基于数据挖掘算法的地铁站能耗时序预测方法*

罗启鉴[1] 龙 静[2] 陈焕新[1] 刘江岩[1] 李正飞[1]

(1. 华中科技大学能源与动力工程学院,430074,武汉; 2. 广州市地铁集团有限公司,510030,广州//第一作者,硕士研究生)

摘 要 建立了误差反向传播神经网络(BPNN)、决策树分类与回归树(CART)、支持向量回归机(SVR)三种普通的输入-输出预测模型,对地铁站能耗进行预测。基于数据挖掘算法对三个模型进行改进,得到了三种模型基于时间延迟的预测结果,对比了改进前后的预测结果,并确定了最佳的时间延迟。结果表明:普通的输入-输出模型中,SVR 对能耗的预测更加精确;基于时间序列的能耗预测模型对 BPNN 预测模型的提升最大;滞后时长为 5 min 时,三种模型的预测精度最高;基于决策树 CART 算法的时序能耗预测模型对时间延迟的敏感度最高。

关键词 地铁站; 总能耗; 数据挖掘; 时间序列

中图分类号 TK01+8

DOI:10.16037/j.1007-869x.2020.06.006

Time Series Prediction of Subway Station Energy Consumption Based on Data Mining Algorithm

LUO Qiyin, LONG Jing, CHEN Huanxin, LIU Jiang yan, LI Zhengfei

Abstract Three general input-output prediction models: back propagation neural network (BPNN), classification and regression tree (CART) and support vector regression (SVR) are established to predict the energy consumption of subway station. The data mining algorithm is used to improve the three models and the prediction results of them based on time delay are obtained. Through comparing the results before and after the improvement, the optimal time delay is determined. Results show that among the general input-output models, the prediction of SVR model is the most accurate in terms of the energy consumption. The energy consumption prediction model based on time series contributes to the maximum improvement of BPNN prediction model. When the time delay is 5 min, the three models could achieve the best prediction accuracy, but the time series prediction model based on CART is the most sensitive one to time delay.

Key words subway station; total energy consumption; data mining; time series

First-author's address Energy and Power Engineering Institute, Huazhong University of Science and Technology, 430074, Wuhan, China

由于地铁站客流量随时间(一日内的早晚高峰、一周内的工作日与休息日)变化的变动较大,因而站内暖通空调(HVAC)系统冷供应量随时间的变化起伏也较大。如果能提前得知需要提供给乘客的最适宜温度[1],再计算所需的供冷量,就能优化 HVAC 的控制系统,减少机组因滞后指令而浪费的能耗[2-3]。因此,若能提供一种精准、简易的预测工具来预测地铁站能耗,则可优化地铁站的 HVAC 系统、减少额外能耗。

本文采集的数据包括了车站的客流量、室内温度、室外温度、空气湿度等。本文基于数据挖掘算法建立了三种普通的输入-输出模型以及基于时间序列算法的预测模型。通过对这些模型的评价指标进行比较,选择出地铁站能耗预测的最优化模型。

1 地铁站总能耗预测模型的建立

常用的数据挖掘算法分为三类:监督式算法,非监督式算法以及半监督式算法[4-5]。本文采用的三种算法均为监督算法。

1.1 支持向量回归机(SVR)模型

SVR 模型算法建立在统计学理论、VC 理论和结构风险最小原理的基础上[7]。对于给定的训练集 T:

$$T = \{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\} \in (X \times Y)^l \quad (1)$$

其中:

* 国家自然科学基金项目(51576074);华中科技大学自主创新研究基金项目(5003120005);华中科技大学国家级大学生创新训练项目基金项目(16A245)

$$x_i \in X = R^n \quad (2)$$

$$y_i \in Y = \{-1,1\},\ (i = 1,2,\cdots,l) \quad (3)$$

式中：

X——广泛的输入空间；

X_i——X 中任意一个点；

R^n——广泛的输入空间，即特征空间；

y_i——第 i 个学习目标；

Y——学习目标；

l——自然数序列。

式(1)～(3)中，x_i 由 n 个不同的属性特征配对组成。在特征空间内，存在有一个实值函数 $g(x)$，对函数值一一进行分段处理后得到 $f(x) = \mathrm{sgn}(g(x))$。最后根据一定的分类方法将学习目标归类为负数与正类。

置信风险表示机器对位置样本进行分类所得到的误差，经验风险表示训练好的机器对训练样本重新分类得到的误差，而结构风险则为上述两者之和。SVR 的优化目标是将结构风险最小化，而不是传统的经验风险最小化，因此，SVR 具有优秀的泛化能力(即对新样本的适应能力)。

1.2 误差反向传播神经网络(BPNN)模型

BPNN 是一种多层的前馈网络，包含了输入层、隐含层以及输出层[6]。神经网络一般采用梯度下降的方法来控制、调整模型的权值或者阈值，以此使模型的实际输出值和期望输出值的均方误差达到最小。对于隐含层，其表达式为：

$$h_j = f\left(\sum_{i=1}^{n} \omega_{ij} x_i + \alpha_j\right),\ (j = 1,2,\cdots,m) \quad (4)$$

式中：

h_j——输入的加权之和；

ω_{ij}——输入层第 i 个节点和隐含层第 j 个节点之间的连接权值；

f——隐含层激励函数；

α_j——隐含层第 j 个节点的阈值。

输出层的表达式为：

$$Z_k = \sum_{j=1}^{m} h_j \omega'_{jk} + b_k,\ (k = 1,2,\cdots,t) \quad (5)$$

式中：

Z_k——输出层的输出值；

ω'_{jk}——隐含层第 j 个节点和输出层第 k 个节点之间的连接权值；

b_k——输出层第 k 个节点的阈值。

1.3 决策树 CART 模型

当决策树的算法类型为分类与回归树(CART)时，将采用基尼系数作为节点分裂的依据；当 CART 为回归树时，将采用样本的最小方差作为节点分裂的依据。

若采用分类树，其节点的纯度越低，则基尼系数的值越大。基尼系数计算公式为：

$$G = 1 - \sum_{i \in I} p_i^2 \quad (6)$$

式中：

G——基尼系数；

p_i——第 i 个类别的输出概率。

若采用回归树，方差越大则说明节点的数据越分散，其预测的效果越差。回归方差计算公式为：

$$\sigma = \sqrt{\sum_{i \in I} (x_i - \mu)^2} = \sqrt{\sum_{i \in I} x_i^2 - n\mu^2} \quad (7)$$

式中：

μ——回归函数值；

n——节点总数。

如果节点的数据相差很大的话，输出的值则很有可能与实际值相差较大。因此，不管是分类树还是回归树，CART 都需要选择使节点的基尼系数或者回归方差最小的属性作为划分方案。

1.4 时间序列模型

时间序列模型作为一种数据挖掘算法中的常用工具，建立在系统观测数据基础上的时间序列预测模型可以用来预测系统的运行状态或者未来的输出值[7]。其中经典的预测模型有自回归模型(AR)、移动平均模型(MA)、差分整合移动平均自回归模型(ARIMA)以及自回归滑动平均模型(ARMA)，这些算法都已经取得了不小的研究成果[8-9]。但在实际的工程应用中，数据都是非线性的，因此需要结合数据挖掘算法建立非线性的系统。非线性的时间序列预测模型包括有外在输入的非线性自回归模型(NARX)、非线性的自回归模型(RAX)以及非线性的输入—输出模型。

本文基于数据挖掘法，建立了非线性的自回归模型。通过设置一定的滞后时长 t_d，建立地铁站能耗与自身前 d 个值相关的时间序列预测模型。

$$y_n \sim x_1 + x_2 + x_3 + \cdots + x_n$$

$$y_{n+1} \sim x_1 + x_2 + x_3 + \cdots + x_n + y_{n-d} \quad (8)$$

$$t_d = fd \quad (9)$$

式中：

y_n——车站总能耗预测值；

x_n——车站能耗相关的自变量；

y_{n-d}——前 d 个时刻的实际能耗累计值；

t_d——滞后时长；

f——时间分辨率；

d——自然数序列。

1.5　能耗预测模型评价指标介绍

本文采取了五个不同的模型性能评价指标，对上述模型进行评价。这五个指标分别为平均绝对误差（δ_{MAE}）、平均绝对百分误差（δ_{MAPE}）、均方根误差（δ_{RMSE}）、相关系数（cor）以及决定系数（R^2）。

1）δ_{MAE}通过将离差的绝对值化，避免了正负差之间的相互抵消，其计算公式为

$$\delta_{MAE} = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \tag{10}$$

式中：

X_i——第 i 个预测值；

Y_i——第 i 个实测值。

2）δ_{MAPE}是平均绝对误差的另一种更为精准的具现化方式。其计算公式为：

$$\delta_{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|X_i - Y_i|}{Y_i} \tag{11}$$

3）δ_{RMSE}用于描述实测值与预测值之间的偏差大小，与其实测值和预测值有着相同的量纲。计算公式为：

$$\delta_{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}} \tag{12}$$

4）相关系数（cor）反应的是实测值与预测值相互之间的密切程度，可以用于比较预测值相对于实测值的程度高低。其计算公式为：

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{13}$$

式中：

$\overline{X}$—— 预测值的期望值；

$\overline{Y}$—— 实测值的期望值。

5）决定系数（R^2）是模型性能评价指标，用于反映预测值与实测值之间的吻合程度。其计算公式为：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{14}$$

2　地铁站总能耗数据分析

本文所采用的数据来自于北京某地铁站的实际数据，时间分辨率为 1 min[10]。在对原始数据进行加以补差整理后，用以对基于数据挖掘法的时间序列预测方法进行验证。该地铁站室内分隔成两部分（A 部分和 B 部分），而每部分又可再细分为大厅和候车区（简称为 A 厅、A 区、B 厅、B 区）。根据上文所述及其他文献参考，车站的室内和室外温度、室外湿度对 HVAC 的能耗均有所影响，进而影响到整座地铁站的总能耗，故本文采取以上变量作为自变量。

2.1　数据描述

车站客流量的变化存在明显的早高峰变化趋势，这对车站的室内温度有很大的影响，故增加客流量作为自变量。因此，本文采用的自变量包括车站室外温度、车站室外湿度、A 厅温度、A 区温度、B 厅温度、B 区温度以及车站日客流量。所采集数据为同一区域内连续两次遥感观测的最小时间间隔为 5 min 的数据，共收集了 153 个样本。

2.2　未改进总能耗预测模型的结果

采用上述三种不同的算法对地铁站的总能耗进行预测。本研究将数据集进行 7∶3 划分为训练集与测试集，其中：训练集用于对预测模型的训练与优化；测试集则用于对预测模型的结果进行评估。

图 1 反映了三种预测模型在普通输入-输出情况下预测值与实际测量值的吻合情况。决策树

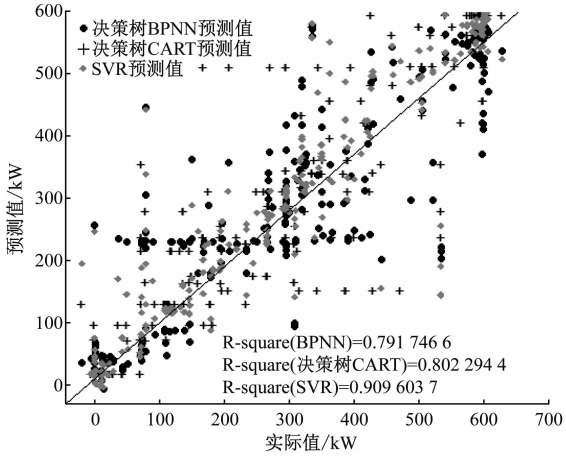

图 1　三种能耗预测模型下的 R^2 图

CART 模型与 BPNN 模型的预测值较实际测量值整体偏大。SVR 模型基本分布在中心线的附近,吻合程度相对较高。

如图 1 所示,三种模型算法的预测结果与实际结果的吻合程度分别为 0.791 746 6(BPNN)、0.802 294 4(决策树 CART)与 0.909 603 7(SVR)。由此可见,SVR 算法在此阶段要全面优于另外两种算法。但这三种模型预测的结果与实际的结果拟合程度均不高。因此对于地铁站的总能耗预测,在输入-输出模型的基础上仅仅采用以上的输入变量是远远不够的,只有提升训练数据的质量或添加更多的输入变量才能达到更为理想的预测效果。若考虑实际因素的影响,这三种模型在未进行改进前均不能达到用于预测地铁站总能耗的目的。

2.3 总能耗预测模型结果的改进

由于上文建立的模型预测效果不理想,故基于数据挖掘法,建立了含外在输入的非线性自回归模型。根据设定一定的滞后时长 t_d,来建立起地铁站总能耗与自身前 d 个值相关的基于三种算法的时间序列预测模型。

在 t_d 为 5 min 的情况下,BPNN 模型在三种模型中受时间序列预测模型的影响最大,效果也最好。图 2 表示了当 t_d 为 5 min 时基于数据挖掘算法的时间序列预测模型产生的能耗预测值与实际值的对比。从图 2 可知,三种预测模型的预测精度都有了很大的提升,改进后的预测值在极值处较原来的预测值更接近于实际值。如表 1 所示,改进后三种模型的 δ_{MAE}、δ_{MAPE} 与 δ_{RMSE} 都低于原来数值的一半,R^2 也都提升到了0.95以上。其中,相较其他算法而言,BPNN 模型的 R^2 提升最大,由最低的 0.791 746 6提升至0.986 128 9,高于 SVR 模型的 R^2(0.984 342 2)。

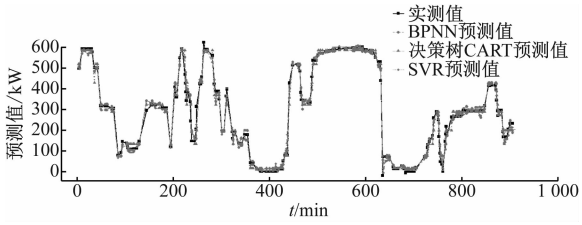

图 2 改进后能耗预测值与实际值的对比

表 1 改进前后各能耗预测模型性能评价指标对比

算法	对比时间	δ_{MAE}	δ_{MAPE}	δ_{RMSE}	r	R^2
BPNN	改进前	63.856 55	0.581 073 7	90.730 45	0.998 627 7	0.791 746 6
	改进后	13.749 79	0.144 934 4	23.415 99	0.993 118 3	0.986 128 9
决策树 CART	改进前	52.626 13	0.430 239 7	88.402 91	0.999 999 1	0.802 294 4
	改进后	20.248 46	0.311 348 0	31.508 68	0.987 770 5	0.974 884 2
SVR	改进前	36.162 43	0.238 404 0	59.776 75	0.999 995 6	0.909 603 7
	改进后	17.639 76	0.138 585 2	24.878 42	0.992 158 7	0.984 342 2

图 3 展示了三种模型的预测吻合程度,可以看出 BPNN 与 SVR 的精确度要优于决策树 CART。决策树 CART 的偏离程度相对较高,但同样能够预测能耗的大致走向与趋势。

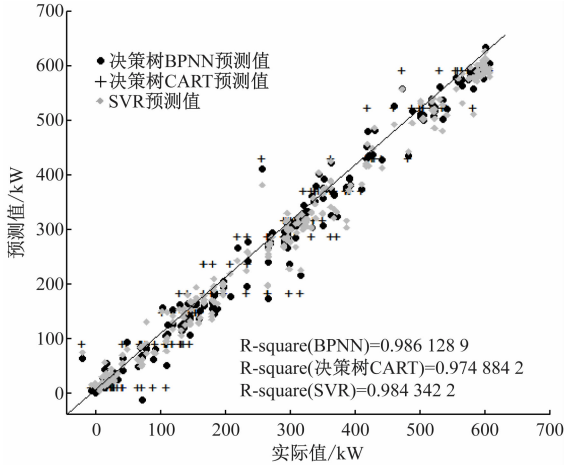

图 3 改进后三种能耗预测模型下的 R^2 图

2.4 滞后时长对模型影响

上文采用的是 t_d 为 5 min 时的预测情况,接下来试验其他情况。表 2 为 t_d 取不同值时三种模型 R^2 的变化情况。随着 t_d 的增加,三种模型的预测精

表 2 不同滞后时长下三种能耗预测模型的 R^2

滞后时长	BPNN	决策树 CART	SVR
5 min	0.986 128 9	0.974 884 2	0.984 342 2
10 min	0.964 335 7	0.949 174 9	0.965 267 3
15 min	0.934 322 7	0.920 015 8	0.957 050 4
20 min	0.915 197 2	0.875 516 1	0.954 246 7

度均在不断下降,其中决策树 CART 算法的变化程度最大,SVR 的变化影响最小。

3 结语

本文分别采用了误差反向传播神经网络(BPNN)、决策树 CART、支持向量回归机(SVR)三种不同的算法,建立了基于数据挖掘算法的时间序列预测模型,对地铁站内的能耗走向进行了预测。利用实际测量某地铁站的总能耗结果,验证了三种预测模型所得到的预测值精度,并通过五个不同的模型性能指标对这三个模型进行了对比及测评,得到结论如下:

1)普通的输入-输出模型中,SVR 模型对能耗的预测最精确;

2)对基于时间序列的能耗预测模型进行改进后,BPNN 模型的提升最大;

3)滞后时长为 5 min 时,三种模型的预测值与实际值的拟合效果最好;

4)基于决策树 CART 算法的时序能耗预测模型对滞后时长的敏感度最高。

参考文献

[1] 朱培根,王春旺,仝晓娜,等. 地铁站乘客动态热舒适评价研究[J]. 暖通空调,2016(2):101.

[2] 张华廷,田雪刚,向灵均. 地铁站空调系统节能潜力分析[J]. 暖通空调,2016(4):7.

[3] 王春,李楠,刘志军,等. 重庆地铁站通风空调系统节能改造[J]. 暖通空调,2017(1):91.

[4] LEUNG P C M, LEE E W M. Estimation of Electrical Power Consumption in Estimation of Electrical Power Consumption in Subway Station Design by Intelligent Approach[J]. Applied Energy, 2013,101(1):634.

[5] WANG Y, FENG H, SONG L, et al. On Energy Saving of Subway HVAC System: Investigation and Autonomous Control [D]. Beijing: Tsinghua University, 2016.

[6] 黄文,王正林. 数据挖掘:R 语言实践[M]. 北京:电子工业出版社,2014:242.

[7] 陈茹雯,黄仁. 非线性自回归时序模型研究及其预测应用[J]. 系统工程理论与实践,2015(9):2370.

[8] WANG Y, WANG J, ZHAO G, et al. Application of Residual Modification Approach in Seasonal ARIMA for Electricity Demand Forcasting[J]. A Case Study of China,2012,48(9):284.

[9] PAPPAS S S, EKONOMOU L, KARAMPELAS P, et al. Electricity Demand Load Forecasting of The Hellenic Power System Using an ARMA Model[J]. Electric Power Systems Research, 2010,80(3):256.

[10] WANG Y, FENG H, XIAO Q. SEED Public Energy and Environment Dataset for Optimizing HVAC Operation in Subway Stations[D]. Beijing: Tsinghua University,2013.

(收稿日期:2018 - 05 - 28)

(上接第 22 页)

被边缘化,若按照上述方法以 B2 型车 190 kW 牵引电机或其他电机替换既有的 180 kW 电机,其加速度亦可适当提高。

3 结语

本文认为 4M2T 与 3M3T 列车都是地铁运营不可或缺的列车。4M2T 列车的粘着力大,爬坡能力较强,适合在线路坡度较大的山城地铁、市域线、城际线,或行车密度大于 30 对/h 的系统运营;3M3T 列车适合在平原城市地铁线路、郊区线路和平均站间距 1km 左右的线路上运营。各个城市应结合自身特点、具体线路条件,合理选择车辆编组及动力配置,降低地铁建设工程投资和运营成本,构建资源节约型社会。

参考文献

[1] 施仲衡. 抓住机遇再创十三五城市轨道交通新辉煌[J]. 都市快轨交通,2016,29(1):卷首语.

[2] 中国城市轨道交通协会. 2017 年统计报告[R/OL]. (2018-01-16)[2018-04-19]. http://www.camet.org.cn/index.php?m=content&c=index&a=show&catid=18&id=13532

[3] 刘书斌. 二线城市轨道交通发展浅思[J]. 都市快轨交通,2016, 29(2):100.

[4] 侯秀芳,左超,李楠. 城市轨道交通 2016 年统计和分析[J]. 都市快轨交通,2017, 30(3):1.

[5] 陆缙华. 关于 6 辆地铁列车编组的动车与拖车配置[J]. 都市快轨交通,2006,3(19):15.

[6] 杨颖,陈中杰. 国内地铁车辆动力配置研究[J]. 城市轨道交通研究,2009(11):26.

[7] 李春成. 关于地铁列车的动力配置[J]. 城市轨道交通研究,2011(2):5.

[8] 冯伯欣. 地铁车辆配置中的新技术应用[J]. 都市快轨交通,2012,25(4):99.

[9] 梁广深,黄隆飞. 地铁 B 型车牵引能耗与再生制动节能效果初探[J]. 城市轨道交通研究,2016(2):27.

[10] 中华人民共和国住房和城乡建设部. 城市轨道交通工程项目建设标准:JB 104—2008[S]. 北京:中国计划出版社,2008:9.

[11] 中华人民共和国住房和城乡建设部,中华人民共和国国家质量监督检验检疫总局. 地铁设计规范:GB 50157—2013[S]. 北京:中国建筑工业出版社,2013:249.

(收稿日期:2018 - 09 - 05)