Hans Journal of Data Mining 数据挖掘, 2016, 6(1), 1-8
Published Online January 2016 in Hans. http://www.hanspub.org/journal/hjdm
http://dx.doi.org/10.12677/hjdm.2016.61001

Hans汉斯

Forecast Analysis of Shanghai Composite Index Based on Machine Learning Method

Rengkang Wu

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: wurengkang@163.com

Received: Dec. 25th, 2015; accepted: Jan. 11th, 2016; published: Jan. 14th, 2016

Abstract

The Shanghai composite index is an important index that general investors pay close attention to. Shanghai composite index, which not only reflects the basic situation of the stock market in our country, but also takes an important guiding role to our economy. Prediction of Shanghai composite index and trend analysis plays an important role to stabilize market and guide investors. And stock market data are a typical nonlinear system; traditional statistical forecasting methods predict a low accuracy. In this paper, we use R software comprehensively and combine with the latest six kinds of methods in machine learning field, decision tree, boosting, bagging, random forests, support vector machine (SVM), neural network to train the training set, respectively, get the corresponding model. And set up the corresponding ten-fold cross validation to calculate the prediction mean square error of each method for comparison. Select the model with better effect, and make a visualized comparison between prediction data and real data. Analysis shows that the results of random forests, SVM are more fitting, and have high precision.

Keywords

Shanghai Composite Index, Machine Learning, Random Forests, SVM

基于机器学习方法的上证综合指数的预测分析

吴仍康

云南财经大学统计与数学学院，云南 昆明
Email: wurengkang@163.com

文章引用: 吴仍康. 基于机器学习方法的上证综合指数的预测分析[J]. 数据挖掘, 2016, 6(1): 1-8.
http://dx.doi.org/10.12677/hjdm.2016.61001

收稿日期：2015年12月25日；录用日期：2016年1月11日；发布日期：2016年1月14日

摘　要

上证综合指数是广大投资者关注的重要指数。上证综合指数不仅反映了我国股票市场的基本状况，同时对我国经济走向也具有重要的导向作用。对上证综合指数的预测分析以及趋势研判对稳定市场、引导投资者具有重大意义。而股票市场数据是典型的非线性系统，传统统计学预测方法在处理时预测精度较低。本文综合运用R软件并结合目前机器学习领域最新的六种方法——决策树、boosting、bagging、随机森林、支持向量机、神经网络分别对训练集进行训练，得到相应模型。并建立相应的十折交叉验证集计算出每种方法的预测均方误差进行对比。筛选出效果较好的模型，并对预测数据与真实数据进行数据可视化对比。对结果分析可知，随机森林、支持向量机两种机器学习方法拟合效果较好，且精度高。

关键词

上证综合指数，机器学习，随机森林，支持向量机

1. 研究背景

从目前学者们对我国股票市场特征的研究来看，许多学者认为我国股票市场还没有达到弱势有效，也就是说用股票交易的历史信息在一定程度上可以预测其未来的走势。同时从投资经验来看，中国的股票市场在一定程度上，特别是在短期内，是具有一定的可预测性的。这里说的是股价的预测，并没有说可以100%准确遇见，而是说其短期趋势具备一定的可预测性。

股票市场作为经济的重要指标，不仅受到广大投资者的关注，也受到各国政府的重视。股票市场是金融市场的重要组成部分，是资本配置的重要手段，对推动国民经济的发展起着举足轻重的作用。同时，股票也是市场经济的产物，股票的涨与跌是由市场多空力量共同作用的结果，股市的暴涨暴跌会严重扰乱国家的金融秩序，甚至导致经济危机。深刻理解股票市场的运行规律，预测分析股票价格的未来走势，无论对广大投资者降低投资风险还是宏观经济管理部门的宏观调控，保障我国证券市场的健康持续发展，都有重要的意义。由此可见，通过对股票趋势预测的研究，不仅可以为广大投资者投资决策提供依据，更可以为国家制定相关经济政策提供参考。

近年来，广大研究人员把随机过程和模糊数学以及信息、控制、人工智能、应用数学等专业大量应用与股票价格的研究上，维纳过程、马尔科夫方法、“黑盒子”理论的灰色系统预测、蒙特卡洛模拟，甚至国际上前沿的混沌理论。也有一些学者从博弈的角度研究股票的价格波动，即庄家、散户、机构，甚至政府之间进行双方和多方的博弈。人们对以股票价格为代表的各种金融资产价格的分析预测方法的研究，从股票市场和各种金融市场的出现开始，就没有停止过，众多金融、计算机学界的专家、学者对此投入了极大的热情，并由此产生了许多优秀的证券预测方法。

1.1. 技术分析方法

在股票市场应用比较多的就是技术分析方法。所谓的技术分析方法，主要根据市场的一些资料(如成交价、成交量)，运用图标、形态、指标等分析手段，对证券价格的发展趋势进行各种有针对性的分析研究，研究市场过去和现在的行为反应，最终判断整个股市和某个股价未来的大致变化趋势[1]。随着技术分析方法的发展，陆续出现了一些技术理论，主要有道氏理论、季节理论、相反理论、相对理论、多数

决定理论、投资者期望理论、江恩理论、亚当理论、平方根理论、四度空间理论、甘氏角度线方法、K线理论等[2]。技术分析方法主要是基于市场资料以及信息的判断，没有严格的数理模型，呈现较大的个体差异化，预测较为模糊。

1.2. 传统统计学预测方法

在传统统计学预测方法中主要分两大流派：

第一，基于拟合以及最小二乘原理建立实验体的各种回归、自回归、混合回归的模型进行预测。在此领域国内外学者对在股票预测上的应用进行了大量研究。如 Nelder JA 等人通过放松经典线性模型的假设，提出了广义线性模型，Aaron，Li 和 Duan 对假设条件进一步放松，提出了一般回归模型，他们的研究极大地丰富了回归分析的理论，对该领域的发展产生了深远的影响[3]-[6]。

第二，时间序列预测方法。其主要是通过分析股票数据构成的时间序列，根据时间序列反映出来的发展过程和趋势，进行类推进而可以预测下一段时间可能达到的水平。

这些传统的统计学预测方法操作简单，很容易求得预测值，但是对数据的分布以及其平稳性有着较为严格的要求。又由于股票市场数据往往呈现非线性，因此，预测结果往往不太精确，不具备较强的参考性。

1.3. 基本面分析方法

基本面分析是指证券分析师根据经济学、金融学、财务管理及投资学等基本原理，对决定证券价值及价格的基本要素，如宏观经济指标、经济政策走势、行业发展状况、产品市场状况、公司销售和财务状况等进行分析，评估证券的投资价值，判断证券的合理价位，剔除相应的投资建议的一种分析方法[7] [8]。

基本面分析重点研究股票的内在价值一上市公司经营状况的真实反映。这中分析方法认为股票价格在市场上的频繁波动是受多种内外因素影响的，但股票的市场价格总是围绕这内在价值变化，市场价格和内在价值之间的差距最终会被市场纠正，因此市场价格低于(或高于)内在价值之日，便是买(卖)机会到来之时。

1.4. 机器学习方法

随着数据挖掘以及人工智能技术的发展，国内外的研究学者提出了许多新的机器学习算法，包括：决策树、boosting、bagging、随机森林、支持向量机、神经网络等等。但是，在当前的研究中众多学者多为选择性的实用众多机器学习方法中的一种或两种，并没有讨论其所选方法的合理性以及优越性。更没有将其与其它机器学习方法的学习效果进行对比。

基于此，本文将综合运用以上 6 种机器学习方法(决策树、boosting、bagging、随机森林、支持向量机、神经网)对我国上证综合指数进行学习。针对每一种机器学习方法我们都建立十折交叉验证集，计算其预测的均方误差。由此来对各种机器学习方法进行对比以及选择。

2. 各类机器学习回归方法及其交叉验证

2.1. 数据预处理

现在对 1990 年 12 月 19 日~2015 年 5 月 29 日上海证券综合指数的日 K 线图的数据进行分析。本文仅考虑由当日开盘指数、最高指数、最低指数、以及收盘指数来对下一日的开盘指数进行预测。因此，我们剔除下载数据中其他无关的变量数据，保留我们所需要的数据变量。

2.2. 建立十折交叉验证集

十折交叉验证(10-fold cross-validation)，用来测试算法准确性。是常用的测试方法。将数据集分成十

分，轮流将其中 9 份作为训练数据，1 份作为测试数据，进行试验。每次试验都会得出相应的正确率(或差错率)。10 次的结果的正确率(或差错率)的平均值作为对算法精度的估计，一般还需要进行多次 10 折交叉验证(例如 10 次 10 折交叉验证)，再求其均值，作为对算法准确性的估计。

2.3. 各类机器学习方法 R 软件的实现

2.3.1. 决策树

在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。数据挖掘中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测。故这里应用决策树方法对上证综合指数进行分析和预测。

#分类树回归 R 语言处理主要程序：

library(rpart);library(rpart.plot);(a=rpart(v5~.,w)) ;rpart.plot(a,type=2) 其中决策树图，见图 1。

决策树十折交叉验证均方误差为：0.02662164。

2.3.2. Boosting

Boosting 是一种提高任意给定学习算法准确度的方法。它的思想起源于 Valiant 提出的 PAC (Probably Approximately Correct)学习模型。Boosting 方法也是一种用来提高弱分类算法准确度的方法，这种方法通过构造一个预测函数系列，然后以一定的方式将他们组合成一个预测函数。他是一种框架算法，主要是通过对样本集的操作获得样本子集,然后用弱分类算法在样本子集上训练生成一系列的基分类器。他可以用来提高其他弱分类算法的识别率。

#boosting 回归 R 语言处理主要程序：

library(mboost); gg1=v5~btree(v1)+btree(v2)+btree(v3)+btree(v4);a=mboost(gg1,data =w)

Boosting 十折交叉验证均方误差为：0.002187182。

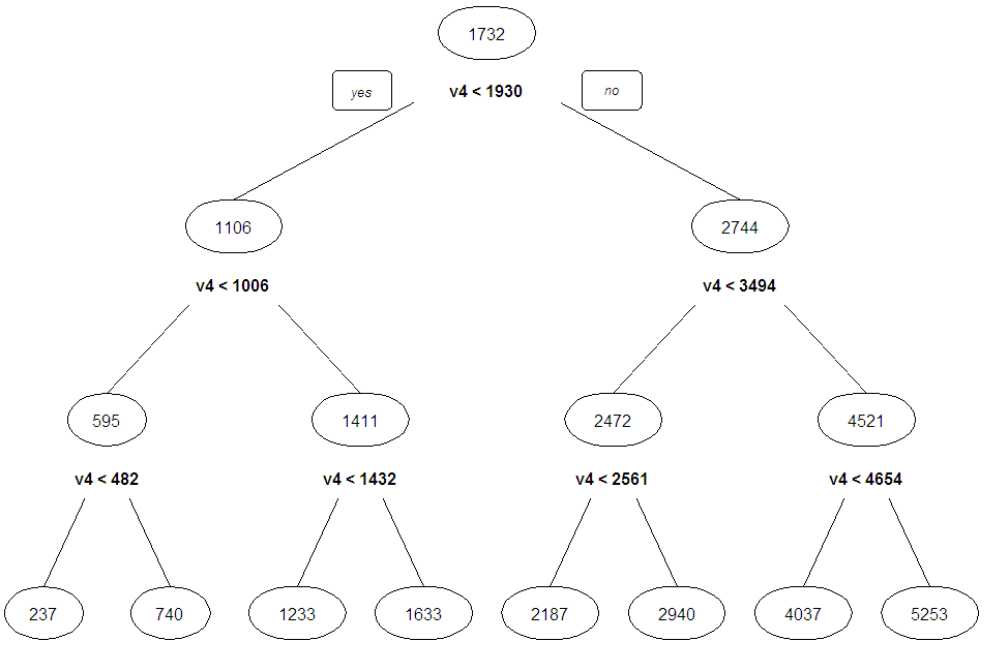

Figure 1. Decision tree diagram
图 1. 决策树图

2.3.3. Bagging

Bagging 是一种比 Boosting 简单的组合方法。在 bagging 中，就是不断放回地对训练样本进行再抽样(自助法样本)，每次再抽样的样本量和原来样本量一样。对每个自助法样本，都建立一棵回归树，最终，对于任何一个观测值，每棵树都给出一个预测值，最终的预测值为这些值的简单平均。Bagging 能用来提高学习算法准确度。

#bagging 回归 R 语言处理主要程序：

library(ipred);set.seed(4410);a=bagging(v5~.,w)

bagging 十折交叉验证均方误差为：0.02317248。

2.3.4. 随机森林

随机森林(random forest)另一种组合方法，也是由随机放回地再抽样的样本形成的决策树组成的，其特点是这些决策树的每一节点的分割变量不是有所有的自变量竞争产生，而是由随即选取的少数变量产生，因此不仅产生每棵决策树的样本是随机的，每棵树的每个节点的产生也是随机的。这些随机产生的决策树数目很大，因此成为随机森林。随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。

#随机森林回归 R 语言处理主要程序：

library(randomForest) ;set.seed(10);a=randomForest(v5~.,w,importance=TRUE,proximity=TRUE)

随机森林十折交叉验证均方误差为：0.0005628885。

2.3.5. 支持向量机

支持向量机(SVM)是 20 世纪 90 年代初 Vapnik 等人根据统计学习理论提出的一种新的机器学习方法，它以结构风险最小化原则为理论基础，通过使当地选择函数子集及该子集中判别函数，是学习机器的实际风险达到最小，保证了通过有限训练样本得到的小误差分类器，对独立测试集的测试误差仍然较小。

其突出的优点表现在：1) 基于统计学习理论中结构风险最小化原则和 VC 维理论，具有良好的泛化能力，即由有限的训练样本得到的小的误差能够保证使独立的测试集仍保持小的误差。2) 支持向量机的求解问题对应的是一个凸优化问题，因此局部最优解一定是全局最优解。3) 核函数的成功应用，将非线性问题转化为线性问题求解。4) 分类间隔的最大化，使得支持向量机算法具有较好的鲁棒性。由于 SVM 自身的突出优势，因此被越来越多的研究人员作为强有力的学习工具，以解决模式识别、回归估计等领域的难题。在本文针对股票预测方法属于回归的范畴。

#支持向量机回归 R 语言处理主要程序：

library(e1071);a=svm(v5~., data = w,kernal="sigmoid")

支持向量机十折交叉验证均方误差为：0.001070409。

2.3.6. 神经网络

人工神经网络，是 20 世纪 80 年代以来人工智能领域兴起的研究热点。它从信息处理角度对人脑神经元网络进行抽象，建立某种简单模型，按不同的连接方式组成不同的网络。在工程与学术界也常直接简称为神经网络或类神经网络。神经网络是一种运算模型，由大量的节点（或称神经元）之间相互联接构成。每个节点代表一种特定的输出函数，称为激励函数。每两个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重，这相当于人工神经网络的记忆。网络的输出则依网络的连接方式，权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

#神经网络回归以及绘图 R 语言处理主要程序：

```
library("neuralnet");library("MASS");v=w;v$v5=v$v5/max(w[,5]);set.seed(1010)#w$v5<=max(w[,5])
b=neuralnet(v5~v1+v2+v3+v4,data=v,err.fct="sse";hidden=6,linear.output=FALSE);plot(b)
```

其中神经网络图，见图 2。

神经网络十折交叉验证均方误差为：0.3003667755。

2.4. 各类机器学习十折交叉验证结果

如表 1 所示，为六种机器学习方法十折交叉验证的均方误差，由表分析可知在各类机器学习方法中“随机森林”、“支持向量机”、“boosting”、“bagging”的学习效果较好。接下来对数据进行预测时我们将选用这四种机器学习方法。

3. 预测数据与真实数据的可视化对比

如图 3 所示，通过对以上四组数据可视化对比图我们可以清晰看出，四种机器学习方法的预测值都能较好的拟合真实上证指数数据。但是“bagging”以及“boosting”拟合的效果稍弱，“支持向量机”以及“随机森林”拟合的效果十分契合真实数据。

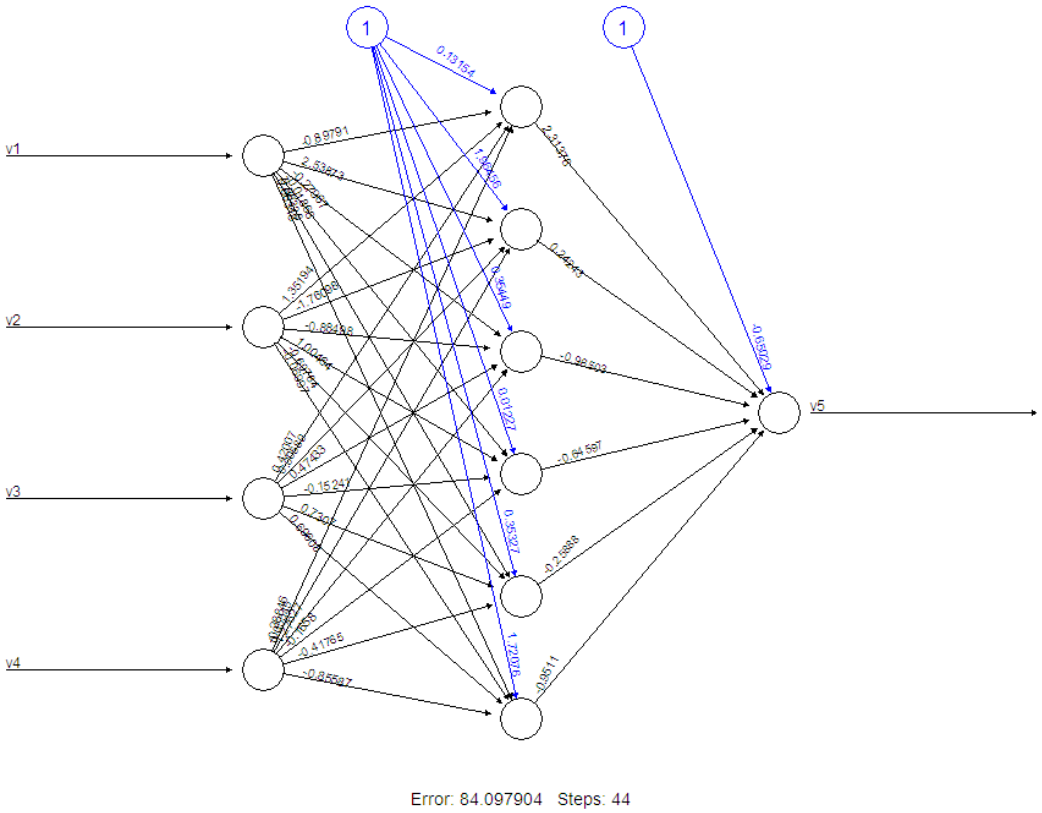

Figure 2. Randon forest diagram
图 2. 神经网络图

Table 1. Mean square error of ten fold cross validation for six kinds of machine learning methods
表 1. 六种机器学习方法十折交叉验证的均方误差

决策树	Boosting	Bagging	随机森林	支持向量机	神经网络
0.026622	0.002187	0.023172	0.000563	0.001070	0.300357

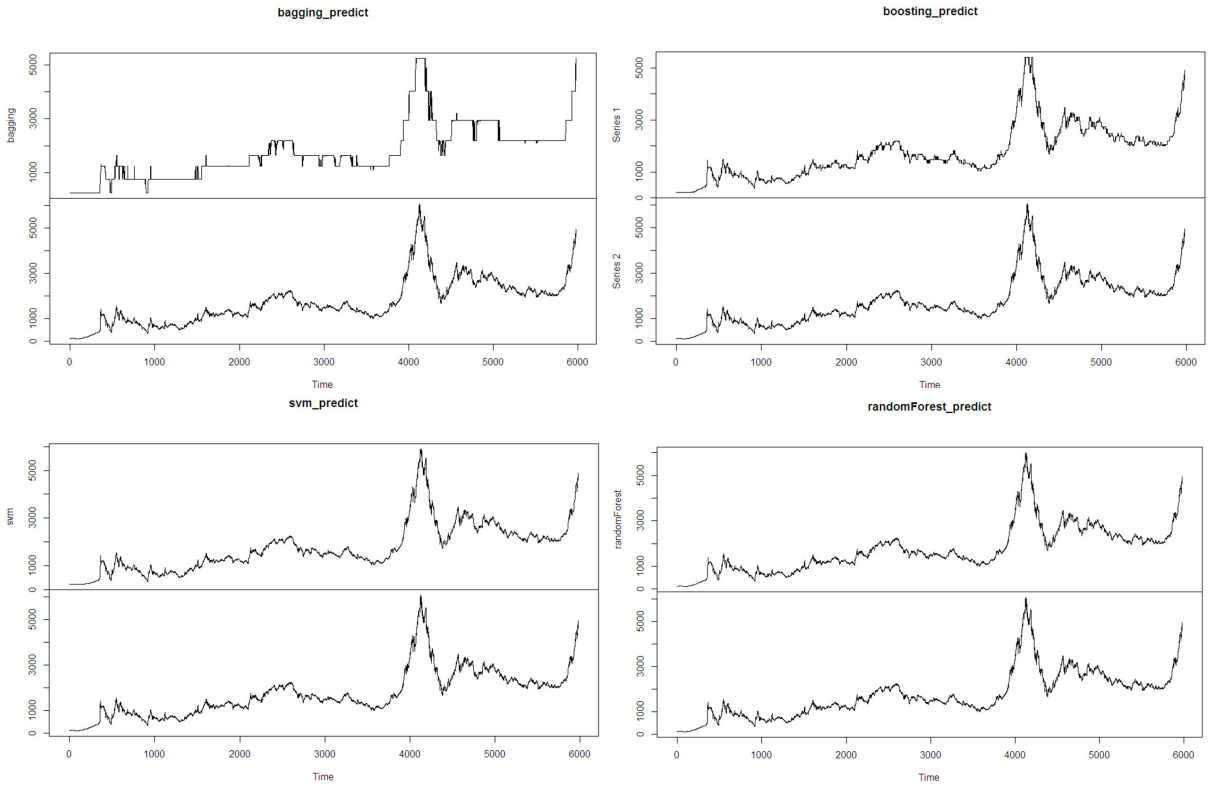

Figure 3. Four machine learning methods to predict the values and the real value of the visual comparison
图 3. 四种机器学习方法预测值与真实值的可视化对比

因此，在上证综合指数的预测分析当中以上 6 种机器学习方法不论是从十折交叉验证的均方误差还是从预测数据可视化对比图来比较，“支持向量机”以及“随机森林”两种机器学习方法都表现的较好。

4. 结论与建议

本文从现实实际问题出发，结合当下广大投资者关注的股市热点。以上证综合指数的日 K 线数据作为训练集。在技术分析方法以及传统统计学预测方法预测精度模糊的基础上，本文综合运用了六种机器学习方法(包括：决策树、boosting、bagging、随机森林、支持向量机、神经网络等)先后对训练集进行训练，得到相应模型。并建立十折交叉验证集，计算出每种机器学习方法十折交叉验证的预测均方误差。选出预测效果较好的四种机器学习方法(boosting、bagging、随机森林、支持向量机)对数据进行预测。最后，我们对四种机器学习方法预测后的数据以及真实数据分别进行数据可视化对比，发现“支持向量机”以及“随机森林”两种机器学习方法拟合效果较好，且精度高。因此，可以运用“支持向量机”以及“随机森林”两种机器学习方法建立的模型，对上证综合指数每日的开盘指数进行初步的预测与研判。

参考文献 (References)

[1] 黄伯中. 技术分析原理[M]. 香港: 明报出版社, 1995: 12-30.

[2] 鲍志强. 证券投资技巧与理论[M]. 南京: 河海大学出版社, 1991: 54-81.

[3] 马超群, 高仁祥. 现代预测理论与方法[M]. 长沙: 湖南大学出版社, 1998.

[4] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. 2nd Edition, Chapman and Hall, London. http://dx.doi.org/10.1007/978-1-4899-3242-6

[5] Granger, C.W.J. (1980) Long Memory Relationships and the Aggregation of Dynamics Models. *Journal of Econome-*

trics, **14**, 227-238. http://dx.doi.org/10.1016/0304-4076(80)90092-5

[6] Bollerslev, T. (1986) Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**, 307-327. http://dx.doi.org/10.1016/0304-4076(86)90063-1

[7] 赵传刚. 我国 A 股市场量价关系的实证分析[D]. 南昌: 江西财经大学, 2007: 20-22.

[8] 曹赛玉. 几种决策概率模型在现实生活中的应用[J]. 理论与实践理论月刊, 2006(5): 91-93.