基于 ARIMA-SVM 的体育彩票销售量预测

鄢正纲
(中南财经政法大学 体育部,湖北 武汉 430064)

摘 要:体育彩票销售量受到多种综合因素影响,呈现出复杂的、非线性的动态变化特性.为了准确刻画体育彩票销售量的变化特征,提出差分自回归移动平均和支持向量机相融合的体育彩票销售量预测模型.首先根据体育彩票销售量时间序列建立 ARIMA 模型,拟合体育彩票销售量的线性变化部分,然后采用支持向量机对差分自回归移动平均的预测残差进行建模,拟合体育彩票销售量的非线性变化部分,最后采用具体体育彩票销售量数据进行仿真实验.仿真结果表明,相对于其它模型,该模型具有更高的预测精度,可以更准确反映体育彩票销售量的变化趋势.

关键词:体育彩票销售量;差分自回归移动平均;支持向量机;组合预测
中图分类号: TP183 文献标志码: A 文章编号: 1671-9476(2017)02-0123-04
DOI:10.13450/j.cnki.jzknu.2017.02.031

随着经济的迅速发展,体育彩票也随之壮大,为国家和社会筹集了大量的公益金,并带动了相关产业发展.对体育彩票的未来销量进行准确预测,为体育彩票的营销推广提供科学依据,对体育彩票产业的健康发展具有十分重要的意义[1].

针对体育彩票销售量问题,国内外学者进行了大量而深入地研究,并取得许多研究成果.体育彩票销售量与经济水平、市场规模、居民收入、节假日相关,变化非常复杂,传统体育彩票销售量模型主要是差分自回归移动平均模型(autoregressive integrated moving average,ARIMA),其主要思想是利用体育彩票销售量前 m 个数据量作为描述因子建立回归模型,对未来体育彩票销售量进行预测,然而 ARIMA 本质上属于线性模型,在描述时间序列的线性特征时有一定的优越性,但在描述其非线性特征时却具有局限性[2].随着机器学习算法的不断成熟,支持向量机(support vector machine,SVM)在体育彩票销售量预测中取得了不错的应用效果[3-6].然而体育彩票销售量受到人的心理、体育赛事等影响,具有明显整体趋势变动性和季节波动性,以及随机性,仅使用 SVM 无法对体育彩票销售量进行高精度地预测[7-9].近年来,根据组合优化理论,研究人员将不同模型组合在一起,实现优势互补,可以提高预测精度,因此组合预测模型为体育彩票销售量预测问题提供了一种新的研究思想.

为了提高体育彩票销售量的预测精度,利用 ARIMIA 和 SVM 的优点,提出了一种 ARIMA 和 SVM 相融合的体育彩票销售量预测模型(ARIMA-SVM),最后利用某地区体育彩票销售量数据进行仿真实验.仿真结果表明,ARIMA-SVM 获得了较高的体育彩票销售量预测精度,同时相对其它体育彩票销售量预测模型,预测和建模效率更高,具有一定的优势.

1 ARIMA-SVM 的工作流程

体育彩票销售量是按照一定的时间间隔收集的数据,是可以采用 $\{x_1,x_2,\cdots,x_n\}$ 来表示,体育彩票销售量受到多种综合因素的影响,不仅具有线性、周期性变化特点,同时也具有非线性变化特征,其预测的数学模型可描述为:

$$\hat{y}=f(x_1,x_2,..,x_n) \tag{1}$$

式中,$\hat{y}$ 表示体育彩票销售量的预测值,f ()表示预测模型.

ARIMA-SVM 的建模与预测思想为:采用 ARIMA 对体育彩票销售量时间序列进行建模,对

收稿日期:2016-07-18;修回日期:2016-11-25
基金项目:湖北省高校省级教学改革研究项目(No.2013160)
作者简介:鄢正纲(1978—),男,湖北武汉人,硕士,讲师,研究方向:体育教育训练学.

预测体育彩票销售量序列的线性特点进行描述，然后，根据残差值包含了体育彩票销售量时间序列的非线性特点，利用支持向量机进行建模，最后将两者进行相加，得到体育彩票销售量的最终预测值. ARIMA-SVM 的具体工作流程如图 1 所示.

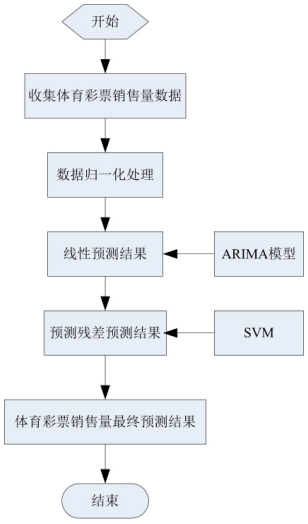


图 1 体育彩票销售量预测模型的工作流程图

2 ARIMA-SVM 的体育彩票销售量模型

2.1 ARIMA 模型

ARIMA 模型是 Box 等提出的一种时间序列建模方法，对原始序列 Z_t 进行 d 阶差分，得到序列 $(1-B)^d Z_t$，设 p 和 q 为阶数，ARIMA (p,q) 模型可以描述为：

$$(1-\varphi_1 B-,\cdots,-\varphi_p B^p)(1-B)^d Z_t=\theta_0+(1-\theta_1 B-,\cdots,-\theta_q B^q)\varepsilon_t \tag{1}$$

式中，B 为滞后算子；ε_t 为白噪声；$\varphi_i(i=1,2,\cdots,p)$和 $\theta_j(j=1,2,\cdots,q)$ 为参数[10].

2.2 支持向量机

支持向量机通过函数 $\varphi(x)$ 对数据进行非线性映射，将问题转化为凸二次规划问题：

$$\min_{\omega,b,\xi} J(\omega,\xi_i)=\frac{1}{2}\omega^{\mathrm{T}}\cdot\omega+C\sum_{i=1}^{l}\xi_i^2$$

$$s.t.\quad \begin{cases} y_i=\omega^{\mathrm{T}}\cdot\varphi(x_i)+b+\xi_i \\ \xi_i\geq 0 \\ i=1,2,\cdots,l \end{cases} \tag{2}$$

式中，ξ_{i2} 为训练误差；参数 C 为惩罚因子[11-12].

引入对偶问题的 Lagrange 约束规划，具体如下：

$$\begin{cases} L(\omega,b,\xi,\alpha)= \\ J(\omega,\xi)-\sum_{i=1}^{l}\alpha_i(\omega\cdot\varphi(x_i)+b+\xi_i-y_i) \\ \alpha_i\geq 0 \\ i=1,2,\cdots,l \end{cases} \tag{3}$$

式中，α_i 为 Lagrange 乘子.

对式(3)中的 ω,b,ξ_i,α_i 进行求偏导得到：

$$\begin{cases} \frac{\partial L}{\partial\omega}=0 \\ \frac{\partial L}{\partial b}=0 \\ \frac{\partial L}{\partial\xi_i}=0 \\ \frac{\partial L}{\partial\alpha_i}=0 \end{cases}\Rightarrow\begin{cases} \omega=\sum_{i=1}^{l}\alpha_i\varphi(x_i) \\ \sum_{i=1}^{l}\alpha_i=0 \\ \alpha_i=c\xi_i \\ \omega^{\mathrm{T}}\varphi(x_i)+b+\xi_i-y_i=0 \end{cases} \tag{4}$$

式中，$i=1,2,\cdots,l$.

消去 ω 和 ξ_i，得矩阵方程为：

$$\begin{bmatrix} 0 & \boldsymbol{I}^{\mathrm{T}} \\ \boldsymbol{I} & \varphi(x_i)^{\mathrm{T}}\varphi(x_l)+c^{-1}\boldsymbol{I} \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix}=\begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{5}$$

式中，设 $Z=[\varphi(x_1),\varphi(x_2),\cdots,\varphi(x_l)]^{\mathrm{T}}$，$Y=[y_1,y_2,\cdots,y_l]^{\mathrm{T}}$，$\rho=[\boldsymbol{I}_1,\boldsymbol{I}_2,\cdots,\boldsymbol{I}_l]^{\mathrm{T}}$，$\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_l]^{\mathrm{T}}$，$\xi=[\xi_1,\xi_2,\cdots,\xi_l]^{\mathrm{T}}$，$\boldsymbol{I}$ 为单元矩阵.

根据 Mercer 条件，可得：

$$\Omega_{il}=\varphi(x_i)^{\mathrm{T}}\varphi(x_l)=K(x_i,x_l) \tag{6}$$

解上述方程组得：

$$\begin{cases} b=\frac{\rho^{\mathrm{T}}(ZZ^{\mathrm{T}}+c^{-1}\boldsymbol{I})Y}{\rho^{\mathrm{T}}(ZZ^{\mathrm{T}}+c^{-1}\boldsymbol{I})\rho} \\ \alpha=(ZZ^{\mathrm{T}}+c^{-1}\boldsymbol{I})-1(Y-b\rho) \end{cases} \tag{7}$$

根据上述求解可得支持向量机的预测模型为：

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)k(x_i,x)+b \tag{8}$$

2.3 ARIMA 和支持向量机的体育彩票销售量预测

体育彩票销售量数据 Z_t 可以描述为

$$Z_t=f(L_t,N_t) \tag{9}$$

式中，L_t 和 N_t 分别代表线性和非线性变化规律.

体育彩票销售量预测模型的工作步骤如下：

(1)利用 ARIMA 模型对 L_t 建模，有：

$$L_t=\hat{L}_t+e_t \tag{10}$$

式中，$\hat{L}_t$ 为 ARIMA 的估计，e_t 为估计残差.

对(10)式进行分析可以发现，e_t 隐含 Z_t 的非线性变化特点，因此，N_t 可看作残差序列和原体育彩票销售量序列的非线性函数，即：

$$N_t=f^1(e_{t-1},\ldots,e_{t-n},Z_{t-1},\ldots,Z_{t-m}) \tag{11}$$

式中，f^1 是非线性变化部分的拟合函数；n,m 均为正数.

因此，由(3)～(5)式有：

$$Z_t=f(e_{t-1},\ldots,e_{t-n},\hat{L}_t,Z_{t-1},\ldots,Z_{t-m}) \tag{12}$$

(2)根据式(12)，利用支持向量机进行建模.将

$L_t, e_i(i=t-1, t-2, \cdots, t-n)$ 和 $Z_i(j=t-1, t-2, \cdots, t-m)$ 作为支持向量机的输入变量,Z_t 为输出变量,根据支持向量机进行训练建立体育彩票销售量预测模型.

(3)利用训练好的模型进行体育彩票销售量预测.

3　仿真实验

3.1　仿真环境及对比模型

为了验证 ARIMA-SVM 的体育彩票销售量预测性能,在 Pentium(R) 双核 2.8 GHz、4G RAM、Windows 7 的操作系统计算机上,采用 VC++进行仿真实验.为了测试 ARIMA-SVM 的优越性,选择 ARIMA 和 SVM 在相同条件下进行仿真实验,其中支持向量机参数采用遗传算法进行优化.采用均方根误差(RMSE)和平均相对百分比误差(MPAE)对体育彩票销售量预测结果衡量.

3.2　数据来源

仿真数据资料来源于 2010－2013 年某地区的体育彩票月销售量,共 48 个数据,具体如图 2 所示,其中以前 24 个月体育彩票销售量作为训练样本进行建模,最后 24 个体育彩票销售量作为测试样本测试模型的泛化和推广能力.

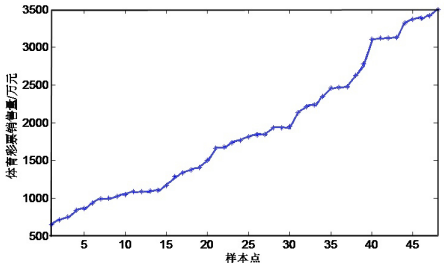


图 2　2010－2013 年某地区的体育彩票销售量图

3.3　模型的实现

对体育彩票销售量预处理,采用 DPS 6.5 软件作为建模工具,通过 ARIMA 模块建立体育彩票销售量偏相关和自相关图,如图 3 所示.

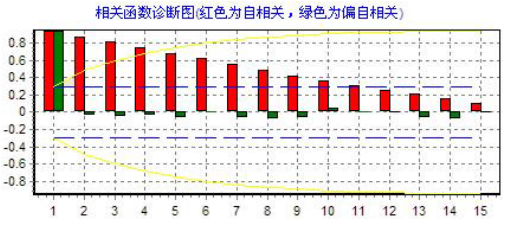


图 3　原始体育彩票销售量偏相关和自相关图

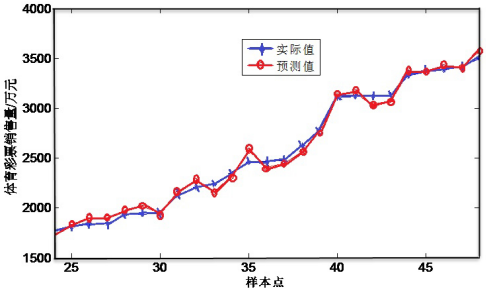
从图 3 可知,该体育彩票销售量自我相关性极高,自相关性呈下降趋势,有拖尾现象,对其进行差分处理,使其变成平稳时间序列,在进行一阶差分后,1 阶偏相关和自相关图如图 4 所示,从图 4 可知,体育彩票销售量基的阶数 $d=1$,根据 AIC 准则和 SC 准则 ARIMA 模型为 ARIMA(3,1,2).

图 4　体育彩票销售量 1 阶偏相关和自相关图

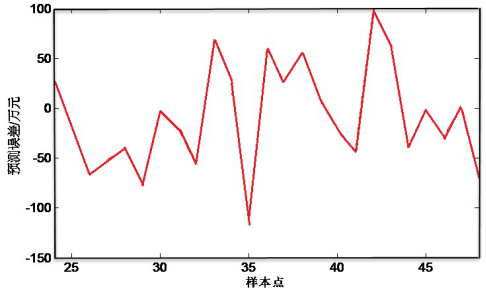
3.4　结果与分析

3.4.1　单步预测结果

ARIMA、SVM 以及 ARIMA-SVM 的体育彩票销售量单步预测结果如图 5 所示,可以明显看出,相对于对比模型,ARIMA-SVM 大幅度降低了体育彩票销售量预测误差,提高了体育彩票销售量的预测精度.



(a)预测值与实际值的变化曲线



(b) 预测误差的变化曲线

图 5　ARIMA-SVM 的体育彩票销售量单步预测结果图

表 1　ARIMA-SVM 与对比模型的单步预测误差

预测模型	RMSE	MAPE/%
ARIMA	68.39	2.63
SVM	58.01	3.68
ARIMA-SVM	46.24	1.91

3.4.2　多步预测结果

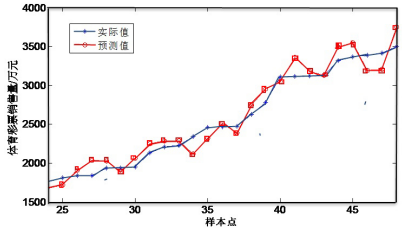
ARIMA-SVM 提前 2 和 4 步预测结果及预测误差的变化曲线如图 6～图 7 所示.从图 5 可以清楚看出,ARIMA-SVM 的预测误差率范围均小于 10%,预测结果符合要求,可用于体育彩票销售量的预测.

多步体育彩票销售量预测误差见表 2,从表 2 可以得到结论为:

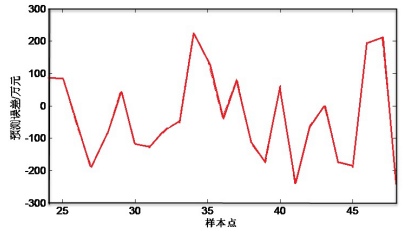
(1)预测步长越大,ARIMA、SVM 的预测误差

增加幅度相当，预测精度低，难以描述体育彩票销售量变化趋势，预测结果没有什么实际应用价值.

(2)相对于 ARIMA、SVM，ARIMA-SVM 的预测精度得到相应提高，预测误差相对较小，这主要是由于 ARIMA-SVM 基于组合优化理论，从不同的方面对体育彩票销售量的变化趋势进行预测，预测结果更加可靠，可以获得较理想的体育彩票销售量预测结果.

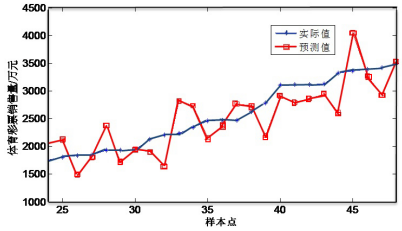


(a)预测值与实际值的变化曲线图

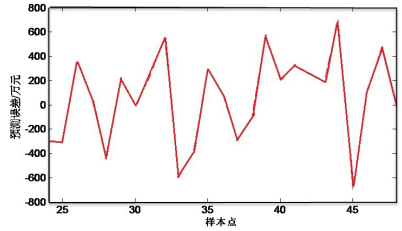


(b)预测误差的变化曲线图

图 6 ARIMA-SVM 的体育彩票销售量提前 2 步预测结果



(a)预测值与实际值之间的变化曲线图



(b)预测误差的变化曲线图

图 7 ARIMA-SVM 的体育彩票销售量提前 4 步预测结果

表 2 不同模型的体育彩票销售量多步预测误差对比

预测模型	2 步预测		4 步预测	
	RMSE	MAPE/%	RMSE	MAPE/%
ARIMA	330.38	14.23	530.38	23.66
SVM	328.48	13.81	428.48	23.52
ARIMA-SVM	126.49	7.73	226.51	9.80

4 结束语

将 ARIMA 与 SVM 技术进行融合，建立体育彩票销售量组合预测模型，并采用具体体育彩票销售量数据进行仿真实验，仿真结果表明，ARIMA-SVM 集成了 ARIMA 和 SVM 的优势，可以描述体育彩票销售量的变化特性，获得了更理想的体育彩票销售量预测结果，具有更高的实际应用价值.

参考文献：

[1]Ariyabuddhiphongs V. Lottery gambling: a review[J]. Journal of gambling Studies, 2011, 27(1):15－33.

[2]谢琼桓. 关于发行体育彩票的若干问题[J].体育科学, 2000, 20(3):7－9.

[3]刘炼，王斌. 基于计划行为理论的体育彩民购彩行为研究[J]. 上海体育学院学报, 2014,38(4):42－46.

[4]李海，陶蕊，傅琪琪，等. 上海市体育彩票问题彩民现状[J]. 体育科研, 2011, 32 (3): 43－49.

[5]史文文，王斌，刘炼，等. 体育彩票消费中问题彩民判断标准的研制[J]. 北京体育大学学报, 2013,36(6): 22－26.

[6]杨亚莉，程林林，张永韬. 体育彩票销量的计量模型及促销策略研究—以四川省为例[J]. 成都体育学院学报, 2012, 38(9): 1－7.

[7]李刚. 彩票人均销量的决定因素和我国彩票市场发展趋势的预测[J]. 体育科学, 2006, 26(12): 38－45.

[8]史文文. 问题彩民的购彩心理与行为特征[J]. 心理科学进展, 2012,20(4) :592－597.

[9]李海. 我国体育彩票问题彩民现状调查—以上海、广州、郑州、沈阳、成都为例[J].成都体育学院学报,2011, 37(5):9－13.

[10]吴殷，李海. 基于 ARIMA 模型的体育彩票销量预测—以上海为例[J]. 体育科研,2013,34(5):23－26.

[11]罗赟骞，夏靖波，王涣彬. 混沌－支持向量机回归在流量预测中的应用研究[J].计算机科学,2009,6(7):244－246.

[12]张培林，钱林方.基于蚁群算法的支持向量机参数优化[J]. 南京理工大学学报(自然科学版), 2009,33(4): 464－468.

Sales volume prediction of sports lottery based on autoregressive integrated moving average and support vector machine

YAN Zhenggang

(P.E. Department, Zhongnan University of Economics and Law, Wuhan 430064, China)

Abstract: Sales volume of sports lottery is comprehensively influenced by a variety of effects and has complex dynamic and nonlinear variation features, in order to accurately describe the sales volume of sports lottery, a sales volume prediction model of sports lottery based on autoregressive integrated moving average and support vector machine is proposed in this paper. Firstly, ARIMA model is used to predict linear structure of sports lottery sales volume, and then support vector machine is used to model the prediction residual of autoregressive integrated moving average, finally the specific sports lottery sales data is used to test the performance by simulation experiment. The simulation results showed that, compared with other models, the proposed model has higher prediction accuracy and can more accurately reflect the change trend of sports lottery sales volume.

Key words: sports lottery sales volume; autoregressive integrated moving average; support vector machine; combine prediction