基于时空特征的地铁客流预测

张和杰 马维华
(南京航空航天大学计算机科学与技术学院 南京 211106)

摘 要 随着城市轨道交通的迅速发展,地铁短期断面客流的预测有利于运营部门观测客流的实时变化,从而调整调度策略。客流具有时空特征,在10 min粒度时间片下,客流变化存在周期性,在空间上客流波形存在差异性。使用凝聚层次聚类算法对不同站点在一周内的客流进行聚类分析,得到贴近站点特征的客流分类结果。根据分类结果,对不同类别客流时间片分别进行相关性分析,提出一种基于SVM的预测模型,将强相关性的时间片序列作为模型输入。同时,提出一种基于协同自适应调整的双种群萤火虫算法以寻优模型参数,算法中引入混沌吸引度来提高算法的全局搜索能力,避免由于初始值陷入局部最优;加入自适应搜索步长,以加快算法的收敛速度并提高求解精度。与其他模型和优化算法的对比表明,本模型具有较好的预测精度、稳定性和鲁棒性。

关键词 客流预测,支持向量机,时间序列,萤火虫算法,混沌

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.045

Subway Passenger Flow Forecasting Model Based on Temporal and Spatial Characteristics

ZHANG He-jie MA Wei-hua
(School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract With the rapid development of urban rail transit, the short-term passenger flow forecast of the subway is conducive to the operation department to observe the real-time changes in passenger flow and adjust the scheduling strategy. This paper studied the temporal and spatial characteristics of passenger flow. Under the 10-minute granular time slice, there is a periodicity of passenger flow changes, and there are differences in the waveform of passenger flow in space. This paper used agglomerative hierarchical clustering algorithm to analyze the passenger flow of different stations for a week, and obtained the results of passenger flow close to the characteristics of the station. According to the results of classification, correlation analysis was performed on time slices of different types of historical passenger flow, and prediction models based on Support Vector Machine were proposed, regarding time slice sequences with strong correlation as input. Besides, a parameter optimization model for double-population firefly algorithm based on cooperative self-adaptive adjustment was proposed, in which the chaotic attraction was introduced to improve the global search ability, avoiding the initial value being trapped into a local optimum. The adaptive search step length was added to improve the convergence speed and solution accuracy. Compared with other models and optimization algorithms, the proposed model has better prediction accuracy, stability and robustness.

Keywords Forecast of passenger flow, SVM, Time series, FA, Chaos

1 引言

随着经济的快速发展,城市现代化建设的步伐加快,城市轨道交通被规划到越来越多的城市发展计划当中。地铁作为城市轨道交通的重要组成部分,得到了快速的发展;随着地铁线路的增多,地铁线路结构越来越复杂,客流量逐步增大,其中客流量是影响地铁运营调度的主要因素之一。客流预测中,短期客流逐渐成为研究的热点,短期客流的预测能有效提高车辆调度水平,疏散客流,确保市民安全出行。

当前短期客流预测中,主要有基于线性预测、非线性预测的方法等。基于线性理论的交通流预测方法主要包括历史客流预测方法、时间序列预测方法和卡尔曼滤波模型预测方法等,但此类方法预测性能一般。由于客流部分属性呈非线性关系,非线性预测方法被运用到了客流预测中,并且具有较高的准确度,但此类方法,如支持向量机[1]、灰色理论[2]、模糊逻辑[3]等,随着客流样本的增多,计算复杂性升高,需要进一步研究和完善。

文献[4]采用Spearman相关系数分析,将站点待预测时

到稿日期:2018-06-05 返修日期:2018-10-16

张和杰(1992—),男,硕士生,主要研究方向为城市轨道交通客流预测、嵌入式系统,E-mail:978508554@qq.com;**马维华**(1960—),男,硕士,教授,主要研究方向为嵌入式系统及应用,E-mail:mwhua@nuaa.edu.cn(通信作者)。

段的前 4 个时间片客流和上一周相同时间段前后的时间片客流作为 BP 神经网络的输入特征，采用基因遗传算法与 BP 神经网络的组合模型使得 BP 神经网络不易陷入局部最优，相对于单独的 BP 模型，MAPE 预测误差大幅降低。文献[5]采用 AP 聚类算法对兰州交大站的客流样本进行聚类，将一周的客流样本分为 2－6 类，对每类样本分别用支持向量机(Support Vector Machine，SVM)预测，结果表明分为 6 类时预测性能最优，构建的混合 SVM 的客流预测性能比单 SVM 预测模型的性能好。文献[6]基于人工神经网络(ANN)的人体感觉模型，从每个乘客主观感受的便利性出发，对客流进行分段预测，取得了较好的预测效果。文献[7]采用 BP 神经网络算法，在分析轨道交通人员上下班的基础上，建立了短期客流预测模型，实验结果表明神经网络在短时交通流的预测上具有一定优势。文献[8]针对 SVM 中由于参数选择不当会导致过度拟合、成对分类和参数正则化等问题，引入萤火虫算法(Firefly Algorithm，FA)来优化 SVM 的参数，通过评估其性能，推导出其优于其他元启发式算法的性能。文献[9]提出使用滚动式时间序列方法对轨道交通客流进行预测，并使用实际数据验证了该预测方法相对于传统方法具有更好的预测精度。

综上所述，根据客流的时空特征，本文提出一种基于时间序列的客流预测模型，利用 SVM 在解决非线性问题时的优势，将非线性历史时间序列作为 SVM 的输入参数来回归预测下一若干时间片的客流，即对客流时间序列进行预测，并提出一种改进的萤火虫算法来寻优 SVM 参数，最终得到双种群自适应混沌萤火虫(Double-population Adaptive Cooperative Firefly Algorithm，DACFA)SVM 模型，以预测客流量。

2　客流特征分析

地铁客流是指在单位时间内，线路上乘客的流动方向及人数的总和。城市轨道交通客流在概念上不仅揭示了乘客在空间上的位移和数量，而且强调了乘客的位移具有方向性和起讫位置[5]。

在不同的时间尺度以及空间地点，地铁客流都展现出不同的特征。根据时间尺度，可以将其分为一周客流、全日运营时间段客流、高峰期客流。根据空间位置，可以将客流分为站点客流和断面客流，其中，站点客流可以细分为换乘客流和出入站客流[5]。

2.1　客流时空特征

地铁客流在时间上的分布是指全日客流量在一天内各个时间段上的分布。本文主要分析一天内小时客流量波形和一周内日客流量波形的变化规律。

地铁线路客流在空间上的分布呈现不均衡的分布特征，空间上不同站点的日客流在时域上表现出不同的波形特性。

根据不同站点在一天内的进站客流数据统计，客流单位时间粒度下的断面客流分布曲线主要分为以下几种类型。

(1)单峰型

在潮汐现象较为明显的地区，站点容易出现单峰型的客流，该车站的进出站高峰客流在早晚将会错开，客流一般相对集中。一般情况下，可将该类客流分为两种类型：1)早高峰进站人数较多，而晚高峰出站人数较多；2)早高峰出站人数较多，晚高峰进站人数较多。单峰型客流示意图如图 1 所示。

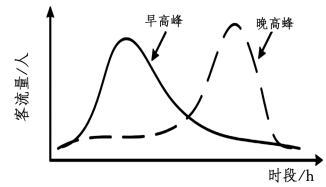


图 1　单峰型客流

Fig. 1　Single peak passenger flow

(2)双峰型

位于城市繁华商业区域、学校等地区的站点，其早、晚高峰客流明显。双峰型客流示意图如图 2 所示。

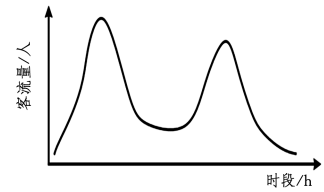


图 2　双峰型客流

Fig. 2　Double peak passenger flow

(3)平稳型

一般在火车站、机场、高流量的地区，客流为一种平稳状态，不会出现明显的高峰客流。平稳型客流示意图如图 3 所示。

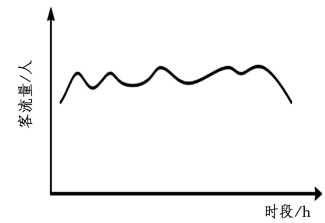


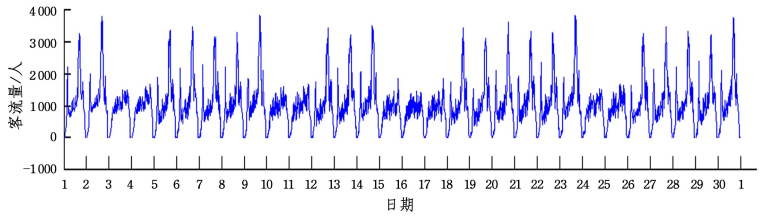
图 3　平稳型客流

Fig. 3　Smooth passenger flow

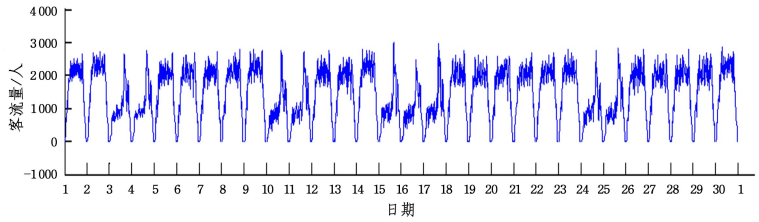
2.2　客流聚类分析

以南京三号线为例，在 2016 年 9 月，站点 S_1，S_2 进站客流的时间序列如图 4 所示。该线路的运营时间为 6:00－23:00，由于进站客流具有提前性和滞留性，将研究时间段设为5:00－24:00，以 10 min 为时间粒度，将运营时间段划分 114 个时间片。

从图 4 中可以看出，除了中秋节 15 号(周四)、16 号(周五)、17 号(周五)3 天假期之外，地铁总体客流整体上表现为按一周时间呈现周期变化，且一周内的客流呈现不同特征，存在明显的双峰型客流、平稳型客流和单峰型客流，不同站点的客流特征也不尽相同。根据客流的周期性变化，可使用聚类方法对客流进行聚类分析，以得出一周内客流的类别特征。



(a)S_1 客流分布情况



(b)S_2 客流分布情况

图 4 部分站点 9 月的客流量分布

Fig. 4 Distribution of passenger flow in some sites in September

层次聚类就是对数据集采用某种方法逐层地进行分解或者汇聚,直到分出最后一层的所有类别数据满足要求为止[10-13]。按照分解或者汇聚原理的不同,层次聚类可以分为凝聚和分裂两种方法[14]。一般对于非线性、流型结构的样本,采用凝聚层次聚类算法十分有效,且稳定性较好[15]。地铁客流时间序列属于非线性流型结构,因此,本文采用凝聚层次聚类方法对客流进行聚类分析。

首先将所有周一至周日的客流数据分别求均值,并整理为时间序列$\{x_1,x_2,\cdots,x_7\}$,假设 x_1 是长度为 114 的所有周一数据的均值序列。利用凝聚层次聚类算法对客流进行聚类,算法的具体步骤如下:

1)把每一个样本分为一类,即将一周分为 7 类,$G_i=\{x_i\}$ $(i=1,2,\cdots,7)$;

2)求出各类间的欧氏距离矩阵$(D_{ij})_{c\times c}$,其中 c 为当前分类数;

3)找出$(D_{ij})_{c\times c}$中的最小值D_{nm},然后将类 G_n 和 G_m 合并为新类 $G_p=\{G_n,G_m\}$,$c=c-1$;

4)若 c 为指定聚类数,则聚类结束,否则跳转至步骤 2)。

根据聚类算法,将最终样本的分类数设为 $c=1$,得到如图 5、图 6 所示的系谱图。

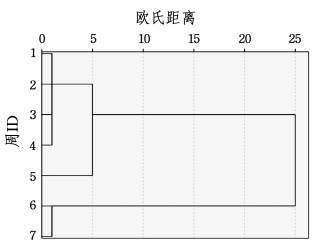


图 5 S_1 车站客流聚类系谱图

Fig. 5 Passenger flow clustering spectrum at S_1 station

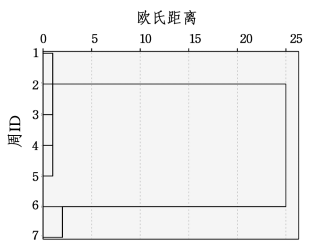


图 6 S_2 车站客流聚类系谱图

Fig. 6 Pedigree spectrum of passenger flow in S_2 station

系谱图的横坐标表示类间欧氏距离,纵坐标表示周一到周日。根据欧氏距离的大小,在 S_1 站点上,最终可将周一至周四归为一类,将周五归为一类,周六、周日归为一类。文献[5,14]指出周五客流与周一至周四客流的相异性较大,与休息日或节假日前一天有关,所以通过系谱图可得出:周五客流时间序列较周一至周四客流时间序列具有更大的欧氏距离,表明这两类客流间的特征差异较大。经过聚类可将 S_1,S_2 站进行分类,结果如表 1、表 2 所列。

表 1 S_1 车站客流类别

Table 1 Passenger classes of S_1 station

客流类别	周类别
A	周一至周四
B	周五
C	周六和周日

表 2 S_2 车站客流类别

Table 2 Passenger classes of S_2 station

客流类别	周类别
A	周一至周五
B	周六和周日

由此,可对不同站点的客流分别建模,以解决因站点空间特征的不同导致模型在部分站点的性能表现较好而在另一些站点的性能较差的问题;同时将客流时间特征细化,对不同波形的客流进行有效建模,以提高预测性能。

3 萤火虫算法

3.1 基本萤火虫算法

2009 年,剑桥学者 Yang 根据自然界中萤火虫的发光行为提出萤火虫算法[10]。该算法把解空间中的各点看成萤火虫,利用了发光强的萤火虫会吸引发光弱的萤火虫的特点,在发光弱的萤火虫向发光强的萤火虫移动的过程中,萤火虫的亮度和吸引度不断更新,完成位置的迭代,从而找出最优位置。其中,亮度取决于萤火虫所处位置的优劣并决定其移动方向,吸引度决定了萤火虫移动的距离,萤火虫的亮度和吸引度与萤火虫之间的距离成反比[10]。

根据萤火虫的位置 $x_i=(x_{i1},x_{i2},\cdots,x_{id})$(x_{id} 为第 d 维数据),定义任意两只萤火虫 i 和 j 的相对亮度为:

$$I_{ij}(r_{ij})=I_o e^{-\gamma\cdot r_{ij}^2} \tag{1}$$

其中，I_o 为萤火虫的绝对亮度值；γ 为光吸收系数，因为荧光会随着距离的增加和传播媒介的吸收逐渐减弱，所以将 γ 设为常数；r_{ij} 为萤火虫 i 和 j 间的距离，可由式(2)得到：

$$r_{ij}=\| x_i-x_j \|=\sqrt{\sum_{k=1}^{d}(x_{i,k}-x_{j,k})^2} \tag{2}$$

定义萤火虫的吸引度为：

$$\beta_{ij}=\beta_0\times e^{-\gamma\cdot r_{ij}} \tag{3}$$

其中，β_0 表示最大吸引度，即光源处的吸引度；$\beta_{ij}(r_{ij})$ 为萤火虫 i 对 j 的吸引度。

根据萤火虫会向吸引度高的个体靠近的特性，当萤火虫 i 向萤火虫 j 移动时，定义萤火虫 i 的位置更新公式如下：

$$X_i(t+1)=X_i(t)+\beta_{ij}(r_{ij})(X_i(t)-X_j(t))+\alpha\varepsilon \tag{4}$$

其中，t 表示迭代次数；$X_i(t)$ 和 $X_j(t)$ 表示萤火虫 i 和 j 的位置；α 为步长因子，取值范围为 $0<\alpha<1$；ε 是服从均匀分布的随机因子。

在式(4)中，步长因子 α 影响着种群的收敛性和寻优精度。α 较大时，萤火虫个体间的移动距离大，全局搜索能力强，迭代后期可能造成在最优解附近震荡，导致搜索精度不佳，甚至无法收敛；α 较小时，个体移动距离小，局部搜索能力强，搜索精度高。

3.2　改进萤火虫算法

3.2.1　萤火虫吸引度的改进

在标准 FA 算法中，式(3)吸引度函数的变化只与两萤火虫之间的距离有关。在搜索初期，萤火虫之间的距离较大，个体具有较强的自主性、移动性，以及较好的全局探索能力；随着迭代次数的增加，种群向发光最亮的个体移动，使得两只萤火虫的距离越来越小，最终吸引度维持在 β_0 左右，自主性、移动性减弱，种群快速融合、渐进稳定，寻优值基本确定。此时，如果最优值为局部最优，则由于移动性较弱而无法跳出局部搜索，出现早熟现象。为了降低搜索时出现早熟的概率，使 β 在 $(0,\beta_0)$ 间随机波动，β 在迭代后期依然具有较好的自主性和全局搜索能力。

当今，混沌理论不断发展与完善，已经被运用于算法优化、密码学等领域。混沌的特性主要有随机性、遍历性、分维性、标度律和对初始条件的敏感性[11]。利用混沌序列的随机性可对吸引度 β 进行混沌调制，使其随机波动，提高自主移动性。引入如下混沌调制函数：

$$\theta(t+1)=\sin\left(\frac{\pi}{2}*\theta(t)\right) \tag{5}$$

其中，t 为迭代步数，$t=0$ 时，$\theta(0)=\theta_0$，初始 θ_0 为 $(0,1)$ 间的任意数。θ 经映射之后，在 $(0,1)$ 间随机波动。

将混沌调制函数代入吸引度函数中，使得吸引度随机波动，改进公式如下：

$$\beta_{ij}=\theta\beta_0\times e^{-\gamma\cdot r_{ij}} \tag{6}$$

当 $\theta=0$ 时，萤火虫的吸引度为 0，萤火虫的全局搜索能力最强，萤火虫处在局部最优位置时可跳出局部最优解，向全局最优解靠近。当 $\theta=1$ 时，萤火虫吸引度为实际值，萤火虫向最优值移动。

3.2.2　搜索步长的改进

在传统萤火虫算法中，α 为常数，显然 α 值过大或过小都会影响搜索性能，因此，很多算法提出自适应的搜索步长策略。大部分研究基于迭代次数来自适应地改变搜索步长，即随着迭代步数的增加，逐步减小搜索步长。此类方法取得了较好的效果，但当要求得到的最优解精度较高，即迭代步数增加时，算法的复杂度会随迭代步数的增多而上升。同时，当某个个体在全局最优值附近时，由于迭代初期搜索步长较大，此类自适应方法会使萤火虫在最优解附近震荡飞行，从而影响其他萤火虫的飞行方向。因此，此类方法需要慎重使用。本文从萤火虫个体适应度变化的角度出发，提出了一种自适应地搜索步长的方法，具体公式如下：

$$\alpha(t)=\begin{cases}\alpha_0, & t=1\\ \alpha_0\times\left\|1-\dfrac{F(t)}{F(t-1)}\right\|, & t>1\end{cases} \tag{7}$$

其中，定义迭代适应度函数为 $F(t)$，α_0 为常数步长因子。当迭代次数为 1 时，初始步长为 α_0。从迭代第二步开始，将本次迭代适应度与上一次迭代适应度进行比较，如果本次适应度值较上一次有较大提升，则下次搜索步长增加，以提高全局搜索能力；如果变化较小，表明萤火虫个体进入最优解附近，搜索步长变小，以提高搜索精度。

3.2.3　双种群自适应混沌萤火虫算法

为了保持萤火虫种群的多样性，避免因初始值陷入局部最优，本文对传统单种群算法进行改进，引入了双种群机制，使得种群间协同进化。本文使用全局搜索种群和局部搜索种群，采用不同的位置更新公式(见式(8)、式(9))来保证种群多样性和算法的收敛性。

$$X_i(t+1)=M_{ij}+\beta_{ij}(r_{ij})(M_{ib}(t)-X_i(t))+\alpha\varepsilon \tag{8}$$

$$X_i(t+1)=M_{ij}+\beta_{ij}(r_{ij})(M_{jb}(t)-X_i(t))+\alpha\varepsilon \tag{9}$$

其中，$M_{ij}(t)=\dfrac{X_i(t)+X_j(t)}{2}$，$M_{ib}(t)=\dfrac{X_i(t)+X_{\text{best}}(t)}{2}$，$M_{jb}(t)=\dfrac{X_j(t)+X_{\text{best}}(t)}{2}$。$X_{\text{best}}$ 表示种群中最佳萤火虫的位置。若萤火虫 i 向萤火虫 j 移动，表明 j 的位置优于 i，M_{ij} 保存了 i 和 j 个体的有益位置信息，所以 M_{ij} 优于 i 的位置信息；M_{ib} 保存了个体 i 和最优个体的有益位置信息，M_{ib} 优于个体 i；M_{jb} 保存了个体 j 和最优个体的有益位置信息，主要是为了加大个体 j 对个体 i 的优势，使个体 i 向最优解个体靠近。

定义种群 P_1 和 P_2 具有相同规模，P_1 采用式(8)的位置更新算法，并采用较大的初始搜索步长，进行全局搜索；种群 P_2 采用式(9)的位置更新算法，采用较小搜索步长，进行局部搜索。在搜索过程中，P_1 种群对 P_2 种群的寻优起指导作用，当种群 P_1 得到全局最优解时，将最优位置信息更新到 P_2，使得种群 P_2 向 P_1 的最优解附近移动并进行局部搜索，以求得更精确的最优解。

算法的步骤如下：

1)设待优化的函数为 $f(X)$，初始化种群 P_1 和 P_2 的规模，设置最大迭代步数为 T，搜索步长分别设为 α_1 和 α_2，初始化 β_0，γ，θ_0 等参数，初始迭代步数 $t=1$；

2)对两个种群中的吸引度分别进行 θ 混沌映射,得到混沌吸引度值;

3)根据混沌吸引度值,种群内萤火虫开始移动,更新位置信息和亮度信息,并获取 P_1 和 P_2 种群最优适应度($f(X)$的最优值)时的位置 X_{b1} 和 X_{b2},根据适应度值更新搜索步长 α_1 和 α_2;

4)如果 $f(X_{b1})>f(X_{b2})$,使种群 P_2 向 P_1 移动,令 $X_{b1}=X_{b2}$;

5)$t=t+1$,如果 $t>T$,进入步骤 6),否则进入步骤 2);

6)输出最优适应度时的位置 X_{b2}。

4 支持向量机的原理

支持向量机是由 Vapnik 等提出的一种机器学习方法[12],它是建立在统计学习理论的 VC 维(Vapnik Chervonenkis Dimension)和结构风险最小化原理的基础上的学习方法[13]。其原理是通过线性变换从线性可分转换到线性不可分,同时还可以扩展到非线性函数中。

若样本线性可分,则存在一个超平面:

$$y=w^{\mathrm{T}}\cdot x+b \tag{10}$$

该超平面能够正确划分$(w^{\mathrm{T}}\cdot x)+b=k$ 和$(w^{\mathrm{T}}\cdot x)+b=-k$。令 $w^{\mathrm{T}}=\frac{w^{\mathrm{T}}}{k}$,$b=\frac{b}{k}$,这两条直线可以表示为:

$$(w^{\mathrm{T}}\cdot x)+b=1$$

$$(w^{\mathrm{T}}\cdot x)+b=-1$$

两条直线的间隔为 $\frac{2}{\|w\|}$,对于 $y_i=1$ 的样本有$(w^{\mathrm{T}}\cdot x_i)+b\geqslant 1$,对于 $y_i=-1$ 的样本有$(w^{\mathrm{T}}\cdot x_i)+b\leqslant -1$,则 $y_i\cdot((w^{\mathrm{T}}\cdot x_i)+b)\geqslant 1$,$(w^{\mathrm{T}}\cdot x_i)+b\geqslant 1$。因此,该最优化问题转为线性规划问题:

$$\max \frac{2}{\|w\|} \quad \text{s.t. } y_i(w^{\mathrm{T}}\cdot x_i+b)\geqslant 1 \tag{11}$$

将式(11)转为 $\min \frac{1}{2}\|w\|^2$,引入拉格朗日乘子来求解问题,得到函数:

$$L=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i((y_i(w^{\mathrm{T}}\cdot x_i)+b)-1) \tag{12}$$

为了让 L 关于 w,b 达到最小化,首先设 α 为常数,对 w 和 b 求导数:

$$\frac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{l}\alpha_i y_i x_i \tag{13}$$

代入 L 得到:

$$L=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j x_i^{\mathrm{T}}x_j \tag{14}$$

则可求解 $\alpha=(\alpha_1,\cdots,\alpha_l)^{\mathrm{T}}$ 值,将其代入式(12)计算出 w 值。通过计算 $b=y_j-\sum_{i=1}^{l}\alpha_i y_i(x_i\cdot x_j)$ 的值,构造一个超平面,用于得到决策函数 $f(x)=\mathrm{sgn}(g(x))$,其中:

$$g(x)=\sum_{i=1}^{l}y_i\alpha_i(x_i\cdot x)+b \tag{15}$$

5 短时客流预测模型

时间序列来自特定观测对象按时间排序的观测值序列。文献[9]根据客流历史时间序列的相关性建立了基因遗传算法与 BP 神经网络(GA_BP)组合模型,其根据连续历史时间客流序列来预测下一时段的客流,取得了良好的预测效果。本文以南京地铁三号线为例,将 2016 年 9 月 S_1 站的进站客流时间序列的作为研究对象,建立客流预测模型。

5.1 客流时间序列的相关性分析

常用的客流时间序列预测方法主要利用历史时间片序列来预测未来的时间片,因此有必要分析客流时间片序列间的相关性。使用 Spearman 相关系数 R[4] 对一天的客流序列样本进行相关性计算,其中相关系数的取值范围为[-1,1],绝对值越接近于 1 则相关性越强。计算结果如表 3 所列。

表 3 不同类别的客流时间片的相关性系数 R

Table 3 Correlation coefficient R of different types of passenger flow time slices

前 n 时间片	A 类 R 值	B 类 R 值	C 类 R 值
1	0.9652	0.9431	0.9832
2	0.9141	0.8924	0.9532
3	0.8642	0.8575	0.9389
4	0.8538	0.8101	0.8852
5	0.8283	0.7824	0.8682
6	0.7925	0.7521	0.8214
7	0.7648	0.6417	0.8049
8	0.7471	0.6134	0.7821

通常定义相关系数大于 0.8 为强相关性。从相关性分析结果可以看出,A 类客流前 5 时间片具有强相关性,B 类客流前 4 时间片具有强相关性,C 类客流前 7 时间片具有强相关性,这表明了选择历史若干时间片的合理性。

5.2 预测模型的建立

根据客流的聚类结果,对不同类别的客流时间序列分别进行建模预测,从图 4(a)的客流波形中可以看出,在 S_1 站,将 14 号(节假日的前一天)客流数据归为 B 类;15-17 号为节假日(周末和中秋节调休),与周末客流波形相似,归为 C 类;18 号虽为周日,但属于正常工作日,且为假后第一天上班,客流波形与周一至周四客流波形相似,因此归为 A 类。本文根据时间序列的相关性,以 A 类客流为例,其他类客流模型的建立相同。以 A 类客流为例,利用具有高相关性的历史前 5 个时间片客流来预测未来时间片客流,预测模型的计算公式如下:

$$x(k+1)=f[x(k),x(k-1),\cdots,x(k-4)] \tag{16}$$

利用 SVM 在解决非线性问题上的优势,对客流进行非线性回归预测。在传统的基于 SVM 的客流预测中,由于输入客流数据较多,导致训练时间长且数据噪声多,预测效果不佳。经过客流分类和时间序列简短化的处理,训练规模得到减少。

其中,9 月份的 A 类客流量如图 7 所示,共 16 天(即 1 号、5-8 号、12-13 号、18-22 号、26-29 号)。为了使模型具有较好的鲁棒性,将训练集和测试集按3∶1进行分类,选择 5∶00-24∶00 的时间片进行训练和预测。将头尾数据处理后,模型的输入、输出矩阵如下:

$$\boldsymbol{input}=\begin{Bmatrix} x_1 & x_2 & \cdots & x_n \\ x_2 & x_3 & \cdots & x_{n+1} \\ \cdots & \cdots & \cdots & \cdots \\ x_m & x_{m+1} & \cdots & x_{n+m-1} \end{Bmatrix} \tag{17}$$

$$output=\begin{Bmatrix} x_{n+1} \\ x_{n+2} \\ \vdots \\ x_{m+n-1} \end{Bmatrix} \tag{18}$$

其中，训练集中共有 1224 个时间片（即前 12 天），$m=1224$。由于客流的提前性和滞后性，地铁6:00开始运行，6:00前已经有客流在地铁等待，而在23:00以后车站客流只出不进，预测进站客流无意义，但预测出站客流有一定价值，因此测试集选择6:00－23:00来预测进站客流更符合实际需要，故测试集选择后 4 天的 408 个时间片，$m=408$。根据对相关性的研究，选择相关性高的前 5 个时间片进行预测训练，$n=5$。

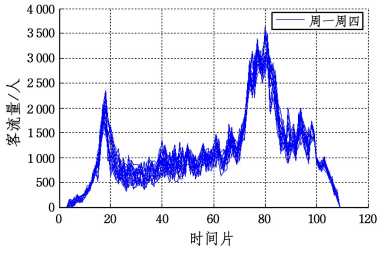


图 7　9 月份周一至周四的客流

Fig. 7　Passenger flow from Monday to Thursday in September

对于训练样本 $S=(input, output)$，在选定训练集合测试后，需要对数据进行归一化处理，使数据分布在[1,2]范围内，以利于数据的处理，加快算法收敛。在使用 SVM 训练前须选择核函数及其参数，实验使用改进萤火虫寻优算法对 RBF 核函数的参数进行优化，得到最优参数。设置完 RBF 核函数的参数后，利用 LIBSVM[16] 库在 Matlab 下进行训练得到回归模型。模型建立流程如图 8 所示。

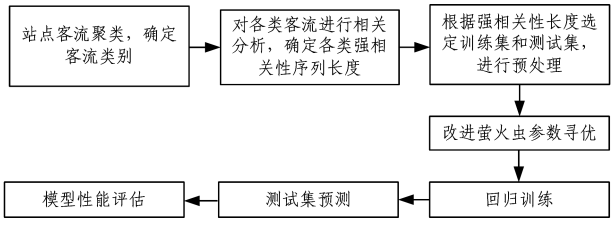


图 8　预测模型流程

Fig. 8　Predictive model flow

5.3　预测性能评估标准

对所提预测模型建立相应的评估标准，以反映模型的性能。本文选择均方根误差 RMSE、平均绝对百分误差 MAPE 作为评估标准。

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-y_i')^2} \tag{19}$$

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\frac{|y_i-y_i|}{y_i'} \tag{20}$$

其中，N 表示测试集中客流的长度，y 表述预测客流，y' 表示实际客流。RMSE 反映了预测的偏差程度，对特大或特小误差非常敏感；MAPE 体现了预测精度，该值越小精度越高。为了得到精度较高的解，将萤火虫算法中的适应度函数选择为 $\frac{1}{MAPE}$。

6　实验结果与分析

在 Matlab 仿真平台上安装 LIBSVM[16] 工具箱，使用工具箱的训练函数对训练集进行训练。SVM 模型中的寻优参数主要有：惩罚系数 c，核函数参数 g，epsilon-SVR 中的不敏感损失函数 p。

c 越大，说明越不能容忍出现误差，越容易过拟合。c 越小，越容易欠拟合。因此，c 过大或过小都会导致泛化能力变差。g 是选择 RBF 函数作为 kernel 后，该函数自带的一个参数，其隐含地决定了数据映射到新特征空间后的分布。G 值越大，支持向量越少；g 值越小，支持向量越多。支持向量的个数将影响训练与预测的速度。p 是回归值与真实值之间允许的最大误差，其大小将影响支持向量的数目，p 越大，支持向量的数量越少。

根据萤火虫算法的定义，每个萤火虫的位置满足 $x_i=\{c, g, p\}$，且属性满足 $c\in[0,1000]$，$g\in[0.1,100]$，$p\in[0.0001, 0.1]$。初始时最大迭代次数为 $T=200$。种群 P_1 和 P_2 的规模均为 40，$\beta_0=1$，$\gamma=1$，$\theta_0=0.5$，$\alpha_1=0.1$，$\alpha_2=0.01$。

为了测试改进萤火虫算法的性能，在迭代过程中记录每次迭代后的吸引度值和 FA 算法的收敛性，这里收敛性取萤火虫个体到最佳个体的平均距离。

从图 9、图 10 可以看出，加入混沌机制的萤火虫吸引度值随机波动，初期经过探索后算法在迭代 50 次时收敛，较未改进算法减少了约 30 次迭代，因此可以有效减少种群的迭代次数，降低算法的时间复杂度。算法收敛后吸引度依旧随机波动，因此可以保持较强个体的搜索能力，降低早熟的概率。与图 10 相比，图 11 中算法加入了自适应搜索步长，具有更快的收敛速度，迭代步数得到进一步优化。

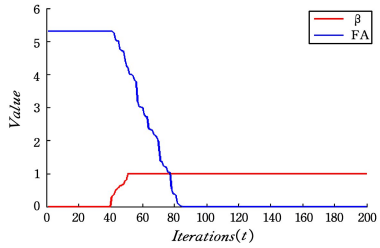


图 9　β 和 FA 收敛性随搜索迭代次数的关系

Fig. 9　Relationship between β and FA convergence and search iterations

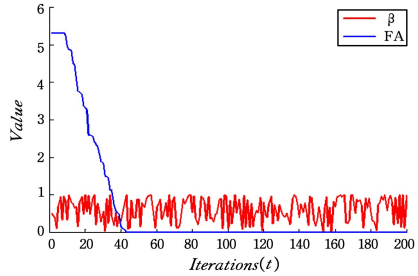


图 10　混沌机制 β 和 FA 收敛性与搜索迭代次数的关系

Fig. 10　Relationship between chaos β and FA convergence and search iterations

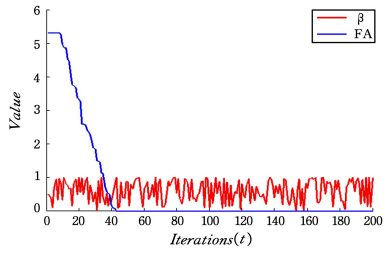


图 11 DACFA 和 FA 收敛性与搜索迭代步数的关系

Fig. 11 Relationship between DACFA and FA convergence and search iterations steps

为了体现模型的性能,将本模型与 ARIMA(自回归移动平均模型)[17]、BP 预测模型[18]、GA_BP 组合模型[4]进行了对比实验,结果如表 4 所列。

同时,为体现改进算法的优势,在 A 类客流预测模型中使用 GA(基因遗传算法)、PSO(粒子群算法)、FA(标准萤火虫算法)和 DACFA 对 SVM 进行优化对比。实验选取测试集的前 50 个时间片进行预测,预测结果如图 12、图 13 所示。从表 4 和图 12、图 13 中可以看出:

1)非线性回归预测方法较传统线性回归方法 ARIMA 具有更高的预测精度,*RMSE* 降幅约为 80 乘次,*MAPE* 改善了 10%左右;较神经网络模型,*RMSE* 降幅约为 8～40 乘次,*MAPE* 提升 2%～8%左右。

2)改进 DACFA 与 GA、PSO 相比,*RMSE* 值降幅为 25～17 乘次,*MAPE* 降低了 3%～6%,使得在预测精度和稳定性上都有较大提升,与真实值较为贴近。

3)加入混沌机制后,*RMSE* 值降幅为 15 乘次,萤火虫算法的预测稳定性明显得到提升;*MAPE* 降低了 6%,预测精度稍有改善。

4)图 12、图 13 表明本模型较其他模型有更好的预测稳定性,与真实值较为贴近。

表 4 地铁客流预测模型的对比结果

Table 4 Comprison of results of subway passenger flow prediction models

预测模型	A 类客流		B 类客流		C 类客流	
	RMSE	*MAPE*	*RMSE*	*MAPE*	*RMSE*	*MAPE*
ARIMA	142.3214	0.2532	135.8471	0.2571	122.3214	0.2375
BP	96.8631	0.1582	97.7832	0.1632	91.2367	0.1512
GA-SVM	83.7291	0.1365	85.7353	0.1395	82.9632	0.1235
PSO-SVM	74.6521	0.1075	71.3561	0.1123	72.8531	0.1169
FA-SVM	71.5628	0.0821	68.3829	0.0792	69.8622	0.0771
GA_BP 组合模型	65.4329	0.0747	63.8642	0.0751	63.7521	0.0738
DACFA-SVM	57.9646	0.0724	53.7143	0.0727	53.9821	0.0715

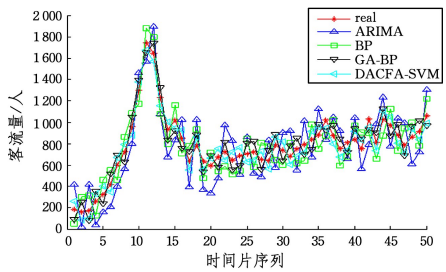


图 12 DACFA-SVM 与其他模型的预测效果对比

Fig. 12 Comparison of prediction results of DACFA-SVM and other models

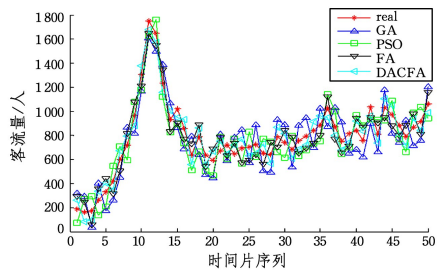


图 13 DACFA 与其他优化算法的预测效果对比

Fig. 13 Comparison of DACFA and other optimization algorithms

结束语 客流表现为时间上的周期变化,一周客流与上一周客流变化具有相似性,且一周内客流波形存在相似性;同时,在空间上,不同站点在一周内存在不同的客流波形特征。因此,本文通过分析客流的时空特性,将凝聚层次聚类算法对不同站点一周的客流进行聚类分析,聚类系谱图表明,不同站点分类出的客流波形存在一定的差异性。现有研究中将一周客流划分为固定类别,从而造成模型的普适性较差,因此本文在不同站点根据聚类系谱图进行不同分类,然后对每类客流分别建模,并建立组合模型,使得预测模型具有较好的普适性和鲁棒性。在一天的短时客流时间序列中,时间片间存在一定的相关性,通过相关性分析,本文提出一种基于 SVM 的回归预测模型,将强相关性时间序列作为模型输入。相较于传统方法,由于对客流进行分类建模,本模型有效地减小了 SVM 的训练规模并减少了训练时间。SVM 模型的性能除了取决于数据规模外,还依靠核函数及其参数的选择。为了更好地选择参数,本文还提出一种双种群自适应混沌萤火虫算法进行参数寻优,相较于其他常用优化算法,其具有更好的预测精度和稳定性。相比于 ARIMA,BP,GA_BP 组合模型等,本模型具有明显优势。本模型在预测选择下一时刻客流时需要历史时间片,若要预测下一连续时间片客流,则会存在累计误差问题,从而影响预测精度。今后将对此问题进行进一步优化;同时,考虑天气、突发事件等因素对客流的影响,以提高预测精度和稳定性。

参 考 文 献

[1] WANG P,WU C,GAO X. Research on subway passenger flow combination prediction model based on RBF neural networks and LSSVM[C]// Control and Decision Conference. Las Vegas: IEEE Press,2016:6064-6068.

[2] AMALIAH B, ZEINITA A, SURYANI E. Dynamics simulation of air passenger forecasting and passenger terminal capacity expansion scenario in Yogyakarta Airport[C]// International Conference on Information & Communication Technology and Systems. Surabaya: IEEE Press, 2017: 187-192.

[3] ESCOLANO C O, DADIOS E P, FILLONE A D. Fuzzy logic controlled adaptive scheduling of public utility buses in Metro Manila[C]// International Conference on Humanoid, Nanotechnology, Information Technology, communication and Control, Environment and Management. Cebu City: IEEE, 2016: 1-5.

[4] DONG S W. Research on short-term passenger flow forecasting method based on improved BP neural network[D]. Beijing: Beijing Jiaotong University, 2013. (in Chinese)
董升伟. 基于改进 BP 神经网络的轨道交通短时客流预测方法研究[D]. 北京：北京交通大学，2013.

[5] YANG X F, LIU L F. Short-time passenger flow forecasting based on AP clustering for bus stations in support vector[J]. 2016, 40(1): 36-40. (in Chinese)
杨信丰，刘兰芬. 基于 AP 聚类的支持向量机公交站点短时客流预测[J]. 武汉理工大学学报(交通科学与工程版)，2016，40(1)：36-40.

[6] LERSPALUNGSANTI S, ALBERS A, OTT S, et al. Human ride comfort prediction of drive train using modeling method based on artificial neural networks[J]. International Journal of Automotive Technology, 2015, 16(1): 153-166.

[7] DOU Y, XIAO Z, XIE Y. Research on Hotspot Short-Term Passenger Flow Forecasting Based on Neural Network[C]// Fifth International Conference on Multimedia Information NETWORKING and Security. Beijing: IEEE Computer Society, 2013: 332-335.

[8] SHARMA A, ZAIDI A, SINGH R, et al. Optimization of SVM classifier using Firefly algorithm[C]// IEEE Second International Conference on Image Information Processing. Paris: IEEE, 2014: 198-202.

[9] JIANG G Y, KONG C L. Traffic Parameters Prediction Method Based on Rolling Time Series[J]. Advanced Materials Research, 2013, 54(6): 2946-2950.

[10] LU K Z, ZHANG Z Q, SUN J. Improved FA algorithm for maintaining individual activity [J]. Journal of University of Science and Technology of China, 2016, 32(2): 120-129. (in Chinese)
陆克中，章哲庆，孙俊. 保持个体活性的改进 FA 算法[J]. 中国科学技术大学学报，2016，32(2)：120-129.

[11] LI W, GE J, DAI G. Detecting Malware for Android Platform: An SVM-Based Approach[C]// IEEE, International Conference on Cyber Security and Cloud Computing. Beijing: IEEE Press, 2016: 464-469.

[12] FLEURY A, VACHER M, NOURY N. SVM-Based Multimodal Classification of Activities of Daily Living in Health Smart Homes: Sensors, Algorithms, and First Experimental Results [J]. IEEE Transactions on Information Technology in Biomedicine A Publication of the IEEE Engineering in Medicine & Biology Society, 2010, 14(2): 274-283.

[13] YILDIZ O T. VC-Dimension of Univariate Decision Trees[J]. IEEE Transactions on Neural Networks & Learning Systems, 2015, 26(2): 378-387.

[14] FENG C, TAGUCHI Y, KAMAT V R. Fast plane extraction in organized point clouds using agglomerative hierarchical clustering[C]// IEEE International Conference on Robotics and Automation. Hong Kong: IEEE Press, 2014: 6218-6225.

[15] ALFRED R, TAN S F, TAHIR A, et al. Concepts Labeling of Document Clusters Using a Hierarchical Agglomerative Clustering (HAC) Technique[M]// The 8th International Conference on Knowledge Management in Organizations. Berlin: Springer Netherlands, 2014: 263-272.

[16] SANTAMARIA-BONFIL G, REYES-BALLESTEROS A, GERSHENSON C. Wind Speed Forecasting For Wind Farms: A Method Based on Support Vector Regression[J]. Renewable Energy, 2016, 85(6): 790-809.

[17] TSEKERIS T, STATHOPOULOS A. Short-Term Prediction of Urban Traffic Variability: Stochastic Volatility Modeling Approach[J]. Journal of Transportation Engineering, 2010, 136(7): 606-613.

[18] YANG W J. Research on Forecast of Railway Passenger Volume Based on BP Neural Network [J]. Cooperative Economy & Technology, 2010, 34(13): 18-19. (in Chinese)
杨伟静. 基于 BP 神经网络的铁路客流量预测研究[J]. 合作经济与科技，2010，34(13)：18-19.