文章编号：2095-2163(2019)05-0231-05 中图分类号：TPTP18 文献标志码：A

基于聚类与SVR的地区支线航空客运市场需求预测

徐梦瑶，赵 鸣，李 洋，安 洋，张友浩
（上海工程技术大学 航空运输学院，上海 201620）

摘 要： 针对支线航空客运市场需求预测问题，某些地区（如海南）缺少足够的历史数据，难以建立准确的预测模型。本文提出基于聚类与支持向量机回归（Support Vector Regression，SVR）预测此类地区航空客运市场需求的方法。首先，基于中国各个地区支线航空客运市场需求的分布比，找出与海南分布比相似的地区，再应用系统聚类法在这些地区中找出与海南聚为一类的地区，作为类比地区。然后，选择类比地区的数据样本，通过K-fold交叉验证（K-fold Cross Validation，K-CV）寻优SVR参数，得到预测模型。最后，预测了2018～2020年海南支线航空客运市场需求，从而为其建设支线机场提供一定的决策参考和可靠的理论依据，具有一定的现实意义和应用价值。

关键词： 支线航空；客运市场需求；预测；系统聚类；SVR

Prediction of regional aviation market demand in specific region based on clustering and SVR

XU Mengyao, ZHAO Ming, LI Yang, AN Yang, ZHANG Youhao
(School of Air Transportion, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 For predicting the market demand of regional air transportation for passengers, some regions (such as Hainan) lack enough available data to establish accurate prediction models. This paper proposes a method based on Clustering and Support Vector Regression (SVR) to predict the market demand of air transportation for passengers in such regions. Firstly, the paper finds the similar regions to Hainan in distribution ratio of the market demand, then compares Hainan with these regions which were clustered together. Secondly, the paper selects the data samples of the similar regions and forms a prediction model after getting the SVR parameters by K-fold Cross Validation (K-CV). Finally, the paper predicts the market demand of Hainan air transportation for passengers from 2018 to 2020. The results could provide theoretical support and guidance for constructing new regional airports, which is realistic and practical value to some extent.

【Key words】 regional aviation; passenger market demand; prediction; hierarchical cluster; SVR

0 引 言

中国民航局的相关资料显示，从2011～2016年，国内支线航线网络增加了27%，运力增加了116%。预计到2020年，支线航空客运量将会突破1亿人次，其发展速度约为干线航空的两倍[1]。国内支线机场在综合交通运输体系中发挥着越来越大的作用，因而为了抓住扩建支线机场数量的最佳有利时机，并减少支线机场建设的盲目性，即需对地区支线航空客运市场需求做出预测，提高针对性，同时也将对地区的支线机场建设和实际生产有着积极的指导意义。但是对于某些地区，如果支线机场通航时间较短，几乎无可用的历史数据，就使得支线航空客运市场需求的预测研究受到了一定的阻碍。

众多学者已经对这种缺少历史数据支撑的客货运需求预测问题展开了大量研究。张娜等人[2]提出了先通过快速聚类找出与新建机场相似的机场，再利用相似机场的航空分担率来预测新建机场客运量的方法。悦慧等人[3]运用动态聚类法找出与新建机场属于同类的机场，基于同类机场的历史数据构建多元回归模型，从而预测新建机场的客运量。但由于航空客运需求预测所受噪声和影响因素较多，并且各因素对支线航空客运市场需求的影响程度也不尽相同，这使得支线航空客运市场需求预测具有高度非线性的特点。故简单的多元线性回归模型已经不能满足预测需求。罗建锋等人[4]将机器学习的方法运用在新建机场货邮量预测上，即先利用相近周边机场航空货运量占社会总货运量的比例关系，并结合本地区GDP与航空货运量、旅客吞吐量与航空货邮量的比例关系进行校核，从而拟合出新建机场航空货邮量的历史数据，再将历史数据带入BP神经网络，预测新建机场的货邮量。BP神经

基金项目： 国家社科基金项目(15BJL104)。

作者简介： 徐梦瑶(1996-)，女，硕士研究生，主要研究方向：载运工具故障诊断与控制、信息处理与模式识别、数据挖掘与决策支持。

收稿日期： 2019-07-16

网络方法虽然能很好地处理非线性问题,但对于航空客运量预测这种影响因素较多且样本量较小的预测问题仍具有较大局限性,其预测出的精度较低[5]。

支持向量机(Support Vector Machine,SVM)是VAPNIK提出的一种建立在统计学理论的VC维理论和结构风险最小化原理基础上的机器学习方法[6]。支持向量机回归(Support Vector Regression,SVR)是由SVM衍生得到的,在解决小样本、非线性、高维度问题中显示出了绝对的优势[7-8]。在SVR的应用过程中,惩罚参数 C 与核函数参数 g 的选取对预测结果的影响很大,如何选取合适的参数成为问题的关键。赵静等人[9]采用了K-fold交叉验证(K-fold Cross Validation,K-CV)模型选择最优参数,提高了预测的精度。在前述研究的基础上,本文提出基于聚类与SVR预测支线航空客运市场需求的研究设计。

1 支持向量机回归(SVR)的基本原理

SVR是SVM衍生得到的,可分为线性SVR与非线性SVR。对此可做研究分述如下。

1.1 线性SVR

设一组样本数为 m 个的样本向量:(x_1,y_1),$(x_2,y_2),\cdots,(x_m,y_m)$,$x_i\in R^n$,$y_i\in R$,$i=1,2,\cdots,m$,样本集中 X 与 Y 的线性关系为:

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

其中,$\boldsymbol{w}$ 为权重系数向量,b 为偏置项。

如若原始数据能如公式(1)所示无误差地用线性关系进行拟合,则可以得到公式(2):

$$\min\frac{1}{2}\|w^2\|$$

$$\text{s.t.}\begin{cases}\boldsymbol{w}^{\mathrm{T}}\cdot x_i+b-y_i\leqslant\varepsilon\\ y_i-\boldsymbol{w}^{\mathrm{T}}\cdot x_i-b\leqslant\varepsilon\end{cases},i=1,2,\cdots,l, \tag{2}$$

其中,ε 可以为任意一个正数。

将Lagrange对数引入公式(1)中,可得:

$$f(x)=\boldsymbol{w}^{\mathrm{T}}\cdot x+b=\sum_{i=1}^{l}(\boldsymbol{a}_i-\boldsymbol{a}_i^*)(x_i\cdot x)+b. \tag{3}$$

其中,$\boldsymbol{a}_i$ 和 $\boldsymbol{a}_i^*$ 为样本支持向量。

1.2 非线性SVR

将样本 x_i 通过 $\varphi(x):x\to H$ 映射到高维的空间。为构造出最优的超平面,在 $\varphi(x)$ 未知的情况下,利用原空间参数实现内积运算。为了解决维数灾难问题,当核函数满足Mercer条件,即可获得内积核函数 $K(x_i,x_j)=\varphi(x_i)\cdot\varphi(x_j)$。同时引入Lagrange变化得到:

$$L(w,\xi,b,a,\beta)=\sum_{i=1}^{l}a_i-\frac{1}{2}\sum_{i,j=1}^{l}y_iy_ja_ia_jK(x_ix_j), \tag{4}$$

最终得到变形后的回归函数,可写为如下形式:

$$f(x)=\boldsymbol{w}^{\mathrm{T}}\cdot x+b=\sum_{i=1}^{l}(\boldsymbol{a}_i-\boldsymbol{a}_i^*)K(x_i\cdot x)+b. \tag{5}$$

SVR不同的内积核函数会形成不同的算法,一般常用的核函数有3种[10],具体如下:

(1)线性核函数:

$$K(x,x_i)=x^{\mathrm{T}}x_i$$

(2)多项式核函数:

$$K(x,x_i)=(\mu x^{\mathrm{T}}x_i+\gamma)^p,\ \mu>0$$

(3)径向基核函数(radial basis function,RBF):

$$K(x,x_i)=\exp(-\mu\|x-x_i\|^2),\ \mu>0$$

2 海南支线航空客运市场需求分析

2.1 研究方法与思路

本文对支线机场的界定需要满足2个条件,对此可阐述为:

(1)年旅客吞吐量占全国旅客总吞吐量的比例小于0.2%。

(2)机场处于非国家中心城市、非省会城市,属于非枢纽性机场[11]。

海南地区的支线机场现有琼海博鳌机场和三沙永兴机场,2个支线机场都是2016年通航,通航时间短,且缺乏历史数据。对数据样本缺乏或较少的通航地区,如海南的支线航空客运市场需求预测就转化为对旅客吞吐量的预测,以整体把握该地区支线航空客运市场未来的发展趋势。

本文研究思路是:首先,基于全国和各地区支线机场旅客吞吐量,提出各个地区支线航空客运市场需求的分布比。接着找出与海南分布比相似的地区,再应用系统聚类法在这些地区中求出与海南聚为一类的地区,作为类比地区。然后,将选定地区的历史值作为训练数据,代入SVR预测模型,通过K-CV寻优SVR参数,确定预测模型。最后,对海南的支线航空旅客吞吐量进行预测,为其建设支线机场提供一定的决策参考。本文的技术研发路线如图1所示。

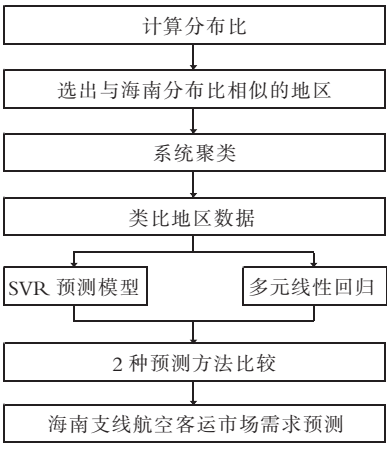

图 1　技术路线图

Fig. 1　Technology route

2.2　类比地区的选择

2.2.1　地区支线航空客运市场需求分布比

中国幅员辽阔、地形地貌差异较大，不同的地域条件导致了中国支线机场分布的不均衡，也影响着机场旅客吞吐量。本文引入分布比的概念[12]，定义了某地区支线航空客运市场需求分布，即某地区支线机场旅客吞吐量与全国支线机场旅客吞吐量的比值。研究推得其数学公式可写为：

$$\gamma=\frac{T_{\text{地区}}}{T_{\text{全国}}}\times 100\%. \tag{6}$$

其中，γ 表示某地区支线航空客运市场需求分布比；$T_{\text{地区}}$ 表示某地区支线机场旅客吞吐量，单位为：人；$T_{\text{全国}}$ 表示全国支线机场旅客吞吐量，单位为：人。

中国各个地区支线航空客运市场需求分布比如图 2 所示。海南位于国内中南地区，由图 2 可知，中南地区与西南地区的支线航空客运需求基本处于相同水平，都在 10% ~23% 之间。故从这 2 个地区中选取贵州、四川、西藏、云南、重庆、广东、广西、海南、河南、湖北、湖南 11 个省份作为类比样本。

2.2.2　分布比相似地区的系统聚类

从影响支线航空客运需求的人口、地区经济发展情况的角度来考虑，选择人口密度、人均 GDP、城镇居民人均可支配收入、城镇居民人均消费支出 4 个指标作为聚类的评价指标。选取 2008 ~2017 年各地区指标值的平均值作为样本数据。用系统聚类法对样本数据进行聚类。由此得到的分布比相似地区的聚类树图即如图 3 所示。

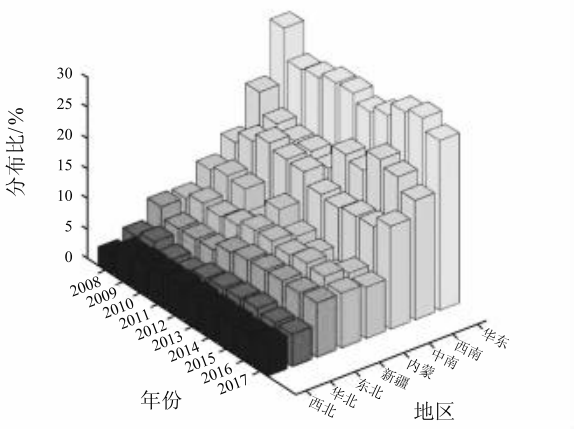

图 2　国内各地区支线航空客运市场需求分布比

Fig. 2　Distribution ratio of the market demand of regional aviation transport for passengers in various regions in China

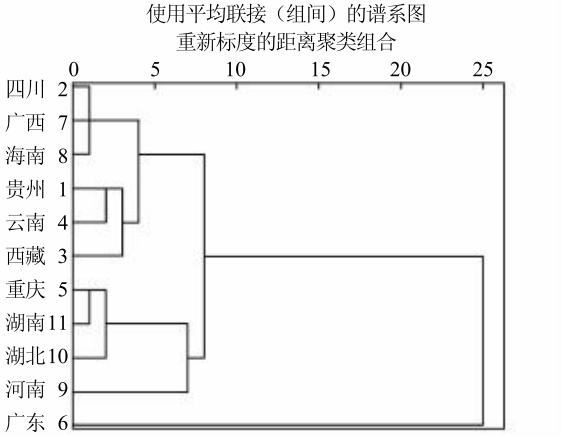

图 3　分布比相似地区的聚类树图

Fig. 3　Cluster tree diagram of region with similar distribution ratio

由图 3 可知，海南、广西、四川聚为一类。由于广西与海南同属于中南地区，且广西与海南地理位置靠近，与海南的人口密度、人均 GDP、城镇居民人均可支配收入、城镇居民人均消费支出也非常相近。故最终选择广西作为海南的类比地区。

3　基于 SVR 的模型构建与预测

3.1　模型构建

因海南与广西同属一类，且广西数据充足，将广西的人口密度、人均 GDP、城镇居民人均可支配收入、城镇居民人均消费支出 4 组数据作为输入特征值，年旅客吞吐总量作为输出特征值。选择广西 2008 ~2016 年的 9 组数据作为 SVR 模型的训练样本，2017 年数据作为测试样本。研发设计步骤可剖析分述如下。

（1）用 Matlab 中的 *mapminmax* 函数来对 10 组

样本数据进行归一化处理,防止特征值范围过大或过小,影响模型的精确度。其中,归一化的范围为[-1,1]。

(2)选择 SVM 的类型为 ε-SVR,核函数选取精度较高的 RBF 函数[13-14]。设置 ε-SVR 中的损失函数 p 的值为 0.1。

(3)采用 K-CV($V=5$,即将测试集分为 5 部分进行交叉验证)的参数优化方法选择一组最优参数($C=1\ 024$,$g=0.001\ 381\ 1$),如图 4 所示。

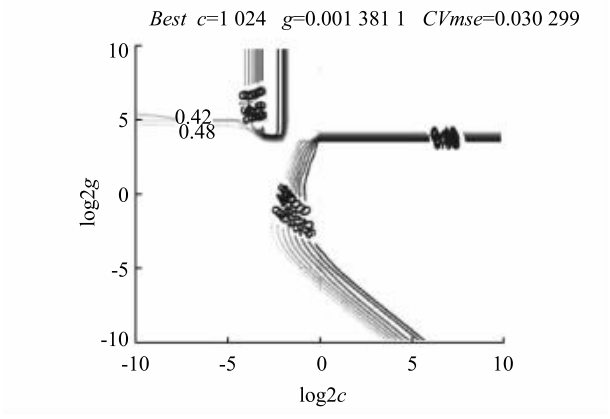

图 4 参数 C、g 寻优的等高线图

Fig. 4 Contour map of optimal parameters C,g

(4)将最佳参数(C,g)和训练样本代入 SVR 中,并得到精度较高的 SVR 模型($MSE=0.007\ 745\ 6$,$R^2=0.977\ 4$)。运行结果详见图 5。

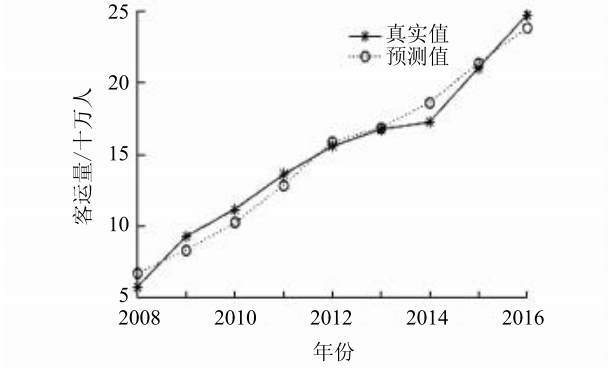

图 5 SVR 模型训练精度

Fig. 5 Training accuracy of SVR

(5)基于测试样本,对此模型进行精度比较,并与多元线性回归模型进行对比,运算对比结果见表 1。

表 1 中展示了广西样本数据分别在 SVR(RBF 核函数)模型与多元线性回归模型下的预测精度,由误差平均值可以看出 SVR(RBF 核函数)模型的预测准确率高于多元线性回归模型,这也说明了 SVR 模型在解决小样本、非线性问题上占有优势。

表 1 不同模型下广西支线机场旅客吞吐量预测值与实际值对比

Tab. 1 Comparison of passengers in Guangxi regional airports under different models

年份	实际值/人	SVR(RBF 核函数)		多元线性回归	
		预测值/人	相对误差/%	预测值/人	相对误差/%
2008	576 927	691 543	19.87	648 556	12.42
2009	929 473	873 182	6.06	807 346	13.14
2010	1 118 315	1 062 274	5.01	1 056 527	5.53
2011	1 358 994	1 301 620	4.22	1 354 023	0.37
2012	1 554 188	1 611 883	3.71	1 659 688	6.79
2013	1 674 935	1 640 147	2.08	1 709 205	2.05
2014	1 723 860	1 779 767	3.24	1 835 212	6.46
2015	2 099 868	2 134 146	1.63	2 114 630	0.70
2016	2 470 677	2 413 330	2.32	2 322 107	6.01
2017	2 998 696	2 650 560	11.61	2 558 673	14.67
平均误差			5.90		6.80

3.2 海南支线航空旅客吞吐量预测

因缺少 2018~2020 年份影响海南支线航空客运市场需求因素的统计数据,将根据 2008~2017 年海南省支线航空客运市场需求影响因素统计数据,建立海南地区支线航空客运市场影响因素与年份间的一元线性关系,预测 2018~2020 年影响因素的指标值,运算预测结果见表 2。

表 2 海南支线航空旅客吞吐量影响因素预测值(2018~2020 年)

Tab. 2 Prediction of the effecting factors of Hainan regional aviation transport(2018~2020)

年份	人口密度/(人/万平方公里)	人均 GDP/元	城镇居民人均可支配收入/元	城镇居民人均消费支出/元
2018	263.220	32 725.584	22 030.089	52 043.578
2019	265.471	34 789.384	23 331.724	55 501.766
2020	267.722	36 853.184	24 633.359	58 959.954

将表 2 中海南支线航空旅客吞吐量影响因素预测值作为输入特征值,即可得到海南旅客吞吐量的输出预测值,详见表 3。表 3 表明了在已确定的 SVR(RBF 核函数)模型下海南 2018~2020 年支线航空旅客吞吐量预测值。

表 3 海南支线航空旅客吞吐量预测值(2018~2020 年)

Tab. 3 Prediction of the passengers of Hainan regional aviation transport (2018~2020) 人

年份	预测值
2018	4 479 303
2019	4 796 637
2020	5 105 165

4　结束语

针对某些地区（如海南）缺少足够的历史数据，难以建立航空客运市场需求预测模型的问题，本文提出基于聚类与 SVR 预测支线航空客运市场需求的方法。根据类比法的思想，首先，选取与海南地区机场旅客吞吐量分布比相似的地区（如贵州、四川、西藏等）进行系统聚类，找出类比地区（广西）。然后，选择广西省 2008～2017 年的数据样本，通过 K-CV 寻优 SVR 参数（$C=1\ 024$，$g=0.001\ 381\ 1$），得到预测模型。将此模型与多元线性回归预测方法进行精度比较，证明 SVR（RBF 核函数）预测模型具有更好的预测效果。基于此模型，预测了 2018～2020 年海南支线航空旅客吞吐量，从而为其建设支线机场提供一定的决策参考和可靠的理论依据，具有一定的现实意义和应用价值。

参考文献

[1] 张一琛. 支线航企如何"叫好又叫座"[J]. 大飞机，2017(3)：28-31.

[2] 张娜，安然. 基于快速聚类分析的航空分担率模型在新建机场客运量预测中的应用[J]. 交通与计算机，2008，26(4)：116-119.

[3] 悦慧，安然. 多元回归模型在新建机场客运量预测中的应用研究-基于动态聚类分析[J]. 现代商贸工业，2010，22(20)：13-15.

[4] 罗建锋，周凌云，李伟. 基于 BP 神经网络的新建支线机场货邮量综合预测[J]. 江苏商论，2012(2)：47-49.

[5] 曾鸣，林磊，程文明. 基于 LIBSVM 和时间序列的区域货运量预测研究[J]. 计算机工程与应用，2013，49(21)：6-10.

[6] VAPNIK V N. The nature of statistical learning theory[M]. New York：Springer，2000.

[7] ABDI M J，GIVEKI D. Automatic detection of erythemato-squamous diseases using PSO-SVM based on association rules[J]. Engineering Applications of Artificial Intelligence，2013，26(1)：603-608.

[8] LIU Zhiwen，CAO Hongrui，CHEN Xuefeng，et al. Multi-fault classification based on wavelet SVM with PSO algorithm to analyze vibration signals from rolling element bearings[J]. Neurocomputing，2013，99：399-410.

[9] 赵静，王选仓，丁龙亭，等. 基于灰色关联度分析和支持向量机回归的沥青路面使用性能预测[J]. 重庆大学学报，2019，42(4)：72-81.

[10] 张文雅，范雨强，韩华，等. 基于交叉验证网格寻优支持向量机的产品销售预测[J]. 计算机系统应用，2019，28(5)：1-9.

[11] 李飞行，宋一鑫，张权. 我国支线机场现状分析及对策研究[J]. 交通运输研究，2018，4(4)：61-68.

[12] 周明妮. 新建支线机场通航可行性论证方法研究[D]. 西安：长安大学，2011.

[13] AYDIN I，KARAKOSE M，AKIN E. A multi-objective artificial immune algorithm for parameter optimization in support vector machine[J]. Applied Soft Computing，2011，11(1)：120-129.

[14] de CASTRO L N，von ZUBEN F J. Learning and optimization using the clonal selection principle[J]. IEEE Transactions on Evolutionary Computation，2002，6(3)：239-251.

（上接第 230 页）

系统中总共分为 3 层，即：controller 层、service 层和 mapper 持久化层。其中，controller 层主要做业务流程控制，service 层进行事务控制，mapper 层则重点实现数据的持久化操作。每一层都通过接口向其调用层提供服务，这样在底层代码发生改变时，只要接口不变，就无需改变上层业务逻辑代码，减少层组件之间耦合度，也方便系统的维护。

系统采用当前比较流行的开发框架 SSM，用 MySQL 存储学生的评价数据，为了增加系统的并发量，提高后台业务系统的读写效率，又采用 redis 作为业务系统的缓存，可以保证系统在高并发的同时，业务系统有着较高的读写效率，提高业务系统响应速度，优化用户感受。

5　系统测试

本系统经过设计和编码，已经完成了大部分的工作。系统在应用阶段最主要的就是保证在高并发的情况下，系统能够正常运转和响应。

本系统采用模拟请求并发的方法进行系统的压力测试。在并发量小于 500 时，系统的响应时间少于 1 s；在并发量为 1 000 时，要超过 1 s；当并发量 2 000内，响应速度不超过 3 s。对于笔者院校规模的学生来说，已经足够应对学生高峰时评价的响应时间，保证了学生对系统的良好体验。

6　结束语

本系统的应用，对本校的教学起到了一定的监督作用。从学生的角度来说，学生可以根据自身实际情况，向授课教师提出建议，授课老师根据学生的建议改进自己的教学。从学院的角度来说，学院也可以实时督查监管本学院教师的工作进展。评价结果也可以为学院领导制定教学决策提供参考。

参考文献

[1] 濮宏积，王娟，黄晓云，等. 微信公众平台在教学质量监控体系中的运用[J]. 曲靖师范学院学报，2016，35(6)：48-52.

[2] 李宗富，张向先. 政务微信公众号服务质量评价指标体系构建及实证研究[J]. 图书情报工作，2016，60(18)：79-88.

[3] 高昱，乔世峰. 基于移动通信平台的教学信息化系统建设研究[J]. 中国管理信息化，2018，21(4)：129-130.