基于神经网络分位数回归及核密度估计的概率密度预测方法

闻才喜[1]，何耀耀[1]，陈录巧[2]
（1. 合肥工业大学管理学院；
2. 徽商银行六安分行）

摘要：本文引入神经网络分位数回归和核密度估计方法，把神经网络强大的非线性自适应能力及分位数回归能更加细致刻画解释变量的优点结合起来，预测未来股票价格连续的分位数；然后运用核密度估计方法，实现未来股票价格连续概率密度曲线图。同时，利用获得的股票价格的概率密度曲线图得到未来股票价格最高概率点下对应的预测值，这样可以获得更高的股票价格预测精度。我们不仅可以获得预测当天股票价格概率密度值，同时也可以获得未来股票价格变化区间的概率密度值，这样可以为投资者在进行投资决策时，提供更多的定量依据和信息，这为投资者创造更大价值提供了帮助。

关键词：股价预测；分位数回归；神经网络；核密度估计；概率密度
中图分类号：C931

Probability Density Forecasting Method Based on Neural Network Quantile Regression and Kernel density estimation

WEN Caixi[1], HE Yaoyao[1], CHEN Luqiao[2]
(1. School of Management，HeFei University of Technology;
2. Hui Merchants Bank Lu'an Branch,Lu’an China)

Abstract: This paper introduced the neural network quantile regression and kernel density estimation method toforecast stock price in the future, which combined the powerful nonlinear adaptive ability of neural network and the advantage of quantile regression descripting explanatory variables ; and then，using the kernel density estimation method obtains probability density curve of the future stock price. Meanwhile, the application of probability forecasting density curve gets the corresponding stock price at the highest probability point value, which can obtain higher prediction accuracy. We can not only obtain the day’s stock price prediction probability density value, can also obtain the interval of the future value of the stock price changes. The method can provide more quantitative basis and value of information to investors when they make investment decisions with the help for investors to create greater value.

Key words: Stock price prediction; quantile regression; neural network; kernel density estimation; probability density

0 引言

随着我国改革开放的深入，我国资本市场也经历着深刻的变动，资本市场不断成熟起来。股票市场是资本市场最主要组成部分，随着资本市场不断成熟和完善，股票市场也在不断发展和完善。股票投资越来越成为人们日常投资生活的主要组成部分，因此关于股票价格的预测也成为更多投资者关注的重点同时也变得越来越有重要的实际意义。但是，由于股票的价格受多种因素的影响，其变化规律具有极复杂的非线性和随机性[1]，因此要建立一个比较有

基金项目：高等学校博士学科点专项科研基金资助课题（20130111120015）
作者简介：闻才喜（1986-），男，硕士研究生，神经网络分位数回归方法研究
通信联系人：何耀耀(1982-)，男，副教授，硕士生导师，主要研究方向：预测理论与方法. E-mail: hy-342501y@163.com

效的预测方法，显得比较困难。

关于股票价格预测方法的研究，国内外学者纷纷提出了众多的预测方法。例如，自回归模型（AR）[2]、差分自回归移动平均模型[3]、支持向量机模型[4]、马尔科夫链[5]、灰色预测[6]、人工神经网络类方法（ANN）[7]、模糊神经网络（FNN）[8]、CARCH 方法[9]等方法。但上述方法基本上都是单一的方法，使用单一方法需要数据满足严格的假设和条件，但是由于股票价格变化的非线性及具有高噪声性，使得单一方法预测效果并不能达到较好的预测效果。所以学者们开始尝试使用更加复杂的模型和方法的结合，实现两种方法的结合来预测多变的股票价格，如朱林等[10]提出了粗集和神经网络相结合的方法，使用该种混合杂交模型来预测股票价格，通过上证综指的实证研究得到该种方法能够取得优于传统 BP 神经网络和 GA 神经网络；王晴[11]提出了一种支持向量机(SVM)和自回归（CAR）相结合预测股票价格方法；于志军[12]提出了误差校正的 GARCH 模型，使用误差校正的思想，把误差校正和 GARCH 模型相结合，取得较好的预测效果。不管是单一预测模型还是两种方法的结合，最后得到的仅仅是关于股票价格的点预测值，都没有给出未来某天股价连续的概率密度，并不能获得更多的其他有用的信息，不能满足投资者更多的投资需要和进行更优的投资决策。

因此本文在上述的基础上，本文尝试将神经网络与分位数回归方法进行结合，结合神经网络极其强大的非线性自适应能力及分位数回归可以更加细致刻画解释变量的优点，我们可以获得未来某天连续的 200 个分位数，然后再利用这些连续的条件分位数，使用核密度估计方法实现未来某天连续的概率密度函数和概率密度曲线图，这样我们不仅可以得到具体的点预测值和股票价格的变化区间，同时也可以得到股票价格预测变化区间各值的概率。我们用最大概率点对应的预测值，得到当天较准确的点预测值。通过上证综指的实证研究，我们不仅取得较好的效果，同时还可以获得未来某天连续的概率密度曲线图，使得股票价格当天价格走势及其对应的概率水平都可以得到很好的反映，能够为投资者进行决策提供更多有用的信息。

1 神经网络分位数回归理论

Kenker[13]为了解决传统均值回归要求样本数据满足正态分布但很多数据不能满足这一特点的缺点，提出分位数回归。分位数回归可以通过计算分布的几个百分位处的回归线而得出更完整的关于解释变量的信息。分位数回归可以克服数据非对称分布和散布较大的情况，可以更加细致的刻画被解释变量与解释变量的之间的关系，能够完整的考察被解释变量完整的条件分布。线性分位数回归的模型为：

$$Q_Y(\theta|\boldsymbol{X})=\beta_0(\theta)+\beta_1(\theta)X_1+\beta_2(\theta)X_2+....+\beta_n(\theta)X_n\equiv\boldsymbol{X}'\boldsymbol{\beta}(\theta) \quad (1)$$

式中，$Q_Y(\theta|\boldsymbol{X})$ 为预测对象的条件分位数，$\boldsymbol{X}$ 为解释变量，θ 为分位点，$\boldsymbol{\beta}(\theta)$ 为分位数回归一系列回归系数向量，其在均值回归中是唯一确定的，但在分位数回归中其是随着分位点的不同而在不断发生变化。而对式子中 $\boldsymbol{\beta}(\theta)$ 的估计可以转化到求解下式：

$$\min_{\beta}\sum_{i=1}^{N}\rho_\theta(Y_i-\boldsymbol{X}_i'\boldsymbol{\beta})=\min_{\beta}\theta|Y_i-\boldsymbol{X}_i'\boldsymbol{\beta}|+\min_{\beta}(1-\theta)|Y_i-\boldsymbol{X}_i'\boldsymbol{\beta}| \quad (2)$$

其中：N 为样本的个数，$\rho_\theta(u)=u[\theta-m(u<0)]$，为示性函数。

$$m(u)=\begin{cases}0, & u\geq 0\\ 1, & u<0\end{cases} \quad (3)$$

分位数回归通过分位点的不同，最后通过内点算法，实现绝对值残差最小，取得参数最

优估计。随着分位点的不同变化，我们可以获得预测对象完整的条件分位数。

人工神经网络是复杂的网络计算系统并由高度相互关联的大量的简单的神经元组成的。人工神经网络是一门活跃的边缘交叉学科，其基本性质有高度的非线性、和自学习性、鲁棒性、及推广性等优点，同时人工神经网络也具有计算的非确定性等特点。人工神经网络常见的形式有：RBF 神经网络、BP 神经网络、Hopfield 神经网络、小波神经网络等形式。本文选用的神经网络隐含层核函数是双曲正切函数，采用这种函数可以对高度复杂的数据进行很好的非线性拟合，建立稳定以及更好预测能力的非线性函数，为提高股票价格精度，提供更好的方式。其中双曲正切函数的形式如下所示：

$$tanh(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{4}$$

其中 tanh(x)为神经网络隐含层期望输出值，x 为输入变量组成的矩阵。

本文的神经网络分位数回归是基于 Taylor 提出的单隐层神经网络模型[14]，运用神经网络分位数回归预测未来股票价格的分位数，然后再利用 Sigmoid 函数作为神经网络隐含层函数，把得到的未来股票价格的预测分位数作为核密度估计的输入变量，实现对股票价格概率密度预测。神经网络分位数回归模型的表达式为：

$$\begin{aligned} Q_{Y}\left(\theta|X\right) &= f\left[\mathrm{x},\ \mathrm{u}(\theta),\ \mathrm{v}(\theta)\right] \\ &= \sum_{j=1}^{J}\left\{\frac{2v_{j}(\theta)}{1+e^{-2\sum_{i=1}^{n}u_{ij}(\theta)X_{i}}}-v_{j}(\theta)\right\} \end{aligned} \tag{5}$$

其中 θ 为分位点，$\boldsymbol{u}(\theta)=\{u_{ij}(\theta)\}_{i=1,2,\ldots,n;\ j=1,2,\ldots,J}$ 为输入层与隐含层之间的待估计权重矩阵；$\boldsymbol{V}(\theta)=\{v_{j}(\theta)\}_{j=1,2,\ldots,J}$ 为隐含层与输出层之间的连接权重向量。为了达到模型（5）的最优参数估计，我们可以可以通过求解优化目标函数：

$$\tilde{E}_{\theta}=\frac{1}{N}\sum_{i=1}^{N}\rho_{\theta}\left\{Y_{i}-f\left[X,\ U(\theta),\ V(\theta)\right]\right\} \tag{6}$$

来实现这一寻优过程。但是，为了使训练的神经网络不进入过度拟合的状态，我们在目标函数函数加入了相应的惩罚参数项，得到新的目标函数:

$$E_{\theta}(U(\theta),V(\theta))=\tilde{E}_{\theta}+\lambda_{1}\sum_{i,j}u_{ij}(\theta)+\lambda_{2}\sum_{i,j}v_{j}(\theta) \tag{7}$$

其中，λ_1、λ_2 为模型惩罚参数，通过确定最优惩罚参数，可以有效防止模型陷入过度对经验数据拟合，减少预测误差，提高预测精度。对式（7）进行优化可以得出 $U(\theta)$、$V(\theta)$ 的最优估计值 $\bar{U}(\theta),\bar{V}(\theta)$。然后将 $\bar{U}(\theta),\bar{V}(\theta)$ 带入式（5)中，得到响应变量条件分位数估计函数[15]。

2 核密度估计概率密度函数

核密度估计由于不需要对随机变量的先验分布进行任何假设，只需要确定好输入变量、核函数以及最优窗宽，就可以通过核密度估计方法，根据预测得到的未来某天连续的股票价格，得到其连续概率密度曲线图。核密度估计思想最本质的思想是要通过核密度估计量，从而估计得到合理的密度函数，核密度估计量形如：

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}k(\frac{X_i-x_0}{h})=\frac{1}{n}\sum_{i=1}^{n}k_h(X_i-x_0) \quad (8)$$

其中$k(\ )$为核函数，$k_h=k（x/h）/h$。本文在采用核密度估计时，选取伊番科尼科夫核（Epanechnikov）核函数形式并用交叉验证法来确定最优窗宽。其中 Epanechnikov 核函数形式为：

$$k(x)=\frac{3}{4}(1-x^2)I(|x|\le 1) \quad (9)$$

在式子中，$I()$为示性函数，当括号中的条件为真时，$I()$取值为 1，当条件为假时，$I()$取值 0。而交叉验证法函数表达式为：

$$CV_f(h)=\frac{1}{nh^2\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{k}(X_i-X_j)}-\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}k_h(X_i-X_j) \quad (10)$$

式中，$\bar{k}(v)=\int k(u)\,\mathrm{k(v-u)}\,\mathrm{du}$是从$k(\ )$导出的卷积核函数，只要$k(\ )$的具体形式给出，我们就能得到$\bar{k}(v)$一个具体的表达式。这样我们可以通过上述方法从而实现未来某天股票价格连续的概率密度函数及概率密度曲线图。

3 上证综指预测实证研究

3.1 数据来源及样本描述

由于我国上海股票市场起步早,数据比较能够全面代表我国股票市场的发展状况,且大盘指数综合了各个方面的影响 ,具有较强的代表性 ,因此本文以上证综指每日收盘指数为研究对象。选择上证综指为预测对象。我们选择从 2008 年 1 月 2 日到 2014 年 6 月 30 日期间指数的收盘价作为样本，我们以样本当日收盘价为模型被解释变量，以当日开盘价和前五日收盘及开盘价为模型解释变量，实行滚动预测的方式，共有 1571 组样本。我们将这 1571 个样本输入到神经网络分位数回归模型，确定好模型结构，对神经网络进行训练，是网络能够稳定下来，满足要求，然后预测获得 2014 年 7 月 1 日到 2014 年 7 月 14 日期间的上证综指的收盘价———每天 200 个连续条件分位数，然后再代入核密度估计模型中，确定未来某天股票价格变化的概率密度曲线图。

表 **1** 上证综指样本描述统计

Tab.1 Shanghai Composite sample descriptive statistics

N	1571
Maximum	5497.9
Minimum	1706.7
Skewness	1.66188
Kurtosis	7.51889

这 1571 个样本具有明显的“尖峰后尾”的特征，其描述统计如表 1 所示。从表 1 中分析，

我们可以得出该样本的峰度值为 7.51889，很明显大于正态分布的峰值度值 3，同时偏度值为 1.66188，这些都充分说明了样本“尖峰厚尾”的特征，而不是服从常见的正态分布。如果采用传统的方法，很难建立响应变量与输入变量之间的关系。而本文提出了神经网络分位数回归与核密度估计方法不仅克服了样本的非正态分布的缺陷，而且还可以获得更多关于未来股价的有用信息。

3.2 模型参数选择

本文构建的模型是基于单隐层的神经网络分位数回归。其中，神经网络的迭代次数为 5000 次，输入层为 11 层，隐含层为 1，输出层为 1，神经网络的结构为 11-1-1；同时为了防止神经网络分位数回归网络陷入过度拟合的情况，模型的惩罚参数 λ_1、λ_2 都设置为 0.1。模型中的分位数点我们选择从 0.0001 开始，一直到 0.9999，间隔 0.005，共选取 200 个分位点，这样神经网络分位数回归模型参数确定下来。而核密度估计形式与最优窗宽选择采用 Epanechnikov 核函数与交叉验证法进行组合，获得其概率密度曲线图及最高概率点对应的点预测值。

3.3 实证结果及分析

由上述内容，我们将滚动方法获得的样本，带入神经网络分位数回归的网络中，训练神经网络结构，并把每天连续的上证综指的 200 个条件分位数带入核密度估计方法中，我们得到未来某天上证综指完整的概率密度曲线图，如图 1 所示。

从下图 1 可以看出，通过本文提出的模型方法，我们可以首先可以获得未来某天股价的连续概率密度函数，通过完整的概率密度函数不仅可以获得每一个股价出现的概率，同时也可以获得股价区间的概率情况，这为投资者进行投资时，提供更多投资决策依据；其次从图上我们可以分析得到，收盘价基本上都处于概率密度曲线图上最高概率点附近处，这样对收盘价真实值获得较好的预测，这也为投资者提供了更有力的决策依据；最后，通过上图也可以看出股票的收盘价、最高价及最低价都出现在估计的概率密度曲线图上，而且都基本出现在最高概率点的附近，这样股票价格每天的价格波动的区间概率密度也可以获得，为投资者预测未来股票价格变动区间提供了比较好的依据。所以采用本文提出的方法，不仅可以实现较准确的预测值，同时也可以为投资者提供更多关于股价变动的有用信息。

通过表 2，运用本文提出的方法可以预测获得未来某天股价的最大相对误差为 1.20%，最小相对误差 0.02%，平均绝对误差为 0.42%。这一结果对上证综合指数的预测精度是比较好的，可以为投资者进行投资决策时，提供一个很好的参考依据。

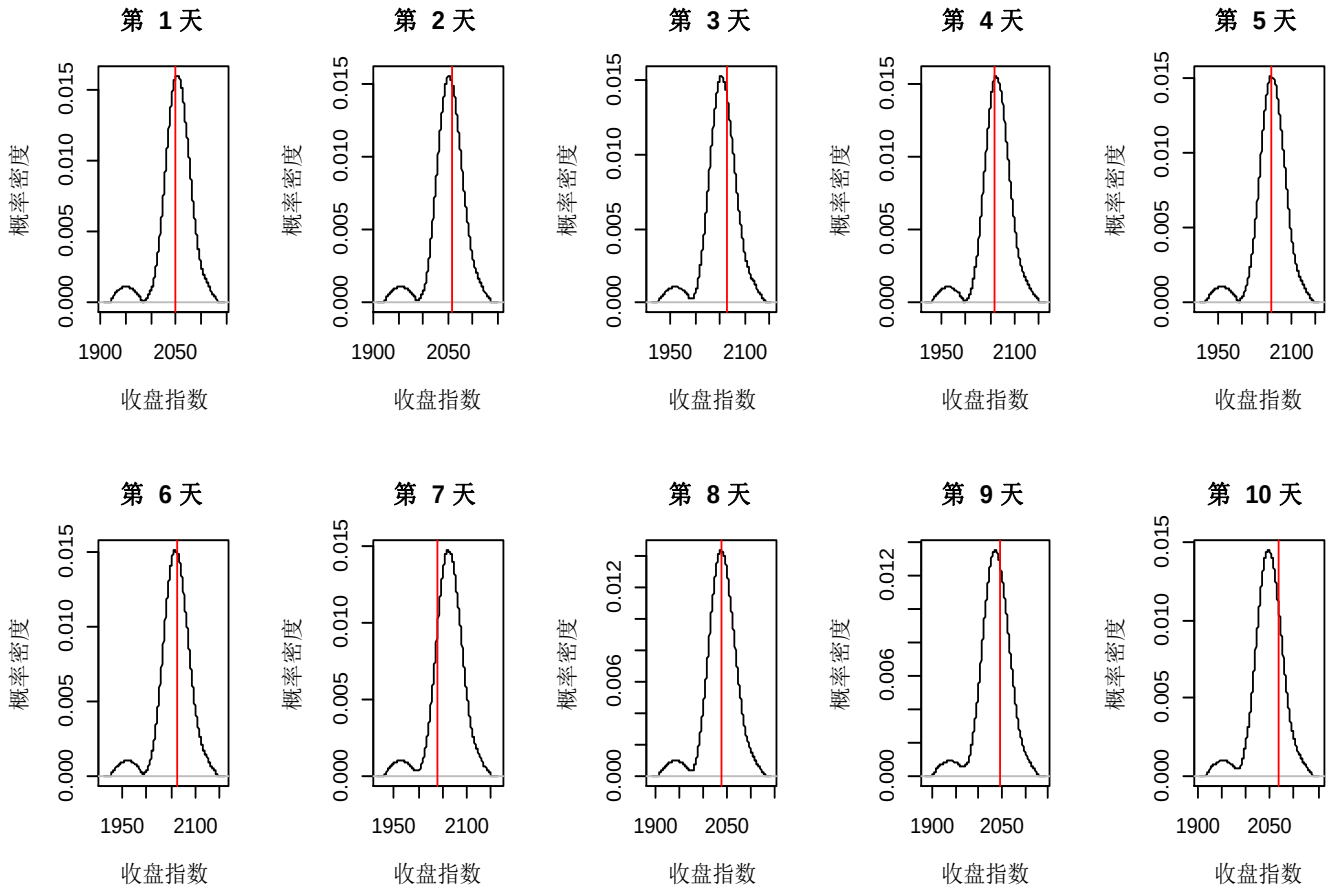

图 1 预测的股票价格概率密度

Fig. 1 Predict stock price probability density

表 **2** 本文模型预测结果及相对误差

Tab. 2 The model prediction results and the relative error of the paper

日期	收盘价格	最高价格	最低价格	预测价格	相对误差/%
2014-7-1	2050.38	2066.64	2041.94	2053.49	0.15
2014-7-2	2059.42	2060.60	2044.04	2048.44	-0.53
2014-7-3	2063.23	2066.64	2048.08	2053.08	-0.49
2014-7-4	2059.38	2065.08	2054.22	2063.42	0.20
2014-7-7	2059.93	2064.04	2050.89	2061.98	0.10
2014-7-8	2064.02	2064.43	2047.20	2059.37	-0.22
2014-7-9	2038.61	2062.47	2038.61	2063.14	1.20
2014-7-10	2038.34	2045.53	2034.96	2038.76	0.02
2014-7-11	2046.96	2051.24	2033.00	2038.09	-0.43
2014-7-14	2066.65	2067.34	2044.90	2048.94	-0.86

4 结论

（1）本文提出的神经网络分位数回归与核密度估计的股票价格概率密度预测方法，不需要对样本数据进行任何假设，能够克服样本数据非正态性分布，很好的刻画了股票价格“尖峰厚尾”的特征。

（2）通过获得预测当天连续的概率密度曲线图，我们不仅可以获得预测当天的收盘价的概率密度，同时也可以获得预测当天股票价格变动区间的概率密度，可以使得投资者能够

较准确的获得未来股票价格变动区间及其概率，投资者可以根据预测获得的信息，进行股票投资决策。也就是说我们即可以实现较准确的预测值，同时又可以为投资者获得更多更有用的信息，为投资者进行投资决策提供更多有力的定量依据。

（3）本文可以得出较准确的预测率，能够减少投资者投资的不确定性，可以让投资者有更多的获利机会。

（4）通过概率密度曲线图，股票价格的最低价和最高价也出现在概率密度曲线图的中部附近，这样投资者可以在获得预测概率密度曲线图时，可以获得股票价格变动区间及其概率密度，这为投资者进行投资决策时，提供更多定量指标。

（5）通过预测概率密度曲线图，我们可以获得未来股价的走势图及其对应的概率密度，这样投资者，可以很好的了解具有多变性股价走势。

[参考文献]

[1] 吴文峰，吴冲锋.股票价格波动模型探讨[J]．系统工程理论与实践，2000，16(4):63-69.
[2] Champernowne D G. Sampling theory applied to autoregressive schemes[J]．Journal of the Royal statist-ical society:series B,1948,10:204-231.
[3] Box G E P, Jenkins G M. Time series analysis:Forecasting and control[M].3th ed. New York :Wiley,1994.
[4] 丁爱霞. 基于支持向量机自回归分析的股市动态预测模型及其应用研究[J].知识经济，2009,15,31-32.
[5] 李东，苏小红，马双玉.基于新维灰色马尔科夫模型的股价预测算法[J].哈尔滨工业大学学报，2003,35(2):244-248.
[6] 陈海明，李东. 灰色预测模型在股票价格中的应用[J].科研管 理，2003，24(2)：28-31.
[7] 常松，何建敏. 基于小波包和神经网络的股票价格预测模型[J] .中国管理科学，2001,9(5)：8-15.
[8] 杨一文，刘贵忠，基于模糊神经网络和 R/S 分析的股票市场多步预测[J] .系统工程理论与实践，2003,3:70-76.
[9] 王军波，邓述慧. 利率、成交量对股价波动的影响-- GARCH 修正模型的应用[J] . 系统工程理论与实践，1999，9，49-57.
[10] 朱林，何建敏，常松. 粗集与神经网络相结合的股票价格预测模型[J] . 中国管理科学，2002,10(4)，27-33.
[11] 王晴.组合模型在股票价格预测中应用研究[J] .计算机仿真，2010,27(12):361-364.
[12] 于志军，杨善林. 基于误差校正的 GARCH 股票价格预测模型[J] . 中国管理科学，2013,21,341-345.
[13] Koenker R W，Bassett Jr G．Regression quantiles [J]．Econometrica，1978，46(1)：33-50.
[14] Taylor J W．A quantile regression neural network approach to estimating the conditional density of multiperiod returns[J]．Journal of Forecasting，2000，19(4)：299-311.
[15] 何耀耀，许启发，杨善林，等．基于 RBF 神经网络分位数回归的电力负荷概率密度预测方法[J]．中国电机工程学报，2013，33(1)：93-98