Computer Science and Application 计算机科学与应用, 2018, 8(4), 421-428
Published Online April 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.84046

Hans汉斯

The Stock Index Prediction Based on SVM

Cunli Zou, Lei Zhang, Yue Wang, Lin Cong

School of Mathematics, Liaoning Normal University, Dalian Liaoning
Email: 1789456398@qq.com

Received: Mar. 19th, 2018; accepted: Apr. 1st, 2018; published: Apr. 8th, 2018

Abstract

With the rapid development of China's economy, more and more people have joined the big family in the stock market. Because of the high noise and uncertainty of the stock market, it is very difficult to predict the price of the stock. And the more accurate prediction of the stock price is conducive to the investment of people. This paper selects 2676 trading days' data from Shanghai and Shenzhen 300 index from January 4, 2007 to December 29, 2017 in Guotai Junan intelligence software as the original analysis data, and builds support vector machine model and ARMA model to analyze and make short-term prediction. Results: the data prediction model of support vector machine and the actual data fitting degree is higher; the relative error is about 4%; the support vector machine model can make more accurate prediction of the price of the stock market, and can provide some reference for the study of Shanghai and Shenzhen stock market stock price.

Keywords

Shanghai and Shenzhen 300 Index, Support Vector Machine, ARMA Model, Stock Prediction, Data Normalization

基于SVM的股票指数预测

邹存利，张　蕾，王　玥，丛　琳

辽宁师范大学数学学院，辽宁　大连
Email: 1789456398@qq.com

收稿日期：2018年3月19日；录用日期：2018年4月1日；发布日期：2018年4月8日

摘　要

随着中国经济的飞速发展，越来越多的人加入到股市这个大家庭中来。由于股票市场具有高噪声、不确定等特性，使得股票的价格预测极为困难。而较为准确的预测股票价格，有利于人们的投资。本文选用

文章引用: 邹存利, 张蕾, 王玥, 丛琳. 基于 SVM 的股票指数预测[J]. 计算机科学与应用, 2018, 8(4): 421-428.
DOI: 10.12677/csa.2018.84046

国泰君安大智慧软件中2007年1月4日至2017年12月29日的沪深300指数中2676个交易日数据作为原始分析数据，通过建立支持向量机模型和ARMA模型进行分析并做出短期预测[1] [2]。实验结果：采用支持向量机模型的预测数据与实际数据的拟合度较高，相对误差控制在4%左右；说明支持向量机模型可以对股票市场做出更准确的价格预测，可以为沪深股票市场股票价格走势的研究提供一些借鉴[3] [4]。

关键词

沪深300指数，支持向量机，ARMA模型，股票预测，数据归一化

1. 引言

股票市场瞬息万变，风险很高，而对股票指数的预测可以为我们从整体上把握股市的变动提供有效的信息。沪深 300 指数是沪深证交所联合发布，以流动性和规模作为两大选样的根本标准，是一个能反映 A 股市场价格整体走势的指标。所以对于沪深指数的预测具有十分重要的意义。基于支持向量机的优良性能，考虑将其应用于股市指数的预测[5]。

支持向量机于 1995 年由 Cortes 和 Vapnik 等人正式发表，由于其在文本分类任务中显示出卓越性能，很快成为机器学习的主流技术，并直接掀起了“统计学习”在 2000 年前后的高潮。Vapnik 等人从六、七十年代开始致力于此方面研究，直到九十年代才使抽象的理论转化为通用的学习算法，其中核技巧才真正成为机器学习的通用基本技术。

支持向量机是建立在统计学习 VC 维理论和结构风险最小原理基础上的，根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳途径，以期获得最好的推广能力的新型神经网络。随着 ε 不敏感损失函数的引入，支持向量机从原来只处理分类问题逐步扩展到也能胜任回归[6] [7]。尤为值得一提的是通过构造核函数能在无需知道映射具体形式的情况下将非线性问题映射到高维线性空间，并对支持向量机的预测性能起到决定性作用[8] [9]。本文希望根据金融时间序列的有关特性，结合支持向量机引入核函数来提高支持向量机算法性能，增强统计学习方法在金融时间序列预测和特征数据分类中的分析能力[10]。

设选定的训练样本表示为：$\{(x_1,y_1),\cdots,(x_n,y_n)\}\in R^n\times R$，在 SVR 中，通过线性回归来找到函数 $f:R^n\to R$：

$$f(x)=\langle w,x\rangle+b,\ w\in R^n,b\in R.$$

并满足如下的优化问题：

$$\min_{w,b}\frac{1}{2}\|w\|^2$$

s.t.

$$\begin{cases}y_i-\langle\omega,x_i\rangle-b\le\varepsilon\\ \langle\omega,x_i\rangle+b-y_i\le\varepsilon\\ i=1,2,\cdots,l\end{cases}$$

其中，$\varepsilon > 0$ 表示线性近似的参数。

股票指数的预测问题更多是非线性回归，将数据用映射ϕ映射至高维空间的方法进行处理，从而在高维空间中可进行相应的线性回归，用核函数代替所涉及到的内积运算。误差的存在是被允许的。这里需要引入松弛变量ξ_i, ξ_i^*，而损失函数则采用ε-不敏感函数，其定义为：

$$|\xi|_\varepsilon := \begin{cases} 0, & \text{如果} |\xi| \le \varepsilon \\ |\xi| - \varepsilon, & \text{否则} \end{cases}$$

容易得到下列优化问题：

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right)$$

s.t.

$$\begin{cases} y_i - \langle w, \phi(x_i)\rangle - b \le \varepsilon + \xi_i \\ \langle w, \phi(x_i)\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases}.$$

转化为对偶形式：

$$\max W\left(\alpha, \alpha^*\right) = -\frac{1}{2}\sum_{i,j=1}^{l}\left(\alpha_i, \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)K\left(x_i, x_j\right) + \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)y_i - \sum_{i=1}^{l}\left(\alpha_i + \alpha_i^*\right)\varepsilon$$

这里
$\alpha, \alpha^* \ge 0$，约束为：

$$\sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right) = 0, \alpha_i + \alpha_i^* \le C, i = 1, \cdots, l$$

$$\alpha_i, \alpha_i^* \ge 0, i = 1, \cdots, l.$$

用求解此二次规划的问题得到α 和α^*的值，综上所述可以得到的回归方程如下：

$$f(x) = \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)K\left(x, x_i\right) + b.$$

2. 模型建立

模型假设：假设股票市场指数每日开盘价格与前一天的开盘价、最高价、最低价、收盘价、成交量、成交金额具有一定的相关性，即把前一天的开盘价、最高价、最低价、收盘价、成交量和成交金额作为自变量，则因变量为当天的开盘价[11]。

如图 1 所示为模型实现流程图：

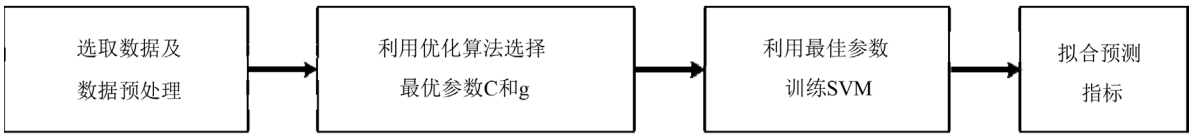

Figure 1. Model building flow chart
图 1. 模型建立流程图

3. 模型实现

3.1. 数据选取

本研究获取了从 2007 年 1 月 4 日至 2017 年 12 月 29 日的共计 2676 个交易日的交易数据作为分析数据。假设沪深指数每日的开盘价与上个交易日的交易情况有关系，而股票交易情况主要反映在每天的开盘价、最高价、最低价、收盘价、成交量、成交金额等指标上，所以本文选择了能够反映交易情况的开盘价、最富价、最低价、收盘价、交易量、交易额六个指标，用前一天的六个指标数据情况来回归预测当天的开盘价。

选取第 1 个到第 2675 个交易日每天的开盘价，最高价、最低价、收盘价、成交量和成交金额 6 个指标作为模型的自变量，该模型的因变量是第 2 个到第 2676 个交易日每天的开盘价，如下图 2 所示为沪深 300 指数每天的开盘价[12]。

3.2. 数据预处理

从样本的历史数据可以观察到，选取的 6 个特征变量在数量级上差别较大，为了防止数据溢出、大值主导小值，提高模型的预测精度，本研究考虑对时间序列数据进行中值归一化处理，其转化函数为：

$$x_i = \frac{x_i - x_{\text{mid}}}{x_{\max} - x_{\min}}$$

x_i 表示归一化的数据中的第 i 个数值，$x_{\max}$ 为每项归一化数据中的最大值，$x_{\min}$ 为每项归一化数据中的最小值，x_{mid} 为区间中间值这里取 0 为中间值,这里的区间取得是[−2, 2]。

将处理后的每日开盘价归一化结果如图 3 所示。

3.3. 参数优化选择

既往研究证实了不同核函数对支持向量机预测性能影响不是很大，但核函数参数的选择却严重的影响支持向量机的推广泛化能力。从已有的研究结果来看，最常用的核函数高斯核函数在大多数情况下都获得了很好的预测效果，因此本研究主要根据既往研究基础上，引用高斯核函数，利用网格搜索法算法对高斯核函数的参数 g 和惩罚参数 c 进行优化选择。

首先采用网格搜索法对高斯核参数 g 和惩罚参数 c 这两个参数进行优化，在最佳参数的基础上训练支持向量机模型。在这里的两个参数用网格搜索法进行优化的基本思想是将参数 c 和 g 的可行区间(从小

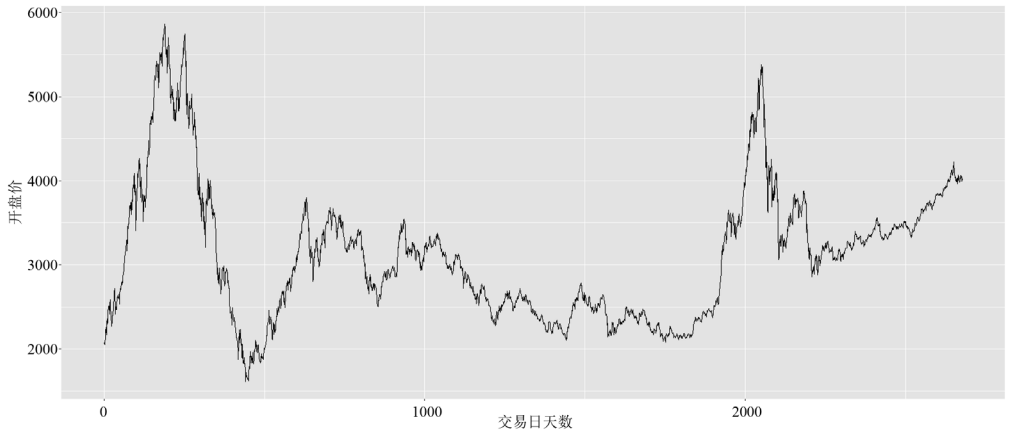

Figure 2. The daily opening price of the Shanghai and Shenzhen index (2007.1.4~2017.12.29)
图 2. 沪深指数每日开盘价(2007.1.4~2017.12.29)

到大)分成一系列小区间，计算出各参数值的组合所对应的误差值并逐一比较选择最佳，从而求得该区间内的最小误差值和它所对应的最优参数值，这种方法使得我们搜索到的解基本上是全局最优解，避免了重大误差的存在。利用网格搜索法对支持向量机参数寻优，并得到最佳参数为：c = 0.5；g = 2。大体如图 4 所示为网格搜索法参数选择结果[13]。

网格搜索法可以找到在交叉验证意义下的最高分类准确率，也就是全局最优解，但若想在更大范围内寻找最佳参数 c 和 g 会很费劲，采用启发式算法不必搜索网格内的所有点，也能找到最优解。

3.4. 利用最佳参数训练 SVM 模型做拟合预测

通过以上参数最优化选择参数 c 和 g 训练支持向量机模型，并根据模型对原始数据进行模型拟合及预测，如图 5 所示利用网络搜素法优化得到的沪深指数开盘价回归预测和原始数据对比情况。并通过基于网络搜索法最终得到的模型拟合结果是，均方误差 MSE = 2.12583e−5，相关系数 R = 99.9521%。

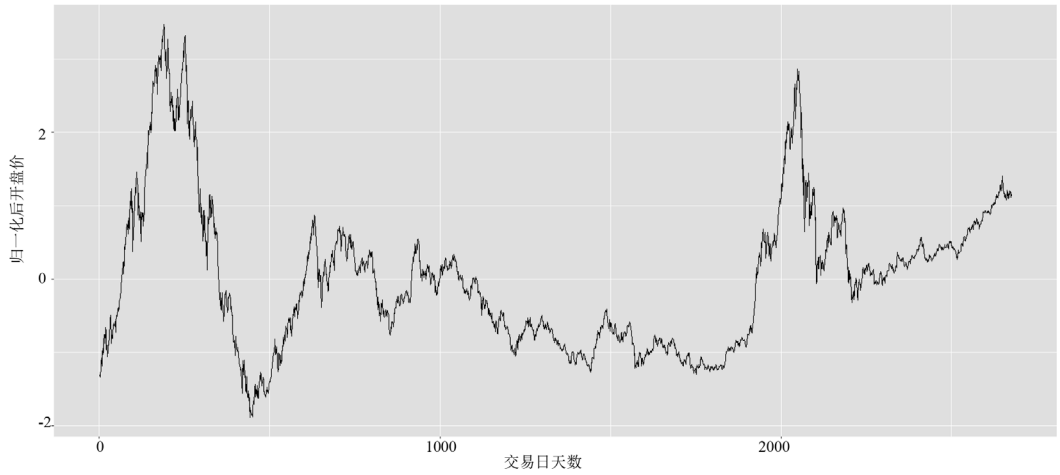

Figure 3. The daily opening price of the Shanghai and Shenzhen index after the normalization of the Shanghai and Shenzhen index (2007.1.4~2017.12.29)
图 3. 沪深指数归一化后每日开盘价(2007.1.4~2017.12.29)

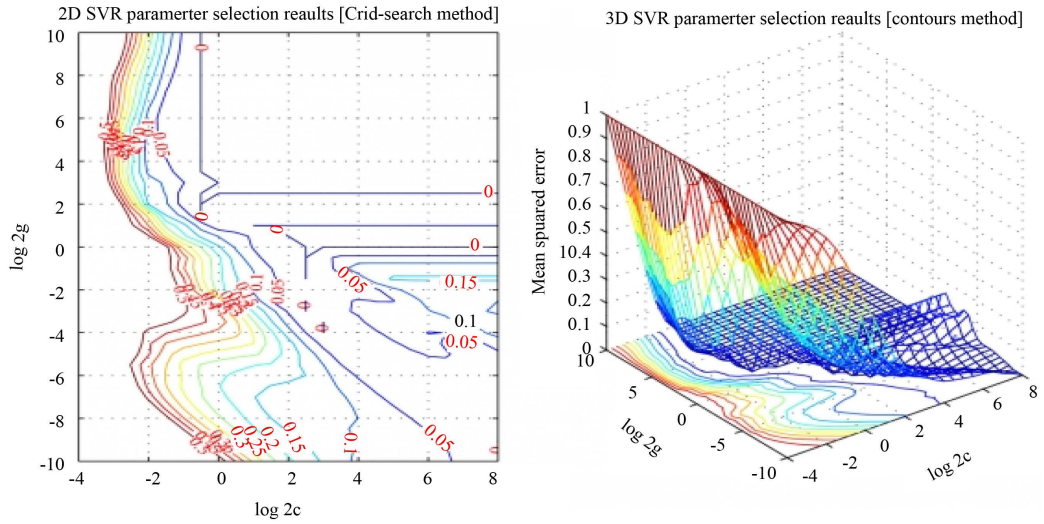

Figure 4. Optimal parameter graph of grid search method
图 4. 网格搜索法最优参数图

另外，预测数据与实际数据之间差异可以通过以下误差图来描述，从图 6 中可以观察到回归预测数据与原始数据的开盘价误差量基本上维持在 70 以内。其次，经过相对误差分析可知，相对误差基本控制在 4%以内，回归的预测精度较高，可显示该方法预测股票指数的效果较好[14]。

4. 实验结果

我们将基于网格搜索法优化的支持向量机模型和时间序列 ARMA 模型的分析结果分别用于沪深 300 指数回归预测，将最终得到的拟合结果进行比较，图 7 所示为两种模型拟合预测结果和原始数据的比较。

实验选取的原始数据前 1000 个交易日数据当作训练样本输入，第 1001 个交易日至 1010 个交易日的收盘价格当作测试样本输出。通过实验模型的拟合，ARMA 模型与支持向量机模型的预测情况。

从图 7 可知，ARMA 模型的预测趋势跟原始数据比较，整体的预测是失效的，并没有做出精准地预测。而对比的基于支持向量机模型预测趋势和原始数据比较，整体的预测准确性较高，并基本与原始数据的上涨或下降趋势基本相同。这验证了将支持向量机应用于股价预测问题很有效。而对于复杂的金融时间序列数据来说，传统的 ARMA 模型已经不能够准确地预测真实的数据走势，而支持向量机股价预测

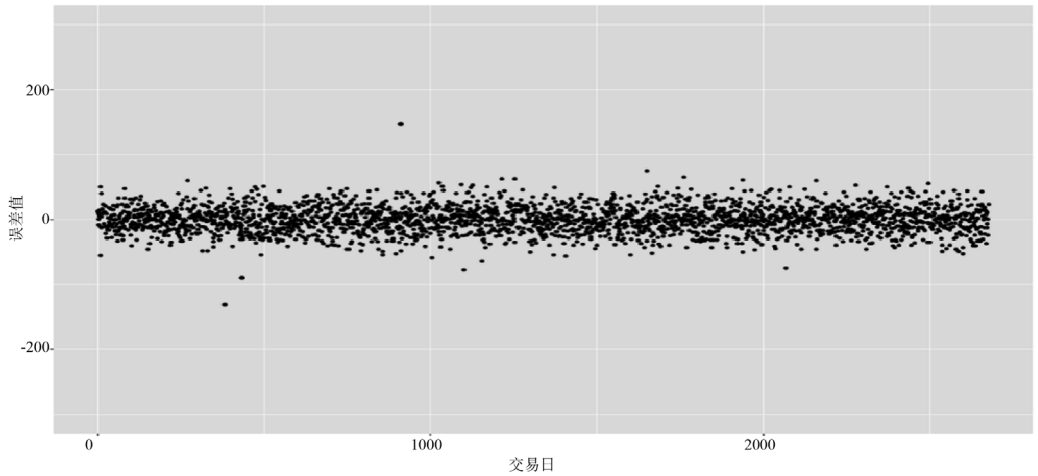

Figure 5. Prediction of error distribution of data and actual data
图 5. 预测数据和实际数据误差分布图

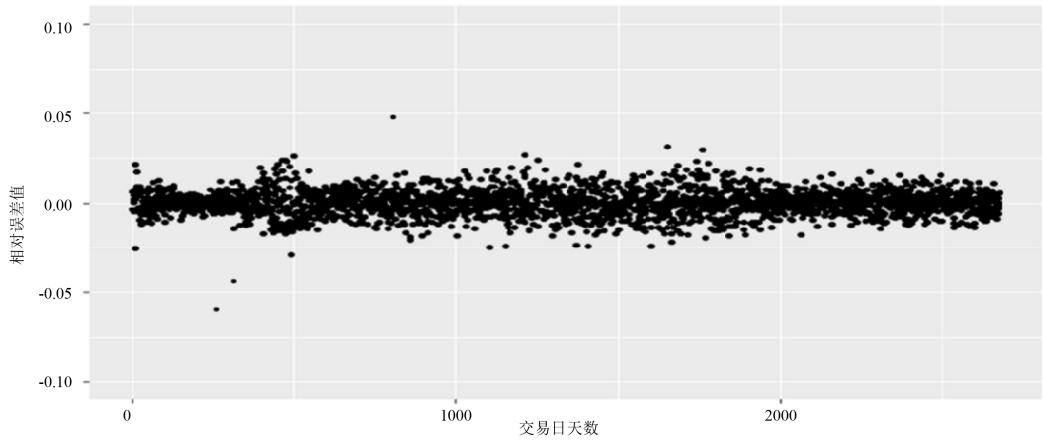

Figure 6. The distribution of relative error between predicted and actual data
图 6. 预测数据和实际数据的相对误差分布图

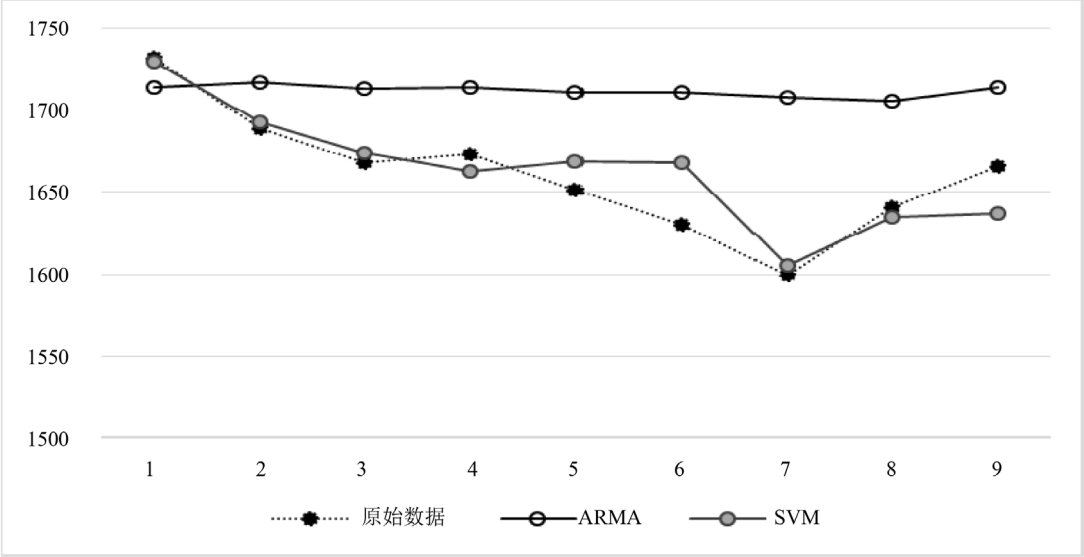

Figure 7. Comparison of ARMA model and SVM fitting model
图 7. ARMA 模型与 SVM 拟合模型的比较

模型具有良好的非线性特征，可以很有效地拟合复杂函数，并对其进行准确预测[15]。

5. 结论

本文选用从国泰君安大智慧软件中下载的 2676 个交易日沪深 300 指数数据作为研究数据。采用建立支持向量机模型和 ARMA 模型的方法对数据进行分析研究，经过对原始数据预处理、最优参数选择、分别进行模型训练及预测、原数据与预测结果的拟合度进行对比分析等步骤来实证分析。结果表明基于支持向量机的模型的预测数据与原始数据拟合度较高；支持向量机模型的预测效果要远远优于 ARMA 模型的预测效果，支持向量机核函数参数的选取很大程度上决定了模型的预测精度，而传统的时间序列模型预测效果不是很好；SVR 在一定时期内，可以为投资者进行投资提供一定的依据。

基金项目

辽宁省自然科学基金资助项目(201602461)。

参考文献

[1] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.
[2] 冯帆, 倪中新. 基于支持向量机的高频金融时间序列预测[J]. 应用数学与计算数学学报, 2017, 31(3): 265-274.
[3] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
[4] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
[5] 孙延风, 梁艳春, 姜静清, 等. 金融时间序列预测中的吉林大学学报[J]. 信息科学版, 2004, 22(1): 48-52.
[6] 汤荣志. 数据归一化方法对提升 SVM 训练效率的研究[D]: [硕士学位论文]. 济南: 山东师范大学, 2017.
[7] 张竞文, 李洋, 孙德山. 时间序列模型在辽宁省 GDP 预测中的应用[J]. 中国集体经济, 2017(7): 61-62.
[8] 曹兆龙. 基于支持向量机的多分类算法研究[D]: [硕士学位论文]. 上海: 华东师范大学, 2007.
[9] 杨琦, 曹显兵. 基于 ARMA-GARCH 模型的股票价格分析与预测[J]. 数学的实践与认识, 2016, 46(6): 80-86.
[10] 孙全, 朱江. 基于遗传神经网络的股票价格短期预测[J]. 计算机工程与应用, 2002, 38(5): 237-239.
[11] 钱峰, 连涛, 吴嘉兴. 一种基于支持向量回归的互联网端到端延迟预测算法[J]. 计算机应用研究, 2012, 29(5):

1850-1853.

[12] 孙德山, 王玥. 基于多种统计分类方法的股票趋势预测[J]. 辽宁师范大学学报(自然科学版), 2017, 40(4): 440-444.

[13] Tang, L.B. and Sheng, H.Y. (2009) Forecasting Stock Returns Based on Spline Wavelet Support Vector. *CINC*09 *International Conference on Computational Intelligence and Natural Computing*, Wuhan, 6-7 June 2009, 383-385.

[14] 郝知远. 基于改进的支持向量机的股票预测方法[J]. 江苏科技大学学报(自然科学版), 2017, 31(3): 339-343.

[15] 刘廷. 基于支持向量机的股票预测[J]. 信息通信, 2015(8): 15-16.