

基于数据挖掘的高校图书馆图书借阅流量建模与分析

尹志强

(徐州医科大学 图书馆, 江苏 徐州 221000)

摘 要: 高校图书馆图书借阅流量受到多种因素的综合作用,具有一定的混沌变化特变性,针对当前高校图书馆图书借阅流量预测模型存在的精度低难题,提出了基于数据挖掘的高校图书馆图书借阅流量预测模型.引入混沌理论对高校图书馆图书借阅流量原始数据进行分析,建立高校图书馆图书借阅流量建模的学习样本,采用数据挖掘技术构建高校图书馆图书借阅流量预测模型,并与其它模型进行了高校图书馆图书借阅流量预测的仿真实验.结果表明,本文模型能够描述高校图书馆图书借阅流量变化趋势,高校图书馆图书借阅流量预测性能更优,可为高校图书馆管理者提供有价值的参考信息.

关键词: 高校图书馆; 图书借阅流量; 建模与分析; 数据挖掘

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2018)11-0095-05

Modeling and Analysis of Book Borrowing Flow in University Library Based on Data Mining

YIN Zhi-qiang

(The Library of Xuzhou Medical University, Xuzhou, Jiangsu 221000)

Abstract: Borrowing flow of library books in the university library is influenced by a variety of factors, and has certain chaos change specificity. In view of the low precision of the current library borrowing flow prediction model, this paper puts forward the prediction model of the borrowing flow for Library drawing books based on data mining. Chaos theory is introduced to analyze original data of library borrowing flow in university library and establish the learning sample of library borrowing flow modeling in university library, data mining technology is used to build the book borrowing flow of University Library. The simulation experiments are carried out to test book borrowing flow prediction model in university library compared with other models. The results show that the proposed model can describe the changing trend of library borrowing flow in university library, and the performance of library borrowing flow prediction in university library is better, which can provide valuable reference information for university library managers.

Key words: University Library; book lending flow; modeling and analysis; data mining

1 引言

在图书馆信息急剧膨胀的过程中,图书馆智能化管理程度要求越来越高,图书管理系统面临巨大的挑战^[1,2]. 图书借阅流量预测是图书馆系统中的重要技术,为此如何对高校图书馆图书借阅流量进行建模和分析成为各高校图书馆热切关注的问题^[3].

高校图书馆图书借阅流量模型主要分为两类:基于统计学理论的模型和基于时间序列的模型^[4-5],其中多元回归分析方法是一种典型的统计学理论模型,多元回归分析方法无法对随机变化特性进行建模,严重影响图书借阅流量的预测精度^[6]. 基于时间序列的模型根据时间先后收集高校图书馆图书借阅流量数据,然后采用灰色模型、神经网络、支持向量机等对高校图书馆图书借阅流量进行建模,对将来

某一时刻的图书借阅流量估计^[7,8]. 在实际应用中, 灰色模型只能对短时间、波动不大的高校图书馆图书借阅流量进行预测, 缺陷十分明显; 神经网络虽然可以对高校图书馆图书借阅流量的随机性变化特性进行拟合, 但其网络结构复杂、收敛速度慢, 经常得到“过拟合”的高校图书馆图书借阅流量预测结果, 预测结果不可靠; 支持向量机虽然不存在神经网络的“过拟合”的高校图书馆图书借阅流量预测缺陷, 但建模效率极低^[9-11]. 最小二乘支持向量机是近几年提出来的新型数据挖掘技术, 建模效率高, 而且继承了为支持向量机的优秀泛化能力, 为高校图书馆图书借阅流量预测提供了一种新的建模工具^[12].

针对高校图书馆图书借阅流量变化的随机性和混沌性, 提出了基于数据挖掘的高校图书馆图书借阅流量预测模型. 首先引入混沌理论对高校图书馆图书借阅流量原始数据进行分析, 建立高校图书馆图书借阅流量建模的学习样本, 然后采用数据挖掘技术构建高校图书馆图书借阅流量预测模型, 最后与其它高校图书馆图书借阅流量预测模型进行仿真对比实验, 以验证本文模型的优越性.

2 混沌分析方法和数据挖掘技术

2.1 混沌分析方法

对于含有 N 个数据点的时间序列 $x = \{x_i | i = 1, 2, \dots, N\}$, 根据混沌定理, 采用嵌入维(m) 和延迟时间(τ) 得到一个与原序列变化特点相同的多维时间序列数据, 具体如下:

$$X = \{X_i | X_i = [x_i, x_{i+\tau}, \dots, x_{i+(m-1)\tau}]^T\} \quad (1)$$

$$i = 1, 2, \dots, M$$

式中, M 为数据重构后的空间规模, 它与 m 和 τ 之间的关系为:

$$M = N - (m - 1)\tau \quad (2)$$

对式(1) 进行分析可以清楚知道, m 和 τ 决定数据重构的质量, 当前 m 和 τ 方法确定有两种理论, 一种为单独分开确定, 另一种方法联合确定, 由于联合确定通用性更高, 为此本文选择联合确定中的 $C-C$ 方法估计高校图书馆图书借阅流量的 m 和 τ 最佳值.

$C-C$ 方法的工作步骤为:

(1) 定义关联积分为:

$$C(m, N, r, t) = \frac{2}{M(M-1)} \times \sum_{1 \leq i < j \leq M} \theta(r - ||X_i - X_j||_{(\infty)}) \quad (3)$$

式中, $||X_i - X_j||_{(\infty)}$ 为无穷范数, 当 $x \leq 0$ 时,

$\theta(x) = 0$, 不然 $\theta(x) = 1$.

(2) 关联积分用于描述任意两点之间距离小于 r 的概率, 那么时间序列检验统计量为:

$$S_1(m, N, r, t) = C(m, N, r, t) - C^m(1, N, r, t) \quad (4)$$

(3) 将 $x = \{x_i | i = 1, 2, \dots, N\}$ 平均划分为 t 个序列, 具体为

$$\begin{cases} x(1) = \{x_1, x_{t+1}, \dots, x_{[N/t]-t+1}\} \\ x(2) = \{x_2, x_{t+2}, \dots, x_{[N/t]-t+2}\} \\ \dots\dots \\ x(t) = \{x_t, x_{t+t}, \dots, x_{[N/t]}\} \end{cases} \quad (5)$$

(4) 计算式(5) 的统计量得到

$$S_2(m, N, r, t) = \frac{1}{t} \sum_{s=1}^t [C_s(m, N, r, t) - C_s^m(1, N, r, t)] \quad (6)$$

(5) 最大统计量和最小统计之间的偏差为:

$$\Delta S_2(m, t) = \max\{S_2(m, r_j, t)\} - \min\{S_2(m, r_j, t)\} \quad (7)$$

(6) 当 $\Delta S_2(m, t) \sim t$ 达到第一个局部极小点时, 可以确定最优延迟时间(τ) 值.

(7) 根据 BDS(Brock Dechert Scheinkman) 理论定义

$$S_{2cor}(r) = \frac{1}{4} \sum_{m=1}^4 \Delta S_2(m, t) + \frac{1}{16} \sum_{m=1}^4 \sum_{m=1}^4 S_2(m, r, t) \quad (8)$$

(8) 搜索 $S_{1cor}(r)$ 的全局最小点, 那么可以获得最优嵌入窗 τ_w , 那么最优嵌入维数 m 如下.

$$m = \tau_w / \tau - 1 \quad (9)$$

2.2 数据挖掘技术

设训练样本集合为: $\{(x_i, y_i)\}, i = 1, 2, \dots, n$, 选择函数 φ 将训练样本映射到空间 f 中, 然后对训练样本进行回归, 具体如下.

$$f(x) = \omega^T \varphi(x_i) + b \quad (10)$$

式中, ω 为权值向量, b 为特征向量.

为了对 ω 和 b 进行求解, 构建一个带约束的优化问题, 具体为:

$$\begin{aligned} \min_{\omega, b, \xi, \xi^*} J &= \frac{1}{2} \omega^T \omega + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ s. t. & \begin{cases} y_i - \omega^T \cdot \varphi(x_i) - b \leq \varepsilon + \xi_i \\ \omega^T \cdot \varphi(x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \\ i = 1, 2, \dots, n \end{cases} \end{aligned} \quad (11)$$

式中, ξ_i, ξ_i^* 为训练误差的上限和下限; ε 表示预测误

差; C 表示对大于 ε 的惩罚度.

基于 KKT 的最优化条件,建立如下的拉格朗日函数.

$$L = \frac{1}{2} \omega^T \omega + C \sum_{i=1}^n (\xi_i + \xi_i^*) - \sum_{i=1}^n \alpha_i (\xi_i + \varepsilon - y_i + \omega x_i + b) - \sum_{i=1}^n \alpha_i^* (\xi_i^* + \varepsilon - y_i + \omega x_i + b) - \sum_{i=1}^n (\eta_i \xi_i - \eta_i^* \xi_i^*) \quad (12)$$

相应的约束条件为:

$$\begin{cases} \sum_{i=1}^n (\alpha_i - \alpha_i^*) \\ 0 \leq \alpha_i, \alpha_i^* \leq C \\ \omega = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \varphi(x_i) \end{cases} \quad (13)$$

式中, $\alpha_i, \alpha_i^*, \eta_i$ 和 η_i^* 为拉格朗日乘子.

对于非线性回归问题,支持向量机通过对式(14)的对偶问题进行求解.

$$W(\alpha_i, \alpha_i^*) = -\frac{1}{2} \sum_{i,j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(\varphi(x_i), \varphi(x_j)) + \varepsilon \sum_{i=1}^n (\alpha_i - \alpha_i^*) + \sum_{i=1}^n y_i (\alpha_i - \alpha_i^*) \quad (15)$$

由于支持向量机的学习效率低下,为了解决该难题,最小二乘支持向量机采用不同损失函数对目标函数进行优化,即:

$$\min_{\omega, b, \xi} J = \frac{1}{2} \|\omega\|^2 + \frac{1}{2} C \sum_{i=1}^n (\xi_i^2) \quad (16)$$

s. t.

$$\begin{cases} y_i - \omega \varphi(x_i) = b + \xi_i \\ i = 1, 2, \dots, n \end{cases}$$

那么其对偶问题相就为

$$\text{Max} L_D = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i y_i y_j k(x, x_i) \alpha_j + \sum_{i=1}^n \alpha_i^* \quad (17)$$

引入拉格朗日函数对式(16)进行优化,即

$$L = \frac{1}{2} \|\omega\|^2 + \gamma \sum_{i=1}^n \xi_i^2 - \sum_{i=1}^n \alpha_i (\omega \varphi(x_i) + b - \xi_i - y_i) \quad (18)$$

式中, γ 表示正则化参数.

最小二乘支持向量机的回归函数如下.

$$f(x) = \sum_{i=1}^n \alpha_i k(x, x_i) + b \quad (19)$$

本文选择 $k(x, x_i)$ 为 RBF 核函数,具体为

$$k(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \quad (20)$$

式中, σ 表示 RBF 核的宽度.

2.3 图书馆图书借阅流量预测模型的工作步骤

(1)针对具体一个高校图书馆,采集其一段时间的图书借阅流量.

(2)根据 C-C 方法估计高校图书馆图书借阅流量的最优 m 和 τ .

(3)根据最优 m 和 τ 建立高校图书馆图书借阅流量预测的学习样本.

(4)确定最小二乘支持向量机的参数 γ 和 σ 值.

(5)采用最小二乘支持向量机对高校图书馆图书借阅流量预测建模的训练样本进行学习,建立高校图书馆图书借阅流量预测模型.

图书借阅流量预测模型的工作流程如图 1 所示.

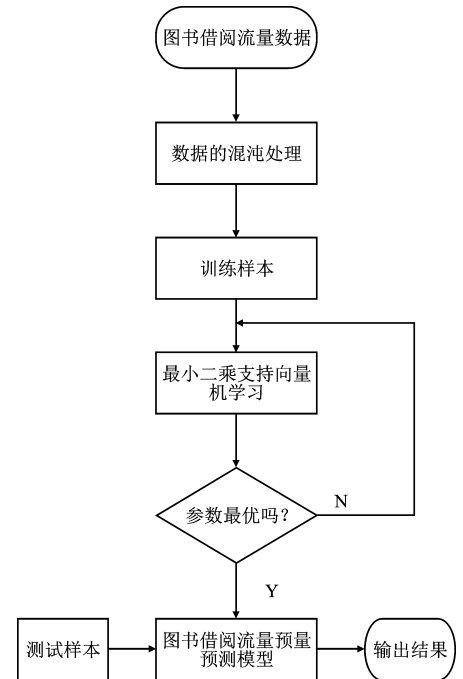


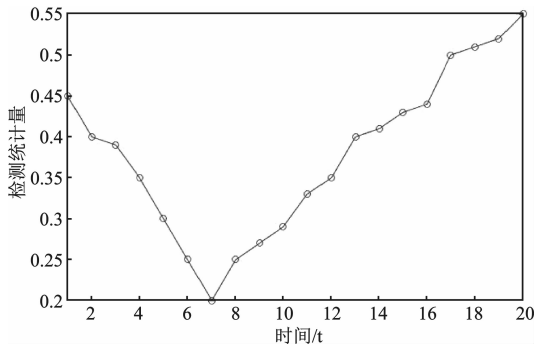
图1 图书馆图书借阅流量预测模型的工作流程

3 图书馆图书借阅流量预测模型的性能测试

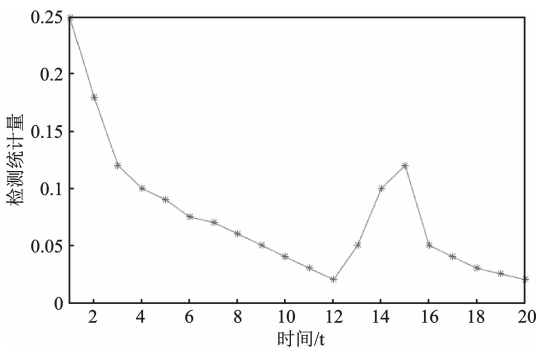
3.1 高校图书馆图书借阅流量学习样本的构建

采用 C-C 方法对高校图书馆图书借阅流量数

据进行混沌处理, ΔS_2 的变化曲线如图 2(a) 所示, 从 ΔS_2 的变化曲线可以看出, 最优 $\tau = 7, S_{2cor}$ 的变化曲线如图 2(b) 所示, 从图 2(b) 可以得到最优嵌入窗 $\tau_w = 12$, 然后根据式(9)可以得到最优 $m = 3$, 基于 $\tau = 5, m = 3$ 对图书借阅流量数据进行重构, 得到高校图书馆图书借阅流量建模的学习样本。



(a) ΔS_2 的变化曲线



(b) S_{2cor} 的变化曲线

图 2 C-C 方法确定最优的 τ 和 τ_w

3.2 结果与分析

3.2.1 本文模型的图书馆图书借阅流量预测结果

确定参数 $\gamma=100, \sigma=1.872$, 最后 50 个图书馆图书借阅流量数据点作为测试样本, 其它用于建立高校图书馆图书借阅流量预测模型, 图书馆图书借阅流量测试样本一步预测值的变化曲线如图 3 所示. 对图 3 的图书馆图书借阅流量预测结果可以看出, 本文模型可以对图书馆图书借阅流量的变化态势进行准确挖掘, 高校图书馆图书借阅流量一步预测结果很好。

图书馆图书借阅流量预测要对将来一段时间变化结果进行跟踪, 一步预测结果仅描述图书馆图书借阅流量下一时刻变化特性, 为此要进行图书馆图书借阅流量的多步建模与分析, 提前 4 步的高校图书馆图书借阅流量预测值变化如图 4 所示, 从图 4 可知, 高校图书馆图书借阅流量预测准确性有了一定的下降, 不过仍然可以反映述高校图书馆图书借

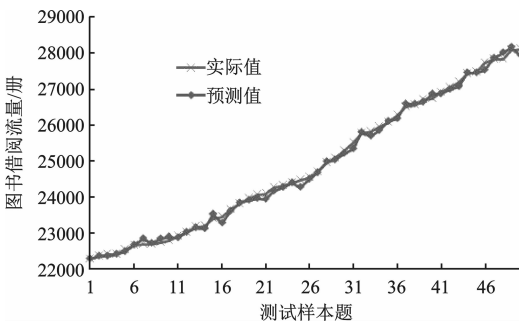


图 3 图书馆图书借阅流量的一步预测结果

阅流量的整体变化趋势, 这是由于本文根据图书馆图书借阅流量时间序列的时变性, 引入混沌理论对其进行分析, 能够更好描述高书馆图书借阅流量变化特点, 并引入最小二乘采用支持向量机校图书馆图书借阅流量的变化趋势进行拟合, 建立了高精度的图书馆图书借阅流量预测模型。

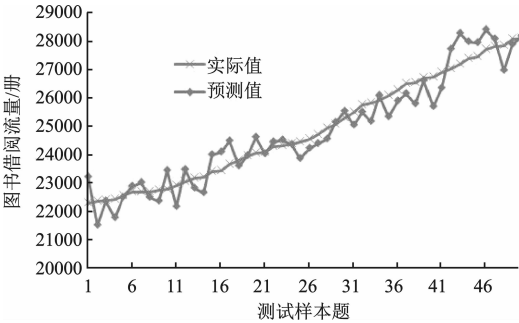


图 4 图书馆图书借阅流量的提前 4 步预测结果

3.2.2 与其它模型整性性能对比

为了使本文的图书馆图书借阅流量预测结果更具说服力, 选择灰色模型、马尔柯夫模型、BP 神经网络、支持向量机建立图书馆图书借阅流量预测的对比模型, 它们图书馆图书借阅流量预测精度如表 1 所示. 从表 1 可以知道, 相对于所有的对比模型, 本文的图书馆图书借阅流量预测精度均有了一定的改善, 解决了当前图书馆图书借阅流量预测错误大的弊端, 同时比支持向量机具有更高的高校图书馆图书借阅流量建模效率。

表 1 与其它模型的整性性能对比

模型	单步		多步	
	预测精度/%	建模时间/s	预测精度/%	建模时间/s
灰色模型	82.81	9.14	67.72	9.97
马尔柯夫模型	88.73	8.89	79.05	8.93
RBFFNN	90.07	8.02	83.00	8.16
支持向量机	96.94	29.00	90.18	35.08
本文模型	96.95	7.70	90.29	6.89

4 结束语

为了准确分析高校图书馆图书借阅流量的变化趋势,提出了基于数据挖掘的高校图书馆图书借阅流量预测模型,并进行了单步和多步预测仿真实验,结果表明,本文方法可以高精度对高校图书馆图书借阅流量变化特点进行逼近,获得了理想的高校图书馆图书借阅流量预测结果,可以应用于实际的图书馆图书管理中.

参考文献:

[1] 吕秀云. 大学图书馆自助式信息服务模式[J]. 图书馆界, 2003, 9(3):27-29, 41.

[2] 周妤. 高校图书馆用户自助服务研究. 图书馆论坛, 2006, 6(3):193- 195.

[3] 姜炳蔚,任玉杰. 回归分析在图书流通中的应用[J]. 职大学报(自然科学版),2003,15(4):6 互 7.

[4] 王家胜,牟肖光. 基于时间序列高校图书馆借阅流量分布统计分析[J]. 农业图书情报学刊,2011, 23(4): 72-75.

[5] 吴红艳. 图书借阅流量行为季节预测模型[J]. 图书情报工作,2007,51(11):98-101

[6] 王静,李丕仕. 基于 Lyapunov 指数的高校图书馆图

书借阅流量混沌预测[J]. 现代情报, 2009, 29(9): 7-10.

[7] 段玮弘. 基于灰色-马尔柯夫模型的图书借阅行为流量预测研究[J]. 鲁东大学学报(自然科学版), 2011, 27(3):207-212.

[8] 田梅. 基于混沌时间序列模型的图书借阅流量预测研究[J]. 图书馆理论与实践, 2013,22(7):1-4.

[9] 陈静荣. 图书借阅分析系统的数据挖掘技术[J]. 农业图书情报学刊, 2017, 29(2): 69-72.

[10] 程泽,李思宇,韩丽洁,蒋春晓. 基于数据挖掘的光伏阵列发电预测方法研究[J]. 太阳能学报, 2017, 38(3): 726-733.

[11] 张凯斐,刘继华,张菊芳. 大规模高维数据集中局部异常数据挖掘算法[J]. 微电子学与计算机, 2018, 35(3): 116-119.

[12] 高雄飞. 基于混沌最小二乘支持向量机的时间序列预测研究[J]. 数学的实践与认识, 2018, 48(8): 219-224.

作者简介:

尹志强 男,(1983-),图书馆员. 研究方向为数据挖掘、图书馆学. E-mail: yzqv@163. com.