控制与决策
Control and Decision

基于MI-SVR模型的航空旅客出行指数预测方法研究

熊红林， 朱人杰， 冀和， 樊重俊， 徐佩

引用本文:
熊红林, 朱人杰, 冀和, 等. 基于MI-SVR模型的航空旅客出行指数预测方法研究[J]. 控制与决策, 2021, 36(7): 1619-1626.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1446

基于MI-SVR模型的航空旅客出行指数预测方法研究

熊红林[1], 朱人杰[1,2], 冀 和[1], 樊重俊[1†], 徐 佩[1]

(1. 上海理工大学 管理学院，上海 200093；2. 同济大学附属东方医院 运营管理部，上海 200120)

摘 要: 航空旅客出行的情况对民用航空机场建设与运营具有重大意义. 针对航空旅客出行情况的预测研究,首先定义一种航空旅客出行指数,通过K-means聚类方法对航空旅客出行指数进行分级;然后基于互信息与相关性原理,选取航空旅客出行情况关键影响特征因子,提出一种基于关键影响因子与航空旅客出行指数互信息的MI-SVR (mutual information-support vector regression)机器学习预测模型;最后通过上海机场旅客出行指数预测实验对模型进行验证,实验结果显示MI-SVR模型具有可行性与有效性,同时,相比传统的预测模型预测效果更优. 此外,实验结果也表明,相对仅基于历史数据进行独立预测,各模型基于互信息引入影响因子进行预测误差更小,研究结果有助于提升机场建设及运营管理水平,同时也可辅助人们选择通过民航交通方式出行的时段.

关键词: 机场运营管理；航空旅客出行指数；机器学习；互信息；支持向量回归；K-均值聚类

中图分类号: TP181 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1446 **开放科学(资源服务)标识码(OSID):**

引用格式: 熊红林,朱人杰,冀和,等. 基于MI-SVR模型的航空旅客出行指数预测方法研究[J]. 控制与决策, 2021, 36(7): 1619-1626.

Air passenger index prediction method based on MI-SVR mode

XIONG Hong-lin[1], *ZHU Ren-jie*[1,2], *JI He*[1], *FAN Chong-jun*[1†], *XU Pei*[1]

(1. Business School，University of Shanghai for Science and Technology，Shanghai 200093，China；2. Operation Management Department, East Hospital Affiliated to Tongji University，Shanghai 200120，China)

Abstract: Air passenger travel is of great significance to the construction and operation of civil aviation airports. This paper studies the prediction of air passenger index, the main research work is as follows: Firstly, the air passenger index is defined, and the air passenger index is classified using the K-means clustering method. Then, based on the principle of mutual information and correlation, the key influencing factors of air passenger index are selected. This work presents a mutual information-support vector regression (MI-SVR) machine learning model based on mutual information between the key influencing factors and the air passenger index, which is used to predict the air passenger index. Finally, the model is validated by passengers throughput data of the Shanghai airport. The experimental results show that the MI-SVR model is feasible and effective, and compared with classical models, the IM-SVR model has better prediction effect. In addition, it is also showed that the prediction effect of each model is better after introducing influence factors based on mutual information. Overall, the study is helpful to the construction and operation of airports, and it can also help people choose the time to travel by air.

Keywords: airport operation and management；air passenger index；machine learning；mutual information；support vector regression；K-means

0 引 言

自人工智能基本理论诞生以来,人工智能技术与应用得到长足的发展. 机器学习是人工智能的重要表现形式,其经历了从浅层机器学习到深度学习两次浪潮[1],随着机器学习理论及应用的不断发展,各种浅层机器学习模型相继被提出,典型机器学习模型之一是Cortes等[2]发明的支持向量机. 机器学习作为目前人工智能的重点研究领域之一,应用于众多领域,包括语音处理、计算机视觉、自然语言处理等[3-6]. 机器学习在回归预测方面也取得了不错的效果,近年来,国内外学者通过运用机器学习模型对机场旅客吞吐量的预测研究取得了一定的成果[7],如支持向量回归[8]、神经网络[9]也都取得了很好的预测精度. 目前,针对机场旅客吞吐量的主要预测方法包含两个方面,即线性预测与非线性预测. 基于线性的方法包括时间序列模型(ARIMA)、灰色模型(GM)[10-14]等,这类

收稿日期: 2019-10-14；**修回日期:** 2020-03-01.

基金项目: 国家自然科学基金项目(71774111)；上海市教育委员会科研创新重点基金项目(14ZZ131).

†通讯作者. E-mail: fan.chongjun@163.com.

方法虽然已取得较好的预测结果,但不能对非线性趋势进行反映,预测精度有待提高;基于非线性的方法包括BP-神经网络[9]、循环神经网络(RNN)、长短期记忆网络(LSTM)、支持向量回归(SVR)[15-21]等模型,这类模型可以对输入与输出之间的非线性关系进行拟合,具有较强的容错能力,也是目前机场旅客吞吐量预测常用的模型. 主要的机场旅客吞吐量预测技术都是建立于具有良好的统计规律性的西方成熟的航空业基础之上[8-14, 22].

机器学习方法在航空旅客运输方面的相关研究越来越深入. 在航线流量方面,从灰色预测模型在航线客流预测[23]到经典神经网络模型与其他组合模型在机场客运与货运的混合预测[24-25]进行了相关探索. 在区域航空客流量方面,以将神经网络与支持向量机结合的方式[3],对区域航空市场趋势、长短期客流量预测并获得了较好的效果;文献[26]对支线航空公司在现有航路网络中增加枢纽的客流影响进行了探讨. 在航空流量特征研究方面,基于时间序列[5, 12, 27-28]与非线性向量自回归神经网络(MIV-NVARNN)[29]对航空客流量特征进行了分析预测. 在客运量影响因素方面,文献[30]采用灰色综合关联方式对影响航空客运量的因素进行分析,找出了影响航空客运量的主要因素,并以此为基础建立多元回归模型. 综上,在航空旅客相关预测方法的研究工作上,近些年国内外很多学者主要致力于提高模型预测的精度,主要从两个方面进行研究,一方面是对单个模型的改进,另一方面是将多个单一模型进行组合以获得更优的效果.

互信息不仅能够反映各变量间的相关性,而且能够表征变量的非线性相关,互信息用于相关性分析得到广泛应用,基于互信息的特征选择[31-32]与机器学习领域[33]的优异性得到广泛体现. 与以往的研究不同,本文不仅聚焦于民航机场旅客吞吐量,而是以旅客吞吐量为基础数据,通过对其进行相应处理达到对航空旅客出行情况的直观体现,同时也为民航机场运营服务提供辅助决策. 此外,以往的方法局限于对历史数据的关注,并未考虑影响航空旅客出行的其他关联因素. 本文在其他学者的研究基础上,引入信息论知识并结合SVR方法,通过计算互信息选取关键影响特征因子构建机器学习预测模型,以寻求更优的预测方法,为旅客出行时间的选择提供参考依据.

1 问题描述

民用航空机场的3大指标(旅客吞吐量、邮货吞吐量、飞机起落架次)的变化规律一直深受国内外研究学者的关注,由于民用航空机场旅客吞吐量的日流量数据较大,并且不同的机场差异更大,仅仅用机场旅客吞吐量来描述机场服务能力,在面向机场运营与公众选择航空出行方式决策方面提供决策支持并不直观. 由此可见,将旅客吞吐量原始数据转化处理为民航机场旅客出行指数的意义不言而喻. 航空旅客出行指数不仅能够体现机场容纳旅客的规模也反映出旅客量的变化趋势,且清晰明了便于理解. 此处先对相关定义进行说明以便接下来进行研究分析.

定义1 设X_t为某单位时间的机场旅客吞吐量数值,则该单位时间内的航空旅客出行指数X_t^*为

$$X_t^* = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}}. \tag{1}$$

其中:$X_{\min}$为一年中旅客量最小值,$X_{\max}$为一年中旅客量最大值. 由式(1)可以发现,X_t^*的取值范围在$[0, 1]$之间,很直观地反映了不同民用航空机场旅客出行指数(文中亦简称“出行指数”).

定义2 设$\{p_1, p_2, \dots, p_t\}$为多个时间单位的航空旅客出行指数序列集,对其聚类后生成的簇为一组数据对象的集合$\{N_t\}$,其中N_t取值为聚类后每一簇的数据元素. 若航空旅客出行指数时间的单位为月,则当机场该月航空旅客出行指数为X_t^*时,其航空旅客出行指数等级记为

$$N_{X_t^*} = \begin{bmatrix} 1, p_t \in (0, i) \\ 2, p_t \in (i, j) \\ 3, p_t \in (j, k) \\ \vdots \\ N, p_t \in (\theta, 1) \end{bmatrix}, \tag{2}$$

其中i、j、k、θ为聚类族边界值.

民用航空机场旅客的出行指数对机场运营管理与旅客出行服务具有重要意义,本文尝试借助互信息理论与SVR方法构建机器学习模型对航空旅客出行指数进行预测,并对结果进行验证说明.

2 相关理论与模型构建

2.1 互信息理论

信息熵是信息论中的一个重要指标,系统越有序,信息熵越小;相反地,系统越混乱,信息熵越大. 因此,信息熵可以作为系统不确定性程度(或者说有序化程度)的度量标准[34-37]. 信息熵可用如下公式表示:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i). \tag{3}$$

其中:$P(x_i)$为样本x_i的概率,n为样本数目. 由此可以看出,某个事件出现的概率越小,即信息的不确定性越大,熵值越高. 设随机向量(X, Y)的联合概率分布为p_{ij},则其二维联合熵为

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_{ij})\log P(x_{ij}). \quad (4)$$

假设X和Y的边缘概率分布分别为$p_{i.}$和$p_{.j}$,可定义条件熵为

$$H(X/Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_{ij})\log\frac{p_{ij}}{p_{.j}}, \quad (5)$$

$$H(Y/X)=-\sum_{i=1}^{n}\sum_{j=1}^{m}P(x_{ij})\log\frac{p_{i.}}{p_{.j}}. \quad (6)$$

因此,互信息可以表示为对于变量X(或Y),由于变量Y(或X)的发生导致其不确定性减少的熵值.

$$\begin{aligned}I(X;Y)&=H(X)-H(X/Y)=\\&H(Y)-H(Y/X)=\\&H(X)+H(Y)-H(X,Y). \quad (7)\end{aligned}$$

综合式(3)～(7)可以推出互信息的完整表达公式为

$$I(X;Y)=\sum_{i,j}p_{ij}\log_2\frac{p_{ij}}{p_{i.}p_{.j}}. \quad (8)$$

$I(X;Y)$值域为[0,1],其值的大小代表变量X与变量Y依赖关系的强弱,0代表两随机变量完全无关,1代表完全相关.

2.2　支持向量回归

支持向量回归(support vector regression, SVR)是机器学习中较为常用的一种预测方法. SVR预测采用的是最小化结构风险原则而非经验风险最小化,有效地克服了“维数灾难”和传统的模式识别等诸多问题[38-45]. 一般的线性回归模型为

$$f(x)=w^{\mathrm{T}}x+b. \quad (9)$$

其中:w表示旅客出行指数输入向量的法向量,b表示偏差值. 只有当$f(x)$与真实值完全一致时,损失才为0. 而在实际的航空旅客出行指数预测中,精确预测到每天的准确值是任何模型都不可能实现的. 但SVR模型通过“软化”预测结果,允许预测值与实际值存在一定的误差ε,等价于以预测值$f(x)$为中心,形成一条宽度为$2\times\varepsilon$的预测误差隔离带,落入该隔离带内的航空旅客出行指数值即为预测正确,损失为0,同时离隔离带最近的旅客出行指数输入向量构成其“支持向量”. 为最小化损失,需最大化两组支持向量与预测中心的距离之和Υ,这可通过最小化法向量w的欧几里得范数实现,则SVR问题可转化为

$$\min\frac{1}{2}w^2+C\sum_{i=1}^{m}l_\epsilon(f(x_i)-y_i),\ C>0. \quad (10)$$

其中:i为输入的第i组输入变量,y_i为实际目标值,$f(x_i)$为SVR模型拟合值;C为正则化常数,用于对前后两项进行折中计算,前项表示在模型结构上尽可能使得所有预测值都落入误差范围内,后项应用ε-不敏感损失函数l_ϵ来刻画模型预测效果与实际的旅客量数据的契合程度.

$$l_\epsilon(z)=\begin{cases}0,\ |z|\leqslant\varepsilon;\\|z|-\varepsilon,\ \text{otherwise.}\end{cases} \quad (11)$$

在实际航空旅客出行指数数据中,可能由于外在原因导致某个值超出正常的趋势,成为离群值. 这种情况下,上述所定义的“硬间隔”不再适用. 因此,在某些与实际值偏离较严重的情况下,引入松弛变量ξ_i和ξ_i^*“软化”间隔,问题即可转化为

$$\begin{aligned}&\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}l_\epsilon(\xi_i-\xi_i^*).\\&\text{s.t. }f(x_i)-y_i\leqslant\varepsilon+\xi_i;\\&\quad y_i-f(x_i)\leqslant\varepsilon+\xi_i^*;\\&\quad\xi_i,\xi_i^*\geqslant0,\ i=1,2,\ldots,m. \quad (12)\end{aligned}$$

利用对偶原理并引入拉格朗日乘子α_i、α_i^*,得到SVR数[15]的对偶问题

$$\begin{aligned}&\max_{\alpha,\alpha^*}\sum_{i=1}^{m}y_i(\alpha_i^*-\alpha_i)-\varepsilon(\alpha_i^*-\alpha_i)-\\&\quad\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y_i(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)x_i^{\mathrm{T}}x_j;\\&\text{s.t. }\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)=0,\\&\quad0\leqslant\alpha_i^*,\\&\quad\alpha_i^*\leqslant C. \quad (13)\end{aligned}$$

当航空旅客出行指数预测值落入ε-软间隔带中,α_i和α_i^*才能取非零值,且一个预测值不可能同时落入两个相对的区域,因此α_i和α_i^*中至少有一个为0. 最终,SVR回归预测函数[15]可表示为

$$f(x)=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)x_i^{\mathrm{T}}x+b, \quad (14)$$

$$b=y_i+\varepsilon-\sum_{i=1}^{m}(\alpha_j^*-\alpha_j)x_j^{\mathrm{T}}x_i. \quad (15)$$

对于航空旅客出行指数时序数据而言,其往往呈现出非线性变化趋势,而SVR可通过非线性映射函数$\varphi(x)$将样本映射到高维空间,然后采用核函数$K(x_i,x_j)$替换高维空间的向量内积$\varphi(x_i)\cdot\varphi(x_j)$. 最常用的核函数是高斯径向基核函数(RBF)[15],其表达式为

$$K(x_i,x_j)=\mathrm{e}^{-\frac{\|x_i-x_j\|^2}{2\sigma^2}}. \quad (16)$$

其中:$\sigma>0$为高斯核的带宽,$\mathrm{gamma}=-1/\sigma^2$为高斯径向基核函数参数.

高斯径向基核函数使SVR具有更强的非线性预

测能力,最终得到SVR回归函数[15]为

$$f(x)=\sum_{i=1}^{m}(\alpha_i^*-\alpha_i)K(x,x_i)+b. \tag{17}$$

2.3 模型构建

在基本预处理的数据基础上,将类别性数据转换为相应的信息熵值;然后,为了消除量纲的影响,将所有数据进行标准化. 由于所收集的数据并非所有维度都与航空旅客出行指数相关,通过计算每个影响因子与航空旅客出行指数等级的互信息值,根据互信息值排除影响因子较小者,为构建基于互信息的机器学习模型建立基础,从而实现在维度较小情况下达到较好的预测结果.

本文基于互信息的MI-SVR航空旅客出行指数预测模型构建步骤如下.

step 1: 根据航空旅客出行指数的定义对机场旅客吞吐量数据进行处理得到原始序列X_t.

step 2: 将航空旅客出行影响因子转换为信息熵,然后对转换后的数据以既定单位长度进行标准化操作,得到标准化信息熵.

step 3: 将航空出行旅客量数据以同样的方式进行标准化转换为航空旅客出行指数X_t^*.

step 4: 应用K-means方法将出行指数、最高气温、最低气温、风力、风向分别进行聚类,形成聚类簇.

step 5: 分别计算转换后的标准化信息熵与出行指数等级的互信息值,并对互信息值进行排序.

step 6: 根据互信息值大小与相关性强弱的特点,选择互信息值较高的最高气温、最低气温、风向3个因素,将这些影响因子加入SVR模型,即构建MI-SVR模型对航空旅客出行指数进行预测.

step 7: 选择样本数据,根据本文航空旅客出行指数的定义及互信息理论进行数据处理,并对MI-SVR模型进行训练,根据拟合的效果设定参数,得到训练效果较好的机器学习预测模型.

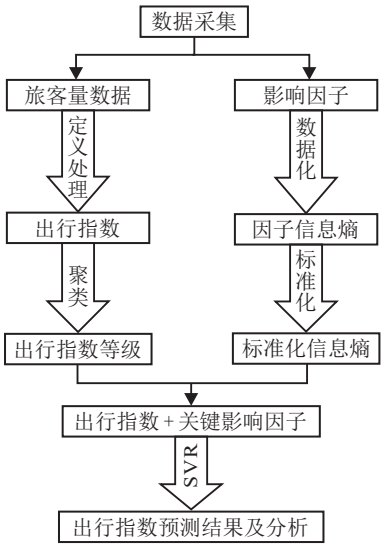

图 1 MI-SVR模型示意图

MI-SVR模型运行流程如图1所示.

3 实证分析

3.1 实证计算

3.1.1 数据准备与处理

根据本文航空旅客出行指数的定义进行数据处理,选取的实验数据集较为完整不存在缺失值,所以直接利用原始数据即可进行计算. 在获得完整的原始数据后,首先将数据集转换为与所构建的模型输入维度一致的类型,即根据前文中对航空旅客出行指数的定义(定义1),将原始的航空旅客吞吐量值转为航空旅客出行指数,本文所选数据集为上海浦东机场从2017年1月1日到2018年8月31日的每天旅客吞吐量、最高气温、最低气温、天气状况、风向、风力的完整原始实验数据. 其中:天气状况、风向、风力为文本型数据,将气温数据进行分级以表示不同的温度级别,数据转换对应表1所示.

表 1 数据预处理

最高气温	最低气温	风力	天气	风向
2～10(冷)	−3～4(寒冷)	1～2级	晴/多云	无持续风向
11～17(微冷)	5～10(冷)	2～3级	阴	东风
18～25(适中)	11～17(凉)	3～4级	小雨	南风
26～32(微热)	18～24(适中)	4～5级	阵雨	西风
33～40(热)	25～31(微热)	5级以上	中雨	北风
			大雪	东南风
			暴雪	西南风
				西南风
				西北风

本文基于信息熵相关理论,对数据进行相应的处理. 首先,将出行指数影响因子5个类别型数据转换为各自对应的信息熵值;然后,对全部旅客吞吐量数据根据式(1)进行计算,并将标准化后的旅客量值作为旅客出行指数运用K-means聚类算法划分等级,分级结果如表2所示;最后,分别计算转换为出行指数等级与5个影响因子的互信息值,得到影响旅客出行指数的重要影响因子依次为最高气温、最低气温、风向、天气状况、风力,这些影响因子与航空旅客出行指数之间的互信息值如表3所示.

表 2 出行指数等级

等级	出行指数范围	说明
1	0～0.13	畅通
2	0.13～0.35	较为畅通
3	0.35～0.52	较为拥堵
4	0.52～0.79	拥堵
5	0.79～1.00	严重拥堵

表 3　多因子与出行指数互信息值及排序

影响因子	与出行指数互信息值	排序
最高气温	0.516	1
最低气温	0.434	2
天气	0.249	4
风向	0.290	3
风力	0.136	5

3.1.2　误差分析方法

为了分析模型的实验结果,本文选取平均绝对百分比误差(MPAE)和均方根误差(RMSE)作为模型误差分析函数,以此评价模型的预测效果,具体公式如下:

$$\mathrm{MPAE}=\frac{\sum_{i=1}^{n}\frac{|y_i-y_i^*|}{y_i}}{n}\times 100\,\%, \tag{18}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-y_i^*)^2}. \tag{19}$$

其中: y_i 为实际值, y_i^* 为预测值.

3.2　结果分析

3.2.1　关键因子选取

实验中,在航空旅客出行指数分别引入最高气温与最低气温单影响因子情况下,SVR模型的RMSE值与MPAE值分别为0.103 0、 11.44 % 和0.101 6、11.18 %. 相比之下,最低气温的影响力略高于最高气温.

从表3可以得出,对旅客出行指数影响最小的是风力,通过对风力因素的统计分析可以发现,风力等级在5级以上的占比仅为2.3 %,而近88 % 的数据其风力级别不高于3级(3～4级风),因此风力对旅客出行指数的影响相对较小.

此外,通过计算发现选取的样本信息中,最高气温与最低气温的Pearson相关系数为0.581,在99 % 的置信区间下,具有显著相关性.

综上,本文最终选择了最低气温、天气、风向3个关键影响因子加入对航空旅客出行指数的预测,以得到精度更优的预测结果.

本文的数据分为两部分: 从2017年1月1日到2018年7月31日的577天的数据用于训练模型,剩余数据集即2018年8月的数据作为测试集用于对模型拟合效果进行验证,通过实验发现一些有意义的结果.

将模型在python3.6软件的基础上运行,经过多次实验,发现当模型的各种具体参数设置在一定的值时,训练集和测试集的整体效果最佳误差最小. 模型相关参数设置如表4所示.

表 4　SVR模型参数设置

函数	gamma值	惩罚系数C	ε
RBF	0.64	0.516	1

3.2.2　对比实验分析

将常用的机器学习模型LSTM与时序模型ARIMA作为对比模型,各个模型分别在有影响因子和无影响因子的条件下实验,对不同模型在测试集上的预测结果进行统计. 为了更加直观地对比预测结果,将6种不同预测模型的实际值与预测值进行可视化处理,结果如图2～图7所示. 在图2中,虚线左侧部分为模型在训练集上的拟合效果,右侧是模型在测试集上的预测结果. 与LSTM和ARIMA模型相比,SVR模型在极限值上表现的更为优异. LSTM和ARIMA只是符合了数据的大致趋势,对较高值与较低值并没有很好的拟合,而MI-SVR模型在较高值和较低值的拟合上较其他两个模型表现更突出,也就是说,在引入出行指数影响因子后构建的模型中,基于互信息选取相关影响因子构建的MI-SVR模型预测效果更好.

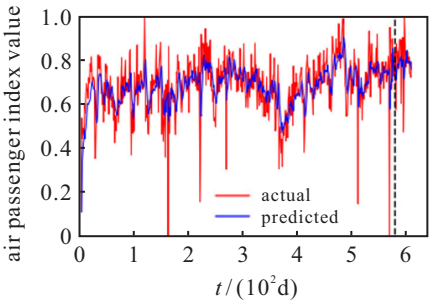

图 2　ARIMA模型实验结果

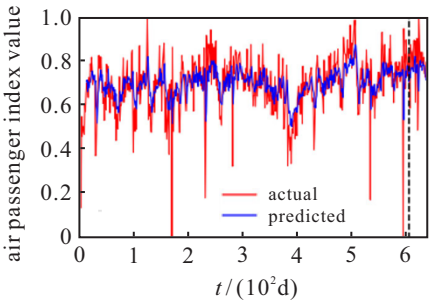

图 3　ARIMA+影响因子模型实验结果

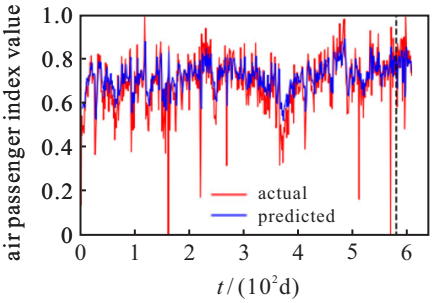

图 4　LSTM模型实验结果

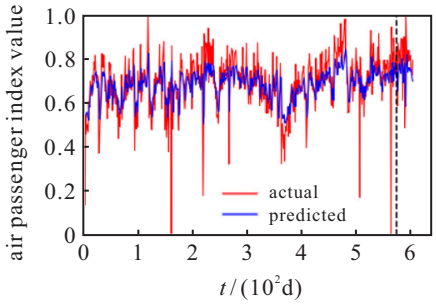

图 5 LSTM+影响因子模型实验结果

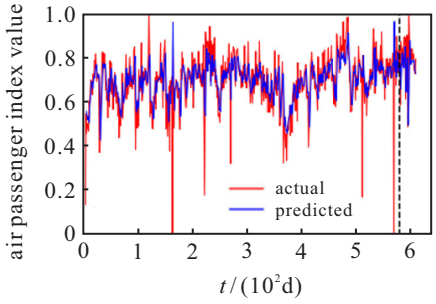

图 6 SVR模型实验结果

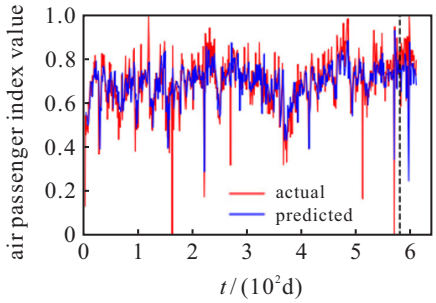

图 7 SVR+影响因子模型实验结果

通过本文3.1.2节误差分析方法计算6个模型的MPAE值与RMSE值,得到不同模型下的误差情况如表5所示. 从表5中分析发现,在增加相关影响因子后的MI-SVR模型进行的实验结果中,RMSE值在0.08以下,且MPAE值也是6种模型中最低的. 由此可以看出,MI-SVR模型对航空旅客出行指数的预测效果相比其他模型优势明显.

表 5 不同模型的RMSE与MPAE对比

模型名称	RMSE	MPAE/%
SVR+影响因子(MI-SVR)	0.078 5	8.04
LSTM+影响因子	0.105 3	11.86
ARIMA+影响因子	0.106 0	11.66
SVR	0.103 1	11.33
LSTM	0.106 4	12.10
ARIMA	0.109 3	12.09

一般而言,组合预测模型较单一预测模型效果要好,为了进一步验证MI-SVR模型在航空旅客出行指数预测问题应用上更优,与其他组合模型预测效果进行对比,选择与文献[18]SVR-ARIMA、文献[19]GM(1,1)-BPNN进行对比分析,结果如表6所示. 通过对比MPAE与RMSE值发现,MI-SVR模型预测拟合度与精度均较优.

表 6 与组合模型的预测效果比较

模型名称	RMSE	MPAE/%
SVR+影响因子(MI-SVR)	0.078 5	8.04
LSTM+影响因子	0.105 3	11.86
ARIMA+影响因子	0.106 0	11.66

4 结 论

航空旅客出行指数的提出为机场运营管理提供了一个新的参考指标. 一方面,机场可以选择在相对不那么繁忙的时间段内安排机场相关设施设备的运维或其他施工活动,这样能最大限度减少此类工作给机场运营管理带来的影响;另一方面,针对旅客出行的高峰时段,充分调配相应的内外部资源,如配套的地面公共交通资源、地勤人员、安保人员等相关资源,最大限度地保障机场的通畅与安全生产. 同时,航空旅客出行指数等级的给出,为广大公众选择出行时机和交通方式提供一定的决策参考.

针对机场航空旅客出行指数预测,本文提出了基于互信息的MI-SVR模型,通过计算各个因子与出行指数的互信息值并比较遴选,选择引入最高气温、最低气温、风向作为模型关键条件影响因子,应用MI-SVR模型对上海机场每天的旅客吞吐量数据进行仿真验证,并将其与其他几种常用的预测模型作对比,结果表明,在引入影响因子情况下,预测准确率明显得到提高,这一结果验证了基于互信息值选取的航空旅客出行指数预测影响因子的科学性与有效性. 同时,对上海浦东机场旅客出行指数进行K-means聚类分析,并给出了科学的出行指数评级. 结果分析发现,上海机场航空旅客出行指数等级情况符合每年人口的出行规律变化,进一步验证了MI-SVR预测模型为机场建设中考虑航空旅客出行指数评估提供了有效的参考方法.

此外,科学的航空旅客出行指数直观反应了各大机场运营的繁忙程度,这一数值指标有助于机场运营管理部门根据有限的服务资源制订相应的服务与管理方案,如机场设施设备的维护日程合理安排、地面服务精准计划、机位分配作出优化决策等,这些机场运营活动对于保障机场绿色通行具有重要意义,将是下一步研究的问题.

参考文献(References)

[1] Kraus M, Feuerriegel S, Oztekin A. Deep learning in business analytics and operations research: Models, applications and managerial implications[J]. European Journal of Operational Research, 2020, 281(3): 628-641.

[2] Cortes C, Vapnik V. Support-vector networks[J]. Machine

Learning, 1995, 20(3): 273-297.

[3] Wason Y. Deep learning: Evolution and expansion[J]. Cognitive Systems Research, 2018, 52: 701-708.

[4] Zou Y, Zhang G, Liu L K. Research on image steganography analysis based on deep learning[J]. Journal of Visual Communication and Image Representation, 2019, 60: 266-275.

[5] Kohavi R, Provost F. Special issue on applications of machine learning and knowledge discovery process[J]. Journal of Machine Learning, 1998, 30(1): 271-274.

[6] Huo Y, Qin R, Xing C, et al. CUDA-based parallel K-means clustering algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery, 2014, 45(11): 47-53.

[7] 焦朋朋. 机场旅客吞吐量的影响激励与预测方法研究[J]. 交通运输系统工程与信息, 2005(1): 107-110.
(Jiao P P. Forecasting method and its mechanism of impacts on airport passenger throughput[J]. Transportation Systems Engineering and Information Technology, 2005(1): 107-110.)

[8] 冯兴杰, 魏新, 黄亚楼. 基于支持向量回归的旅客吞吐量预测研究[J]. 计算机工程, 2005, 31(14): 172-173.
(Feng X J, Wei X, Huang Y L. Predicting the airport passenger throughput based on a support vector regression model[J]. Computer Engineering, 2005, 31(14): 172-173.)

[9] Sun S, Lu H, Tsui K L, et al. Nonlinear vector auto-regression neural network for forecasting air passenger flow[J]. Journal of Air Transport Management, 2019, 78(7): 54-62.

[10] 张丽, 闫世锋. Holt-Winters方法与ARIMA模型在中国航空旅客运输量预测中的比较研究[J]. 上海工程技术大学学报, 2006, 20(3): 280-283.
(Zhang L, Yan S F. Comparison of Holt-Winters and ARIMA methods for forecasting charge of china airline passengers[J]. Journal of Shanghai University of Engineering Science, 2006, 20(3): 280-283.)

[11] 姚晏斌, 高金华. 灰色模型GM(1,2)在机场旅客吞吐量预测中的应用[J]. 中国民航飞行学院学报, 2006, 17(4): 12-16.
(Yao Y B，Gao J H. Gray model GM(l,2) application in airport passenger throughput pre-side[J]. Journal of Civil Aviation Flight University of China, 2006, 17(4): 12-16.)

[12] Gummadi R, Edara S R. Analysis of passenger flow prediction of transit buses along a route based on time series[C]. Advances in Intelligent Systems and Computing. Singapore: Springer Singapore, 2018: 31-37.

[13] Claveria O, Torra S. Forecasting tourism demand to Catalonia: Neural networks vs. time series models[J]. Economic Modelling, 2014, 36: 220-228.

[14] AhmadBeygi S, Cohn A, Guan Y H, et al. Analysis of the potential for delay propagation in passenger airline networks[J]. Journal of Air Transport Management, 2008, 14(5): 221-236.

[15] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 121-135.
(Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 121-135.)

[16] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5/6): 602-610.

[17] Kong W C, Dong Z Y, Jia Y, et al. Short-term residential load forecasting based on LSTM recurrent neural network[J]. IEEE Transactions on Smart Grid, 2019, 10(1): 841-851.

[18] 梁昌勇, 马银超, 陈荣, 等. 基于SVR-ARMA组合模型的日旅游需求预测[J]. 管理工程学报, 2015, 29(1): 122-127.
(Liang C Y, Ma Y C, Chen R, et al. Prediction of daily tourism demand based on svr-arma combination model[J]. Journal of Management Engineering, 2015, 29(4): 122-127.)

[19] 屈拓. 组合模型在机场旅客吞吐量预测中的应用[J]. 计算机仿真, 2012, 29(4): 108-111.
(Qu T. Application of combination model in airport passenger throughput prediction[J]. Computer Simulation, 2012, 29(4): 108-111.)

[20] Petrevska B. Predicting tourism demand by ARIMA models[J]. Economic Research-Ekonomska Istrazivanja, 2017, 30(1): 939-950.

[21] Hong W C, Dong Y, Chen L Y, et al. SVR with hybrid chaotic genetic algorithms for tourism demand forecasting[J]. Applied Soft Computing, 2011, 11(2): 1881-1890.

[22] Sun S, Wei Y, Tsui K L, et al. Forecasting tourist arrivals with machine learning and internet search index[J]. Tourism Management, 2019, 70(2): 1-10.

[23] Xia L, Jie Y, Lei C, et al. Prediction for air route passenger flow based on a grey prediction model[J]. Computer Systems & Applications, 2017, 26(7): 221-226.

[24] Ratna Sulistyowati, Suhartono, Heri Kuswanto, et al. Hybrid forecasting model to Predict air passenger and cargo in Indonesia[C]. International Conference on Information and Communications Technology (ICOIACT). Yogyakarta: IEEE, 2018: 442-447.

[25] Yue G G, Zhang B M, Dai F H, et al. Three dimensional cure simulation of stiffened thermosetting composite Panels[J]. Journal of Materials Science & Technology, 2010, 26(5): 467-471.

[26] Hensher D A. Determining passenger potential for a regional airline hub at canberra international airport[J]. Journal of Air Transport Management, 2002, 8(5): 301-311.

[27] Li Y F, Cao H. Prediction for tourism flow based on LSTM neural network[J]. Procedia Computer Science, 2018, 129: 277-283.

[28] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A Search

space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28(10): 2222-2232.

[29] Sun S L, Lu H X, Tsui K L, et al. Nonlinear vector auto-regression neural network for forecasting air passenger flow[J]. Journal of Air Transport Management, 2019, 78: 54-62.

[30] Li J, Yao X F, Liu Y H, et al. Thermo-viscoelastic analysis of the integrated T-shaped composite structures[J]. Composites Science and Technology, 2010, 70(10): 1497-1503.

[31] Shannon C E. A mathematical theory of communication[J]. ACM SIGMOBILE Mobile Computing and Communications Review, 2001, 5(1): 3-55.

[32] Sharmin S, Shoyaib M, Ali A A, et al. Simultaneous feature selection and discretization based on mutual information[J]. Pattern Recognition, 2019, 91: 162-174.

[33] Chernyshov K R. An information theoretic approach to constructing machine learning criteria[J]. IFAC Proceedings Volumes, 2013, 46(11): 269-274.

[34] 范雪莉, 冯海泓, 原猛. 基于互信息的主成分分析特征选择算法[J]. 控制与决策, 2013, 28(6): 915-919.
(Fan X L, Feng H H, Yuan M. PCA based on mutual information[J]. Control and Decision, 2013, 28(6): 915-919.)

[35] 梁吉业, 冯晨娇, 宋鹏. 大数据相关分析综述[J]. 计算机学报, 2016, 31(1): 1-18.
(Liang J Y, Feng C J, Song P. Summary of big data relevance analysis[J]. Chinese Journal of Computer, 2016, 31(1): 1-18.)

[36] 王泳. 基于互信息与先验信息的机器学习方法研究[D]. 北京: 中国科学院研究生院, 2008.
(Wang Y. Research on machine learning method based on mutual information and prior information[D]. Beijing: Graduate School of Chinese Academy of Sciences, 2008.)

[37] 张春涛, 马千里, 彭宏. 基于信息熵优化相空间重构参数的混沌时间序列预测[J]. 物理学报, 2010, 59(11): 7623-7629.
(Zhang C T, Ma Q L, Peng H. Prediction of chaotic time series based on information entropy optimization for phase space reconstruction parameters [J]. Journal of Physics, 2010, 59(11): 7623-7629.)

[38] 郭明玮, 赵宇宙, 项俊平, 等. 基于支持向量机的目标检测算法综述[J]. 控制与决策, 2014, 29(2): 193-200.
(Guo M W, Zhao Y Z, Xiang J P, et al. Review of object detection methods based on SVM[J]. Control and Decision, 2014, 29(2): 193-200.)

[39] 李萍, 倪志伟, 朱旭辉, 等. 基于分形流形学习的支持向量机空气污染指数预测模型[J]. 系统科学与数学, 2018, 38(11): 1296-1306.
(Li P, Ni Z W, Zhu X H, et al. Air pollution index prediction model of support vector machine based on Fractal manifold learning[J]. Systems Science and Mathematics, 2018, 38(11): 1296-1306.)

[40] 孙涵, 杨普容, 成金华. 基于Matlab支持向量回归机的能源需求预测模型[J]. 系统工程理论与实践, 2011, 31(10): 2001-2007.
(Sun H, Yang P R, Cheng J H. Energy demand prediction model based on Matlab support vector regression machine[J]. Systems Engineering — theory & practice, 2011, 31(10): 2001-2007.)

[41] 许寅. 基于机器学习方法的航天器在轨状态异变趋势预测算法研究[D]. 成都: 电子科技大学, 2017.
(Xu Y. Research on prediction algorithm of abnormal trend of spacecraft in-orbit state based on machine learning method[D]. Chengdu: University of Electronic Science and Technology, 2017.)

[42] 孙轶轩, 邵春福, 计寻, 等. 基于ARIMA与信息粒化SVR组合模型的交通事故时序预测[J]. 清华大学学报: 自然科学版, 2014, 54(3): 348-353.
(Sun Y X, Shao C F, Ji X, et al. Time series prediction of traffic accidents based on ARIMA and information granulation SVR combination model[J]. Journal of Tsinghua University: Natural Science Edition, 2014, 54(3): 348-353.)

[43] Farber S, Ritter B, Fu L W. Space-time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: A social equity perspective[J]. Travel Behavior and Society, 2016, 4(5): 40-48.

[44] 戢晓峰, 刘澜. 基于出行行为的铁路出行信息传递指数模型[J]. 交通运输工程学报, 2008, 8(6): 99-103.
(Ji X F, Liu L. Railway travel information transfer index model based on travel behavior[J]. Journal of Traffic Transportation Engineering, 2008, 8(6): 99-103.)

[45] García Nieto P J, Combarro E F, Del Coz Díaz J J, et al. A SVM-based regression model to study the air quality at local scale in Oviedo urban area (Northern Spain): A case study[J]. Applied Mathematics and Computation, 2013, 219(17): 8923-8937.

作者简介

熊红林(1984−), 男, 博士生, 从事机器学习应用算法、智慧机场的研究, E-mail: honyex@126.com;

朱人杰(1982−), 男, 博士生, 从事智能优化算法、决策支持、智慧医疗的研究, E-mail: renjiezhu@aliyun.com;

冀和(1995−), 男, 硕士生, 从事机器学习、大数据分析方法的研究, E-mail: 1063462486@qq.com;

樊重俊(1963−), 男, 教授, 博士生导师, 从事人工智能算法、机场大数据、智慧医疗等研究, E-mail: fan.chongjun@163.com;

徐佩(1996−), 女, 硕士生, 从事机器学习、智慧机场的研究, E-mail: 2596810883@qq.com.

(责任编辑: 齐 霁)