文章编号：2095-2163(2020)11-0121-05 中图分类号：TP399 文献标志码：A

基于模型树的沪深 300 指数预测

林天华[1]，祁旭阳[1]，张倩倩[1]，赵 霞[2]
(1 河北经贸大学 信息技术学院，石家庄 050061；2 河北经贸大学 经管实验中心，石家庄 050061)

摘 要：针对当前智能算法对证券数据预测准确度不高，以及基于最小损失函数的模型树(Model Tree based on Least Loss Function，MTLLF)预测模型不适用于证券数据的预测的问题，本文提出基于最大离差分裂算法的模型树预测方法(Model Tree based on Deviation Maximization，MTDM)。使用两组包含完整牛熊市的沪深 300 指数日收盘价数据进行分组实验验证，得到的均方误差 MSE(Mean Squared Error)分别为 0.000058 和 0.000140；均方根误差 RMSE(Root Mean Squared Error)分别为 0.007634 和 0.011822；平均绝对百分比误差 MAPE(Mean Absolute Percent Error)分别为 0.011857 和 0.011348 的结果。说明了 MTDM 预测的稳定性较好，且预测准确度较高。并分别与基于长短记忆神经网络(Long Short-Term Memory，LSTM)和粒子群优化算法(Partical Swarm Optimization，PSO)的预测方法进行实验对比，结果表明 MTDM 算法的预测误差显著低于前两者。

关键词：机器学习；模型树；分裂算法；沪深 300 指数预测

CSI 300 index forecast based on model tree

LIN Tianhua[1]，QI Xuyang[1]，ZHANG Qianqian[1]，ZHAO Xia[2]
(1 College of Information Technology，Hebei University of Economics and Business，Shijiazhuang 050061，China；
2 Department of Economic Management Experiment，Hebei University of Economics and Business，Shijiazhuang 050061，China)

【Abstract】 To solve the problem that the accuracy in predicting securities data of current intelligent algorithms is low and the Model Tree based on Least Loss Function (MTLLF) prediction model is not suitable for the prediction of securities data，Model Tree based on Deviation Maximization (MTDM) is put forward. Using two sets of CSI 300 index daily closing price data containing complete bull and bear markets for grouping experimental verification，the obtained MSE (Mean Squared Error) are 0.000058 and 0.000140，the RMSE (Root Mean Squared Error) are 0.007634 and 0.011822 and the MAPE (Mean Absolute Percent Error) are 0.011857 and 0.011348. Indicating that the MTDM prediction has good stability and high prediction accuracy. Compared with the prediction methods based on Long Short-Term Memory (LSTM) and Particle Swarm Optimization (PSO)，the results show that the prediction error of MTDM algorithm is significantly lower.

【Key words】 Machine learning；Model tree；Split algorithm；CSI 300 index forecast

0 引 言

股票是市场经济的重要体现，在一定程度上反映着我国的经济发展状况，在经济发展走势分析中发挥着重要作用。沪深 300 指数是股票市场的重要指数之一，它能够反映沪深两市市场整体表现和价格变动。预测沪深 300 指数在指导沪深两市个股投资和分析沪深市场变化等方面具有重要意义。预测沪深 300 指数的研究方法主要分为三种，分别是基本面分析法、技术分析法和量化分析法。其中量化分析法是利用计算机技术进行统计、数值模拟，进而研究证券数据的一种方法[1]。该方法分析的数据量大、形成的模型严格，因此能够取得较好的分析效果。

将机器学习、神经网络等现代预测方法应用于股指的分析和预测是当前的一个研究热点。熊涛[2]等提出基于自组织神经网络(Self Organizing Neural Network，SOM)和支持向量机(Support Vector Machine，SVM)的多步预测方法，即先用 SOM 对沪深 300 指数序列进行聚类，随后基于划分后的数据集分别构建 SVM，得到多步预测模型，结果表明该模型的预测效果要好于单一的 SVM。唐艳琴[3]等为解决基于 SVM 的预测模型复杂、耗时长的问题，提出了一种基于多输出的学习方法，该模型在预测沪深 300 指数时比 SVM 预测的均值方差提高了约 10 倍，运行时长也减少了近 3/4。文献[4]提出了使用多支持向量机对股指进行混合频率抽样预测方法。文献[5]提出将夏普比率引入到 SVM 股指预测中，提升投资回报。周荣谦[6]提出的基于 Morlet

基金项目：河北省自然科学基金(F2019207061)。
作者简介：林天华(1979-)，男，硕士，副教授，主要研究方向：数据挖掘与分析；祁旭阳(1994-)，女，硕士研究生，主要研究方向：数据挖掘与分析；张倩倩(1996-)，女，硕士研究生，主要研究方向：数据挖掘与分析；赵 霞(1979-)，女，博士，副教授，主要研究方向：数据挖掘与分析。
通讯作者：祁旭阳 Email：424660509@qq.com
收稿日期：2020-09-09

小波核函数 SVM 的沪深 300 指数预测方法,得到了较低的 RMSE,预测效果较好。文献[7]结合小波变异的混合函数连接人工神经网络和粒子群优化算法,对沪深 300 指数进行了预测。文献[8]和文献[9]分别使用 ModAugNet 框架和多隐层人工神经网络混合模型对标准普尔 500 指数进行预测,预测误差均较低。戴德宝等[10]使用文本挖掘和情感分析方法,生成投资者情绪时间序列,并使用 SVM 和神经网络对上证投资者情绪综合指数进行预测。冯宇旭[11]等提出的基于长短期记忆神经网络的沪深 300 指数预测方法,比同一测试集上的 Adaboost 算法得到的 RMSE 要低。文献[12]提出特征值归一化加权多线性主成分分析对恒生指数进行特征提取,并使用 SVM 预测。文献[13]将 logistic 回归(LR)模型级联到梯度增强决策树(Gradient Boosting Decision Tree,GBDT)模型上,由此构成股指预测模型,并对上证指数、纳斯达克指数和标准普尔 500 指数进行预测,预测准确率较高。

综上所述,现有文献中使用机器学习算法对沪深 300 指数预测较少,且仅有的研究得到的预测效果也欠佳。模型树是机器学习中的一种算法,从理论上看,相较于其它机器学习算法,它具有叶子节点是分段线性函数的特性,能够更好得拟合连续型数据,得到较好的预测效果,从而更适用于预测领域。在应用方面,模型树算法在众多数值型变量的预测问题中,证实了其有理想的预测性能。张建明[14]等将模型树算法用于汽轮机汽耗性预测、GOYAL M K[15]等将模型树算法应用于闸下冲刷预测、李建更[16]提出用模型树预测 $PM_{2.5}$ 浓度,均取得了较好的预测效果,证实了它在连续值预测方面的可行性。因此,本文将基于模型树算法构建预测模型,改进模型树的分裂算法,使其适用于沪深 300 指数预测,提高预测的准确度,这在理论分析和实际应用中都具有重要意义。

1 模型树算法

本文使用目前常见的基于最小损失函数的模型树算法进行证券数据分裂,并针对证券数据的特征进行改进,提出了基于最大离差分裂算法的模型树。

1.1 基于最小损失函数的模型树算法

基于最小损失函数的模型树是分类回归树(Classification And Regression Trees,CART)的变体,既可以用于分类也可以用于回归。其对样本数据集进行二分递归分裂,最终形成一棵以叶节点为分段线性函数的二叉树,并对生成的模型树进行后剪枝,得到最优模型树。模型树作为回归模型时,给定数据集 $D=\{(x_1,y_1),\dots,(x_i,y_i),\dots,(x_n,y_n)\}$,则生成初始模型树 MT_0 的步骤如下:

Step 1 求解式(1),得到最优的特征 A 和特征分裂点 s,

$$\min_{A,s}\left[\min_{c_1}\sum_{x_i\in D_1(A,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in D_2(A,s)}(y_i-c_2)^2\right]. \tag{1}$$

其中,c_1 为数据集 D_1 的均值,c_2 为数据集 D_2 的均值。

Step 2 用选定的(A,s)将当前数据集划分成 D_1 和 D_2 两个数据集。

Step 3 分别对 D_1 和 D_2 两个数据集进行线性回归,得到分段线性函数 f_1 和 f_2,作为当前父节点的两个子节点。

Step 4 对每个子节点执行上述步骤,直至满足停止条件。

Step 5 输出生成的模型树 MT_0。

直接采用生成的 MT_0 做预测,往往会产生过拟合现象,需要对其进行剪枝操作,但又要防止剪掉一些节点后导致预测的误差增加。因此,采用代价复杂度剪枝算法进行后剪枝。具体算法如下:

输入 生成的模型树 MT_0

输出 最优模型树 MT

Step 1 设 $k=0$,$MT=MT_0$,$\gamma=+\infty$。

Step 2 自下而上遍历每个内部节点 t,并计算 $C(T_t)$、$|T_t|$ 和整体损失函数的减少程度 $g(t)$。计算公式见式(2)和(3)。

$$g(t)=\frac{C(t)-C(T_t)}{|T_t|-1}, \tag{2}$$

$$\gamma=\min(\gamma,g(t)). \tag{3}$$

其中,T_t 是以 t 为根节点的子树,$C(T_t)$ 是对训练数据的预测误差;$|T_t|$ 是 T_t 的叶节点个数。

Step 3 自上而下访问内部节点。若 $g(t)=\gamma$,则剪去该分支,得到树 MT_t。

Step 4 设 $k=k+1$,$\gamma_k=\gamma$,$MT_k=MT_t$。

Step 5 如果 MT_t 不是由根节点单独构成的树,则回到 Step 3。

Step 6 使用交叉验证法在子树序列 MT_1,MT_2,…,MT_n 中选取最优子树 MT。

1.2 基于最大离差分裂算法的模型树

由于基于最小损失函数的模型树计算得出的分裂点不理想(如图 2),导致预测效果不好,故对其分裂算法进行改进,提出最大离差分裂算法,使得其能

够适用于证券数据的分裂，提高预测的准确度。

基于最大离差分裂算法的模型树的主要算法流程如下：

输入　沪深 300 指数数据集 Y

Step 1　对全体沪深 300 指数数据 Y 进行线性回归，得到初始的线性回归直线 L_{parent} 及对应的线性回归函数 y_{line}。L_{parent} 与实际值的首次和最后一次交点，分别为 *start* 和 *end*。

Step 2　搜索分裂属性。对已构建的线性回归函数搜索分裂属性，并将分裂属性取并集，即回归属性集合。

Step 3　生成分裂点和线性回归函数。若第 i 个交易日在 *start* 和 *end* 之间，即 $i \in [start, end]$，则从沪深 300 指数数据中选择与 L_{parent} 上的点距离最远的点，作为分裂点 *splitPos*，其计算方法如式(4)、(5)。

$$\triangle y_i = | y_i - y_{line\,i} |, \tag{4}$$

$$splitPos = i,$$

$$\text{s. t. } \triangle y_i = \max \triangle y_i. \tag{5}$$

以此将数据分为左右两段，并对两段数据分别进行线性回归，得到 L_{left} 和 L_{right}。线性回归函数为 y_{left} 和 y_{right}，二者分别作为父节点的左右子树。将得到的 L_{right} 作为 L_{parent}，y_{right} 作为 y_{line}。

Step 4　遍历递归，生成模型树。递归执行 Step2 和 Step3，至达到阈值条件，即 $end - start < 10, R > 0.9$。其中 R 为最大相关系数，最后生成的右子树为 L_{latest}。

Step 5　构建好模型树 MT，使用沪深 300 测试集数据进行预测。以 L_{latest} 作线性回归预测，计算并输出预测衡量指标，则算法结束。

最大离差分裂算法流程如图 1 所示。

图 1 中，y_{line} 为原始沪深 300 数据进行线性回归得到的回归方程；i 表示第 i 个交易日；y_i 表示第 i 个交易日的真实值；y_{line_i} 表示第 i 个交易日的线性回归值；*splitPos* 表示分裂点；R 为最大相关系数。

2　实证分析

2.1　预测评价指标

本文使用均方误差 MSE，均方根误差 RMSE 和平均绝对百分比误差 MAPE 作为预测评价指标，用于描述预测值偏离真实值的程度。三者的计算方法如公式(6)~公式(8)。

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y(i) - \hat{y}(i))^2, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y(i) - \hat{y}(i))^2}, \tag{7}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n} | y(i) - \hat{y}(i) | / y(i). \tag{8}$$

其中，$y(i)$ 为第 i 个交易日沪深 300 指数收盘价的真实值；$\hat{y}(i)$ 为第 i 个交易日沪深 300 指数收盘价的预测值；n 为样本总数。

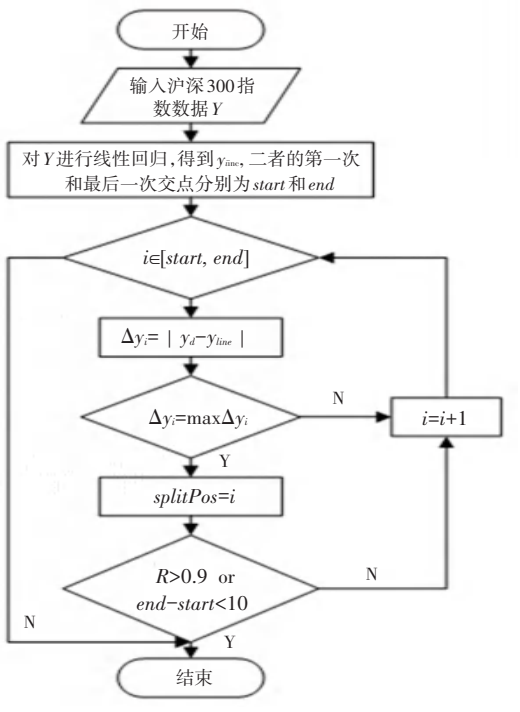

图 1　最大离差分裂算法流程图

Fig. 1　Maximum deviation splitting algorithm flow chart

由上述公式可知，三者的值越小则说明模型预测的结果误差越小，即与真实值越接近，预测效果也越好。

2.2　实验数据及预处理

2.2.1　MTDM 算法分组对比样本数据的选取

本文选取两组时间段的沪深 300 指数日收盘价，作为训练样本数据和测试样本数据。2007 年 8 月 15 日至 2008 年 11 月 6 日的 300 个交易日的收盘价作为第一组的训练样本数据，2008 年 11 月 7 日至 2014 年 7 月 16 日的1 381个交易日的收盘价作为第一组的测试样本数据。2013 年 4 月 20 日至 2014 年 7 月 16 日的 300 个交易日的收盘价作为第二组的训练样本数据，2014 年 7 月 17 日至 2019 年 1 月 4 日的1 092个交易日的收盘价作为第二组的测试样本数据。

在两组数据的测试样本数据中，均包含了完整的上涨牛市数据、下跌的熊市数据以及震荡数据，使得实验能充分包含前述几种情况，更好地验证模型预测的有效性。

2.2.2 MTDM 算法与其他算法对比样本数据的选取

在与其他预测算法进行对比时,保持与原实验一致的时间段数据作为数据样本,即将文献[11]提出的 LSTM/Adaboost、SVR/LSTM/Adaboost 回归集成算法应用于 2012 年 5 月 3 日~2017 年 9 月 4 日的沪深 300 指数的预测;文献[6]提出的 PSO 算法优化,应用于 2015 年 12 月 11 日~2016 年 11 月 12 日的沪深 300 指数的预测。将基于最大离差分裂算法的模型树的沪深 300 指数模型分别用于上述两个时间段,其中训练样本数据在此基础上分别增加 300 个交易日收盘价数据,即 2011 年 2 月 1 日~2012 年 5 月 2 日、2014 年 9 月 17 日~2016 年 11 月 11 日分别作为二者的训练样本数据,从而保持对比实验的一致性。

2.2.3 数据预处理

在预测时,由于原始数据差距较大,直接输入模型树预测模型,预测误差较大。为保证模型测预效果,采用归一化方法处理这些数据,经过线性变换,可以映射到[0,1]范围内,归一化表达式如公式(9):

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{9}$$

其中,x' 为归一化后的数据,$x_{\min}$、$x_{\max}$ 分别为样本数据的最小值和最大值。

2.3 分裂和预测效果

为保证展示效果,在此与 LSTM 算法预测方法对比的数据,以 2011 年 2 月 1 日至 2017 年 9 月 4 日,共 1603 个交易日的沪深 300 收盘价数据为例,说明分裂过程;以该对比实验第一年的测试数据,即 2012 年 5 月 3 日至 2013 年 11 月 5 日共 365 个样本数据说明预测效果。

(1)基于最小损失函数的模型树分裂效果

基于最小损失函数的模型树分裂效果如图 2 所示。

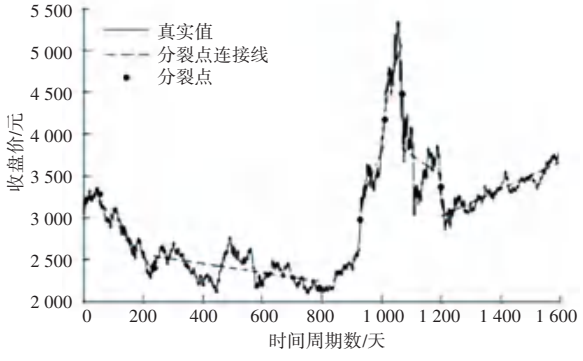

图 2 基于 MTLLF 算法的分裂效果图

Fig. 2 Splitting effect based on MTLLF algorithm

其中,折线表示真实值;圆点表示回归分裂点;虚线表示相邻分裂点的连接线。由图 2 可见,分裂点连接线的走势没有反映沪深 300 指数的走势特征,导致分裂效果不好,不能够很好地应用于证券数据分析当中。

(2)基于最大离差分裂算法的模型树分裂效果

基于最大离差分裂算法的模型树分裂效果如图 3 所示。

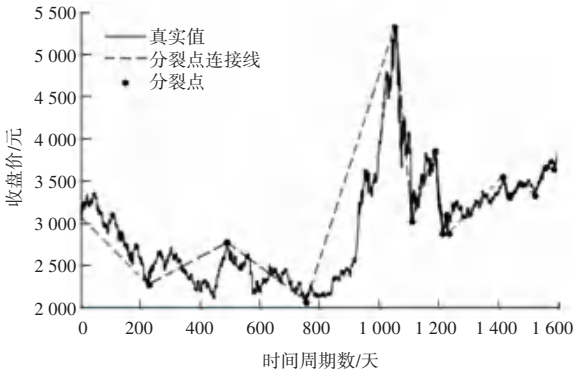

图 3 基于 MTDM 算法的分裂效果图

Fig. 3 Split effect diagram based on MTDM algorithm

由图 3 可以看出,每个圆点都落在代表真实值折线的拐点处,分裂点连接线的走势与沪深 300 指数的走势基本契合,分裂效果理想,适应证券数据的特征,为后续的预测奠定了基础。

(3)基于最大离差分裂算法模型树的预测效果

使用基于最大离差分裂算法的模型树,对沪深 300 数据进行预测,得到的预测结果如图 4 所示。

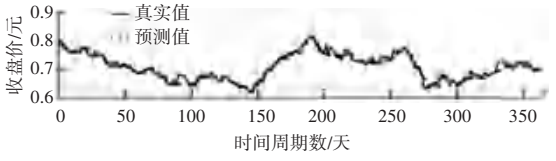

图 4 基于 MTDM 算法的预测效果图

Fig. 4 Forecast effect diagram based on MTDM algorithm

由图 4 可见,基于 MTDM 算法的预测结果接近真实值,与真实值的拟合程度较高,预测效果较好。

2.4 预测性能对比分析

(1)MTDM 算法分组实验预测性能对比

使用 MTDM 算法模型对前述两组实验数据进行预测,得到的 MSE、RMSE 和 MAPE 见表 1。

表 1 分组实验性能对比表

Tab. 1 Performance comparison table of grouping experiment

组别	MSE	RMSE	MAPE
第一组	0.000 058	0.007 634	0.011 857
第二组	0.000 140	0.011 822	0.011 348

由表 1 可知,MTDM 预测方法在不同长度的时

间段内的预测误差变化较小。对于牛市、熊市以及震荡市场数据的预测均具有较好的适用性，预测稳定性和预测精度都有较好的表现。

(2)MTDM 与其他算法预测性能对比

MTDM 算法与基于 LSTM 的预测方法以及 PSO 优化预测方法进行对比，得到的 MSE、RMSE 和 MAPE 分别见表 2、表 3。

表 2　与基于 LSTM 预测方法的性能对比表

Tab. 2　Performance comparison table with the prediction method based on LSTM

方法	MSE	RMSE	MAPE
SVR/LSTM/Adaboost	0.001 279	0.035 758	—
LSTM/Adaboost	0.000 553	0.023 514	—
MTDM(本文算法)	0.000 115	0.010 730	0.011 169

表 3　与 PSO 算法优化预测方法的性能对比表

Tab. 3　Performance comparison table with PSO algorithm optimization prediction method

方法	MSE	RMSE	MAPE
PSO 算法优化	0.000 256	0.016 000	—
MTDM(本文算法)	0.000 091	0.009 527	0.010 255

由表 2、3 可知，MTDM 预测方法的预测误差显著低于其他算法，具有更好的预测效果。

3　结束语

本文使用最大离差分裂算法改进了模型树，使得模型能够适应证券数据的特征，经不同时间段的沪深 300 指数预测实验验证，以及与其他预测方法的对比，表明本模型具有良好的预测准确度和稳定性。

基于最大离差分裂算法的模型树预测模型在找到分裂点并分裂数据后，仅用模型树的最右子树进行预测，丢失了兄弟节点、父节点之间的关系。下一步拟使用多叉模型树，利用节点间的关系、最右子树等所有分裂信息构建预测模型，进一步减小预测误差，提高预测准确率。

参考文献

[1] 张然，汪荣飞，王胜华. 分析师修正信息、基本面分析与未来股票收益[J]. 金融研究，2017(7)：156-174.

[2] 熊涛，鲍玉昆，胡忠义，等. 基于 SOM 和 SVMs 的沪深 300 指数多步预测[J]. 系统工程，2012，30(10)：36-42.

[3] 唐艳琴，潘志松，张艳艳. 基于多输出学习的沪深 300 指数预测研究[J]. 计算机科学，2017，44(S2)：106-109.

[4] PAN Y，XIAO Z，WANG X N，et al. A multiple support vector machine approach to stock index forecasting with mixed frequency sampling[J]. Knowledge-Based Systems，2017(122)：90-122.

[5] HU Z，BAO Y，CHIONG R，et al. Profit guided or statistical error guided? a study of stock index forecasting using support vector regression[J]. Journal of Systems Science and Complexity，2017，30(2)：1-18.

[6] 周荣谦. 基于 Morlet 小波核函数支持向量机的沪深 300 指数预测研究[D]. 天津：天津财经大学，2017.

[7] LU T，LI Z Y. Forecasting CSI 300 index using a Hybrid Functional Link Artificial Neural Network and Particle Swarm Optimization with Improved Wavelet Mutation[J]. International Journal of Advanced Network，Monitoring and Controls，2018，2(3)：241-246.

[8] BAEK Y，KIM H Y. ModAugNet：A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module[J]. Expert Systems With Applications，2018(2)：457-480.

[9] Seo M，Lee S，Kim G. Forecasting the Volatility of Stock Mar ket Index Using the Hybrid Models with Google Domestic Trends[J]. Fluctuation and Noise Letters，2019，18(1)：1-17.

[10] 戴德宝，兰玉森，范体军，等. 基于文本挖掘和机器学习的股指预测与决策研究[J]. 中国软科学，2019(4)：166-175.

[11] 冯宇旭，李裕梅. 基于 LSTM 神经网络的沪深 300 指数预测模型研究[J]. 数学的实践与认识，2019，49(7)：308-315.

[12] GUO Z Q，WANG X. Time series forecasting of stock market index based on ENWMPCA model[J]. Journal of Physics：Conference Series，2019，12-37(5)：1-7.

[13] FENG Z，QunZhang. Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices[J]. Applied Soft Computing Journal，2019，84：1-17.

[14] ZHANG J M，LIU D. Working condition characteristics identification for extraction unit by using M5 model tree and measured data[J]. Proceedings of the CSEE，2017，31(23)：21-26.

[15] GOYALM K，OIHA CSP. Estimation of scour downstream of a ski-jump bucket using support vector and model tree[J]. Water Resources Management，2008(25)，2177-2195.

[16] 李建更，吴水生. 基于 PLS-M5P 模型的 PM_(2.5)浓度预测[J]. 计算机与应用化学，2018，35(12)：959-970.

(上接第 120 页)

[8] MA W，LIU Y，HEAD K L. Optimization of pedestrian phase patterns at signalized intersections：a multi-objective approach[J]. Journal of Advanced Transportation，2014，48(8)：1138-1152.

[9] 杨晓光，马万经，林瑜.两相位信号控制交叉口行人专用相位设置条件研究[J]. 公路交通科技，2005(1)：127-131.

[10] YU C，MA W，YANG X. Integrated optimization of location and signal timings for midblock pedestrian crosswalk[J]. Journal of Advanced Transportation，2016，50(4)：552-569.

[11] YANG Z. Signal timing optimization based on minimizing vehicle and pedestrian delay by genetic algorithm[J]. Guangdong Agricultural Sciences，2010.

[12] LI X，SUN J Q. Intersection multi-objective optimization on signal setting and lane assignment[J]. Physica A：Statistical Mechanics and its Applications，2019，525：1233-1246.

[13] 冯树民，裴玉龙. 行人过街延误研究[J]. 哈尔滨工业大学学报，2007(4)：613-616.