钟云飞. IBM SPSS 数据分析与挖掘实战案例精粹 [M]. 北京：清华大学出版社，2013.

[7] 张良军，陈俊德，刘名军，等. 数据挖掘实用案例分析 [M]. 北京：机械工业出版社，2013.

[8] 陈海燕，刘晨晖，孙 博. 时间序列数据挖掘的相似性度量综述 [J]. 控制与决策，2017 (1)：1-11.

[9] 薛 茹. 数据挖掘技术在油田中的应用 [J]. 微型电脑应用，2018 (5)：26-28.