基于SVM的上证指数预测研究

张晶华[1],莫文柯[2],甘宇健[1]

(1.广西财经学院 信息与统计学院;2.中国工商银行广西分行南宁民族支行,广西 南宁 530003)

摘 要:国家政策或市场经济导向等变动会对整个上证指数产生影响,为了寻找上证指数变化规律,提出基于支持向量机的预测算法。算法首先利用数据挖掘技术在某网站上挖掘相关的上证价格数据,并取一部分上证数据作为支持向量机的训练指数样本,得到支持向量机的训练指数集,然后在训练指数集上利用支持向量机,从而得到上证指数分类的超平面指数函数以及相关的上证指数样本集,最后对所得的上证指数分3个模型进行预测研究,得到下一个开盘日的上证指数变动预测数据。实验结果表明,预测2天后的上证指数趋势只需要前3天的数据作为自变量输入即可,且所得预测值与实际数值的误差率较低。

关键词:上证指数;SVM;数据挖掘;股票预测

DOI:10.11907/rjdk.171351

中图分类号:TP319 **文献标识码**:A **文章编号**:1672-7800(2017)008-0156-04

0 引言

影响上证指数变化的因素诸多,比如企业交易、市场经济、国家政策导向、居民消费能力、国际交易信息、人民币汇率变化等,都会直接或间接地对上证指数变动产生影响。上述因素之间存在着彼此交叉影响,能够对上证指数进行有效预测,这对金融投资者、金融行业,乃至整个股票市场具有重要指导意义和实用价值。

目前,利用支持向量机对股票进行研究的文献不多。文献[1]利用回归预测法对股票进行了短期预测,取得了初步成果,但利用回归预测方法要求的变量多且难确定,所得预测效果不理想。文献[2]利用时间序列预测法对股票进行短期预测,也取得了一定的效果,但文献[2]并没有对股票变化的规律进行深入研究。文献[3]和文献[4]虽然利用SVM对股票进行了研究,但只是简单拿一些数据用SVM方法进行计算,并没有进一步研究用什么样的数据进行预测得到的效果最好。本文在文献[3]和文献[4]的基础上,利用SVM方法,通过对比不同时期的上证指数,并对下一个开盘日上证指数的变化值进行预测。实验结果表明,本文所得的预测结果与实际相差不大,具有一定的实际意义和指导价值。

1 支持向量机

支持向量机[5-14]是在利用统计学分析数据时面对有限样本研究其性能不够高时提出来的新方法,其思路主要是寻找一个超平面,使得正反例之间的距离最大。本文借用支持向量机算法具体过程如下:

首先是训练上证指数集的选取,本文主要通过网络爬虫等数据挖掘软件,对某交易平台的上证指数数据进行挖掘,并对所得数据进行噪音等处理后作为训练指数集,即把所得上证指数数据作为训练指数样本集,设为:

$$D=\{(x_k,y_k)\mid k=1,2,\cdots,M,x_k\in R^n,y_k\in R\}$$

其中,x_k表示上证指数输入数据,y_k表示上证指数输出数据。

其次是支持向量机相关参数选取。针对上述所得训练指数集,结合支持向量机的知识,本文得到关于原始上证指数数据的权ω空间中的上证指数函数求解方程:

$$\min_{\omega,b,e}J(\omega,e)=\frac{1}{2}\omega^T\omega+\frac{1}{2}\gamma\sum_{k=1}^{M}e_k^2 \tag{1}$$

由于上述的权空间上证指数函数是非线性的,一般难以直接求解,目前对其求解的方法是直接转化为相对应的约束条件指数函数的最小值求解:

$$y_k=\omega^T\varphi(x_k)+b+e_k,(k=1,2,\cdots,M)$$

基金项目:国家自然科学基金项目(61662003);广西财经学院数量经济学创新团队2014年开放性课题(2014CX08);广西财经学院博士科研启动资金项目(BS201501);2016年应用统计硕士专业学位点资助学术研究项目(2016TJQN12)

作者简介:张晶华(1981-),男,广西玉林人,博士,广西财经学院信息与统计学院助理研究员,研究方向为模式识别、金融量化投资;莫文柯(1988-),女,广西南宁人,中国工商银行广西分行南宁民族支行职员,研究方向为理财、金融投资;甘宇建(1986-),男,广西玉林人,硕士,广西财经学院信息与统计学院助教,研究方向为模式识别、金融量化投资。

其中：$\varphi(.)$ 代表核空间 R^n 维到 R^m 维的映射关系函数；ω 是 R^m 维的权向量；e_k 是实数域范围内的误差变量；b 代表偏差量；γ 代表可调节的参变量。

为了求得约束条件价格函数的最小值或最优值，本文利用拉格朗日方法构造其拉格朗日上证指数函数方程为：

$$L(\omega,b,e,\lambda)=J(\omega,e)-\sum_{k=1}^{M}\lambda_k\{\omega^T\varphi(x_k)+b+e_k-y_k\} \tag{2}$$

其中，λ_k 为拉格朗日参数。

根据式(2)，利用拉格朗日求解的方法对拉格朗日上证指数函数 $L(\omega,b,e,\lambda)$ 关于所有参数 ω、b、e_k、λ_k 进行偏导数求解，并令所有的偏导数为零，然后消去参量 ω、e，从而得到最优分类上证指数函数方程，为了便于求解，用矩阵的形式给出其方程：

$$\begin{bmatrix}0 & I^T\\ I & \Omega+\dfrac{1}{\gamma}I\end{bmatrix}\begin{bmatrix}b\\ \lambda\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{3}$$

其中：$x=[x_1,\cdots,x_M]$，$y=[y_1,\cdots,y_M]$，$I=[1,\cdots,1]$，$\lambda=[\lambda_1,\cdots,\lambda_M]$，$\Omega_{kl}=\psi(x_k,x_l)=\varphi^T(x_k)\varphi(x_l)$，$(k,l=1,\cdots,M)$。

针对式(3)中的 Ω_{kl} 条件，本文利用 mercer 条件[15]可知，必有一个映射函数 φ 和一个核函数 $\psi(,)$ 使得等式 $\psi(x_k,x_l)=\varphi^T(x_k)\varphi(x_l)$ 成立。

综上所述，本文得到上证指数函数估计的方程为：

$$y(x)=\sum_{k=1}^{M}\lambda_k\psi(x_k,x_l)+b \tag{4}$$

其中，α、b 和 λ_k 利用式(3)求解。若所求得的 λ_k 不为零，则其对应的指数样本集就可以作为支持向量机的训练指数集；若所求得的 λ_k 为零，可以适当采取改变松弛变量 γ 的取值或者改变核函数 $\psi(,)$ 的选取，再进行验算，直至 λ_k 不为零，此时所得的超平面就是最优分类面。

关于式(3)中的核函数 $\psi(,)$，其作用就是对原始上证指数数据的特征进行提取，并将其映射为一个高维的特征空间向量，便于原始上证指数样本分类。目前使用核函数的形式主要有 4 种，比如线性核函数等。

2　模型实验结果与分析

2.1　支持向量机模型

本文选取上海证券交易所 2010 年 4 月 22 日－2014 年 6 月 10 日共 1 000 条数据并对数据进行归一化处理，然后利用支持向量机对归一化的上证指数数据集进行学习训练，得到最优参数的预测模型；最后根据所得最优参数的预测模型对下一个上证指数变化进行预测，得到下一个上证指数变化的预测数据值，并与实际值作误差分析。为了寻找上证指数的变化趋势和 SVM 预测值与实际值之间的关系，本文将所得数据分 3 个模型进行研究。模型一：以昨天的开盘价、最低价、最高价、收盘价和成交量作为自变量，当天的收盘价作为因变量；模型二：以前 2 天的开盘、收盘价作为自变量，当天的收盘价作为因变量；模型三：以前 4 天的收盘价作为自变量，当天收盘价作为因变量。通过上述 3 个模型寻找影响上证指数变化的因素，并建立下一个上证指数预测模型。

不失一般性，设所得的上证指数原始数据为 $\{y_i,i=1,2,\cdots,N。N=1\ 000\}$，并利用 $x_i=\dfrac{y_i-\dfrac{1}{N}\sum_{i=1}^{N}y_i}{y_{max}-y_{min}}$（其中：$y_{max}=\max\{y_i,i=1,2,\cdots,1\ 000\}$，$y_{min}=\min\{y_i,i=1,2,\cdots,1\ 000\}$。）对其进行归一化化简，把上证指数全部数据化简在[0,1]的数，得到上证指数的时间序列为 $\{x_i,i=1,2,\cdots,N。N=1\ 000\}$

2.2　最佳回归参数选取

为了避免出现无解的情况，本文选取适当的惩罚参数 γ 及核函数 $\psi(,)$ 对式(1)－式(4)进行求解，得到最优参数值。本文利用 Matlab 编程来实现程序运行，在求解过程中，主要借用 Matlab 中 libsvm 工具箱中的 SVMcg For－Regress()函数来求解，其中惩罚参数 γ 及核函数 $\psi(,)$ 的取值范围均设置为[－7,7]，步进均取 0.9，得到最佳回归参数 $\gamma=0.48$、$\psi(,)=5.357$。

2.3　数据预测与分析

在上述得到最优参数的支持向量机模型基础上，选取 2014 年 6 月 11 日－2016 年 6 月 24 日共 500 条数据作为预测数据，得到 3 种模型对应的预测值与实际值之间的对比如图 1－图 9 所示。

本文选取相对误差(MAPE)和均方误差(RMS)作误差对比分析，具体如下：

$AE=|y'_i-y_i|(i=1,2,\cdots,N)$；$MAPE=\left|\dfrac{AE}{y_i}\right|(i=1,2,\cdots,N)$；

$$RMS=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y'_i-y_i}{y_i}\right)^2}(i=1,2,\cdots,N)$$

其中：y'_i 为预测值，y_i 为原始值。

利用这些误差进行对比分析，主要是对比用支持向量机预测上证指数的效果，同时，也可以寻找最佳的预测方案。

2.4　三种模型的预测结果与分析

(1)模型一以昨天的开盘价、最低价、最高价、收盘价和成交量作为自变量，当天的收盘价作为因变量，得到如图 1－图 3 所示结果。

实验结果表明，利用昨天的开盘价、最低价、最高价、收盘价和成交量作为自变量，当天的收盘价作为因变量，得到了下一个开盘日上证指数的预测值与实际值，进行误差分析得到均方误差为 0.051 5，平均相对误差为 0.034 8。

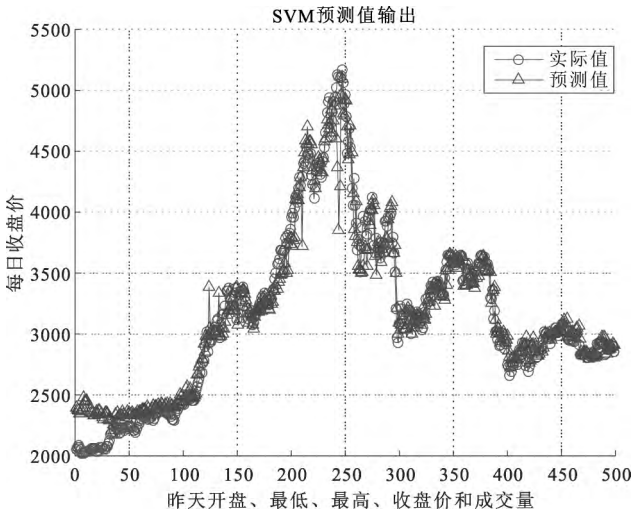

图 1 SVM 预测值与实际值比较

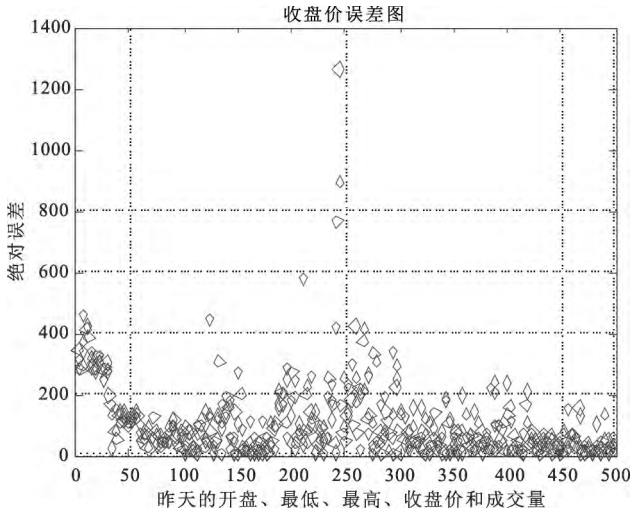

图 2 SVM 预测值与实际值均方误差比较

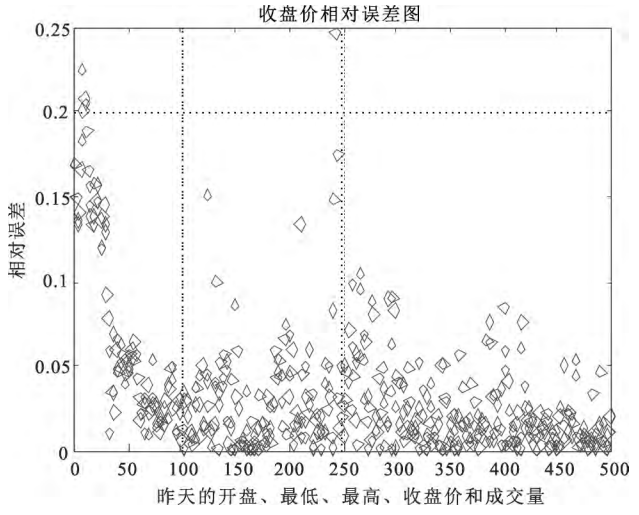

图 3 SVM 预测值与实际值相对误差比较

(2)模型二以前 2 天的开盘、收盘价作为自变量,当天的收盘价作为因变量,得到如图 4—图 6 所示结果。

实验结果表明,利用前 2 天的开盘、收盘价作为自变量,当天的收盘价作为因变量,得到了下一个开盘日上证指数的预测值与实际值,进行误差分析得到均方误差为 0.025 2,平均相对误差为 0.018 3。

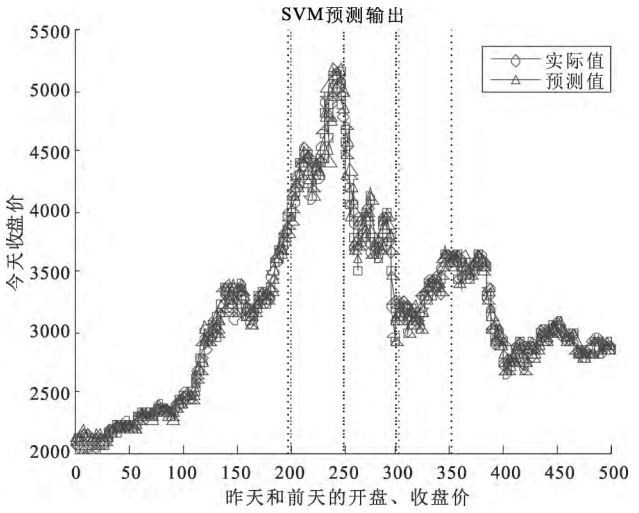

图 4 SVM 预测值与实际值比较

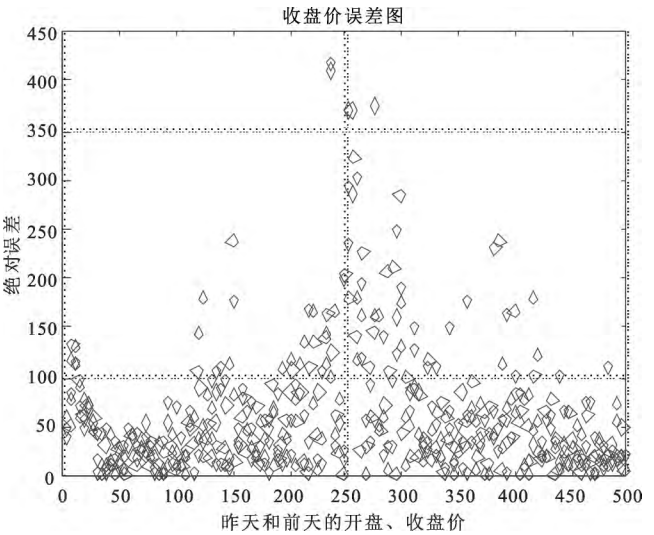

图 5 SVM 预测值与实际值均方误差比较

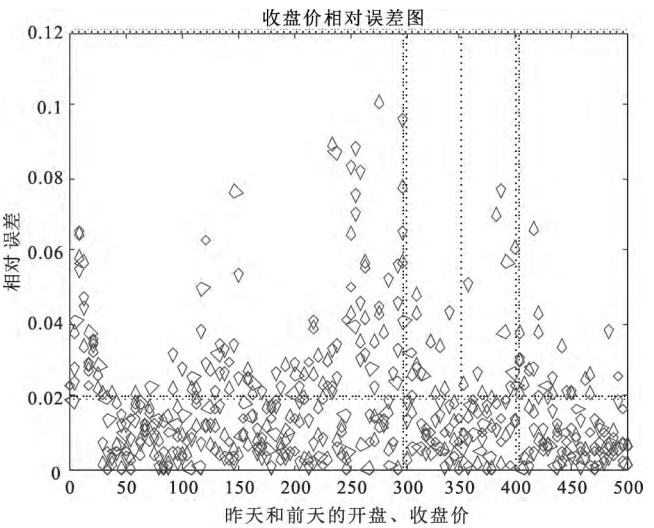

图 6 SVM 预测值与实际值相对误差比较

(3)模型三以前 4 天的收盘价作为自变量,当天的收盘价作为因变量,得到如图 7—图 9 所示结果。

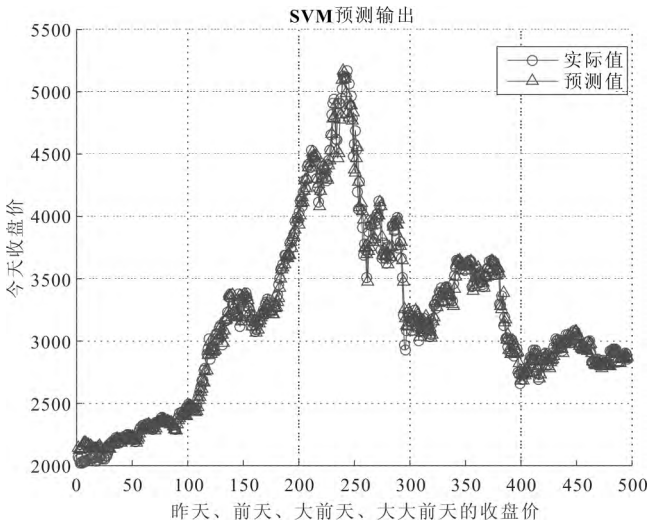

图 7　SVM 预测值与实际值比较

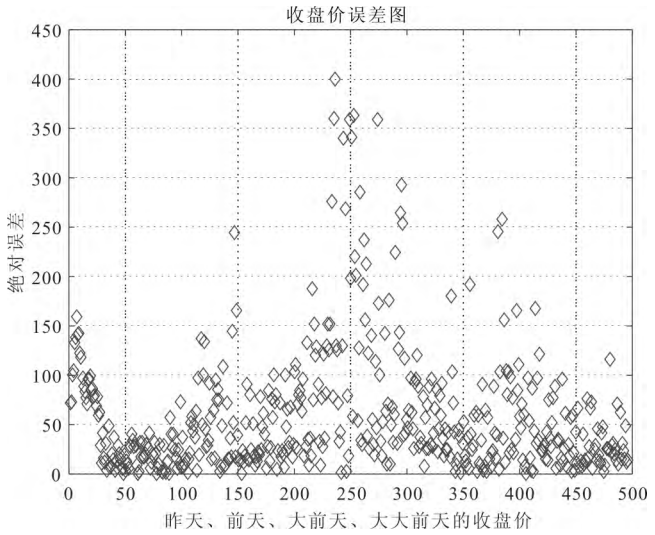

图 8　SVM 预测值与实际值的均方误差比较

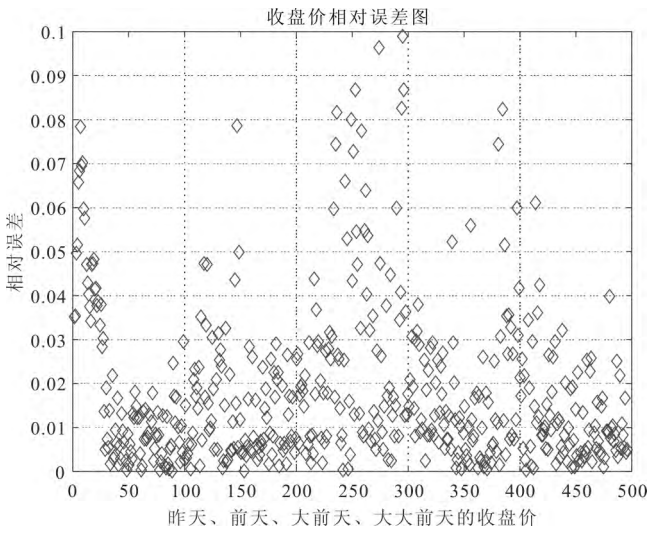

图 9　SVM 预测值与实际值相对误差比较

实验结果表明，利用前 4 天的收盘价作为自变量，当天的收盘价作为因变量，得到了下一个开盘日上证指数的预测值与实际值，进行误差分析得到均方误差为 0.026 1，平均相对误差为 0.018 6。

综上发现，采用昨天的开盘价、最低价、最高价、收盘价作为自变量时，得到的预测值与实际值的误差是最大的，而用前 4 天或者前 3 天作为自变量进行预测，两者所得的结果与实际的均方误差和平均相对误差相差不大。因此，当要预测后天的上证指数变化时，只需要前 3 天作为自变量输入即可，从而降低维数，便于编程计算。

本文利用支持向量机对下一个开盘日上证指数进行预测，得到良好结果，为投资者和持股票者提供了一定参考，可指导投资者适当调整投资，把握投资的盈利关系。同时，市场也可以根据上证指数之间的变化关系，提前对某些股票与投资作出预警，促使投资者合理投资，防止投资风险。但是股票市场变幻莫测，主要是受国家政策、国家事件以及人为等不确定因素所影响，进而影响到预测的准确度，因此对上证指数的变化预测还有待深入研究。

3　结语

本文借助支持向量机方法，对网络数据挖掘的上证指数进行分析，首先利用支持向量机有限样本的特性对所得上证指数数据进行分类，然后对所得上证指数分类集进一步预测下一个开盘日上证指数数值。本文用 3 种模型对下一个开盘日上证指数进行预测，实验结果表明，借助支持向量机方法，发现采用昨天的开盘价、最低价、最高价、收盘价作为自变量时，得到的预测值与实际值的误差最大，而用昨天、前天、大前天和大大前天或者昨天、前天、大前天作为自变量进行预测，两者所得的结果与实际的均方误差和平均相对误差相差不大，因此当要预测后天的上证指数变化时，只需要昨天、前天、大前天作为自变量输入即可，从而降低维数，便于编程计算。通过本文方法，预测到下一个开盘日上证指数的数值且所得数值与实际相差不大，为进一步寻找上证指数变化规律提供了一定的理论依据，由此可见，本文研究具有一定的实用价值。

参考文献：

[1] 杨毓，蒙肖莲. 用支持向量机(SVM)构建企业破产预测模型[J]. 金融研究，2006(10)：65-75.

[2] 邱玉莲，朱琴. 基于支持向量机的财务预警方法[J]. 统计与决策，2006(8)：153-155.

[3] 刘道文，樊明智. 基于支持向量机股票价格指数建模及预测[J]. 统计与决策，2013(2)：76-78.

[4] 张晨希，张燕平，张迎春，等。基于支持向量机的股票预测[J]. 计算机技术与发展，2006(6)：35-27.

[5] 杨一文，杨朝军. 基于支持向量机的金融时间序列预测[J]. 系统工程理论方法应用，2005(2)：176-181.

[6] 吴超鹏，吴世农. 基于价值创造和公司治理的财务状态分析与预测模型研究[J]. 经济研究，2005(11)：99-110.

[7] 杨成，程晓玲，殷旅江. 基于人工神经网络方法的上市公司股价预测[J]. 统计与决策，2005(12)：106-108.

（下转第 163 页）

飞行品质评估的试验方案,并通过具体评估试验实例表明该试验方案的可行性和有效性。该试验方案介绍的评估方法和评估流程为应用民机工程模拟器开展其它飞行员评估试验提供了参考,具有一定的工程实践意义。

表 2　纵向配平飞行员评估意见

飞行员	飞机操纵性	工作负担	HQR	其它意见
飞行员 A	无法控制	过分增加	2	响应良好,可以轻松地推拉杆实现飞机配平
	极难控制	明显增加		
	较难控制	轻微增加		
	正常控制√	不增加√		
飞行员 B	无法控制	过分增加	4	可能受纵向静稳定性影响,速度变化对应的杆位移过小,杆力梯度小,达到目标速度的精细操纵过程中有振荡趋势
	极难控制	明显增加√		
	较难控制√	轻微增加		
	正常控制	不增加		
飞行员 C	无法控制	过分增加	2	高度变化 3 000ft 内完成,杆力不大,配平速率合适
	极难控制	明显增加		
	较难控制	轻微增加		
	正常控制√	不增加√		
飞行员 D	无法控制	过分增加	3	配平速率合适,推杆后飞机响应有延迟,且油门的配平响应偏慢
	极难控制	明显增加		
	较难控制	轻微增加√		
	正常控制√	不增加		
飞行员 E	无法控制	过分增加	3	配平速率合适,油门的配平响应偏慢
	极难控制	明显增加		
	较难控制	轻微增加√		
	正常控制√	不增加		

参考文献:

[1] 向立学. 工程模拟器是现代飞机设计必不可少的工具[J]. 国际航空,1995(7):41-43.

[2] 周冬萍, 郡辉萍, 戚春明. 直升机工程模拟器在飞行控制律设计和操纵品质评估中的应用[J]. 航天与装备仿真, 2010(12): 315-319.

[3] 王维翰. 民用飞机工程模拟器与训练模拟器的区别[J]. 民用飞机设计与研究, 2003(1):1-5.

[4] 章伯定. 工程发展模拟器在飞机研制中的作用[J]. 飞行力学, 1989(2):1-10.

[5] 游崇林. 用工程飞行模拟器支援飞行试验[J]. 飞行力学, 1986(3):61-67.

[6] 李亚男, 刘彩志. 民用飞机飞控系统 MOC8 工程模拟器验证方法分析[J]. 民用飞机设计与研究, 2010(1):33-36.

[7] COOPER G E, HARPER R P. The use of pilot rating in the evaluation of aircraft handing qualities[R]. NASA TN D-5153, 1969.

[8] ROBERT P, HARPER JR, GEORGE E COOPER. Handing qualities and pilot evaluation[C]. Wright Brothers Lectureship in Aeronautics, 1984.

[9] JANN MAYER, TIMOTHY H COX. Evaluation of two unique side stick controllers in a fixed-base flight simulator[R]. NASA/TM-2003-212042, 2003.

[10] 高金源,等. 飞机飞行品质[M]. 北京: 国防工业出版社, 2003.

[11] 王立新. 适航性条例、飞行品质规范和设计准则[J]. 飞行力学, 2000,18(2):1-4.

[12] 张雅妮,李岩,金镭. 电子飞控飞机的飞行品质适航验证[J]. 飞行力学, 2012, 30 (2) : 118-119.

(责任编辑:黄　健)

Research on the Flight Test Program Appropriated for Civil Aircraft Control Law Design and Flying Quality Assessment

Abstract: In order to get better control law of civil aircraft, a flight test program appropriated for control law design and flying quality assessment with pilot in the loop has been proposed in this paper. Assessment tool (engineering simulator), method, process and control law optimization process have been introduced in detail, and the flight test assessment was conducted with 5 pilots in the loop. The results show that the flight test program proposed in this paper is feasible and effective. This conclusion has the real importance for practical engineering design and provides references for conducting other test assessment for aircraft design with pilot in the loop in the engineering simulator.

Key Words: engineering simulator; control law design; flying quality assessment; flight test program

(上接第 159 页)

[8] 蒋艳霞,徐程兴. 基于集成支持向量机的企业财务业绩分类模型研究[J]. 中国管理科学,2009,17(2):42-51.

[9] 李云飞,惠晓峰. 基于支持向量机的股票投资价值分类模型研究[J]. 中国软科学,2008(1):135-140.

[10] LAM M. Neural network techniques for financial performance prediction:integrating fundamental and techqical analysis[J]. Decision support systems,2004,37(4):567-581.

[11] KUAR P R,RAVI V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques——a review[J]. Expert Systems with Applications,2007,180(1):1-28.

[12] WU C H,TZENG G H,GOO Y J,et al. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy[J]. Expert systems with applications, 2005,32(2):397-408.

[13] HUA Z,WANG Y,XU X,et al. Predicting corporate financial distress based on integration of support vector machine and logistic regression[M]. Expert systems with applications,2006,33(2):434-440.

[14] KIM K. Financial time series forecasting using support vector machines[J]. Neurocomputing,2003,55(1-2):307-319.

[15] TAY F E H, CAO L J. Application of support vector machines in financial time series forecasting[J]. International Journal of Management Science,2001,29(4):309-317.

(责任编辑:孙　娟)