Weekly Report of Research Work

WR-ABS-TEMP-2015A-No.014

汤吉(Ji TANG)
Number: WR-ABS-TEMP-2015A, E-mail: tangji08@hotmail.com
Date: 22/2/2016 - 28/2/2016

February 28, 2016

Contents

1 Work

1. Finishing translating the paper
2. Preparing the books for all the students

2 Question

1. Should I translate the function in the paper to document directly or by using the type of latex?
2. How should I deal with the figure in the paper?

3 The translated paper

Hurst指数和金融市场预测

Bo Qian, Khaled Rasheed
计算机科学系，佐治亚大学
Athens, GA 30601
USA
mailto:khaled]@cs.uga.edu

摘要 Hurst指数(H)是一个统计学测量用来分类时间序列。当H=0.5时，表示一个完全随机的序列。而当H>0.5时，表示了一个具有保持趋势倾向能力的序列。H的值越大，这个序列的倾向也越强。我们接下来将要研究如何利用Hurst指数来将不同时期的金融序列数据进行分类。BP神经网络的实验表明，具有高Hurst指数的序列比那些Hurst指数接近于0.50的序列能够被更加精确的预测。因此，Hurst指数提供了一种预测方法。

关键词 Hurst指数，时间序列分析，神经网络，蒙特卡诺模拟，预测

1．介绍 Hurst指数是H. E. Hurst提出用来作分形分析的，现在已经被用在许多研究领域。最近，由于Peter的相关工作，它在金融领域也变的十分热门。Hurst指数为长期记忆和时间序列的分形提供了一种方法。由于它是高鲁棒性的基本系统的几个假设，现在已经被广泛用于时间序列分析。Hurst指数的值在0和1之间。基于Hurst指数H，一个时间序列能够被分为三种类型：(1)H=0.5表明了序列可以用随机游走来描述。(2)0<H<0.5表明了序列具有反持续性。(3)0.5<H<1表明序列具有持续性。一个反持续性序列具有均值回复的特性，即意味着一个上升的值更有可能紧接着一个下降的值，反之亦然。H的值越接近于0.0，序列均值回复的能力也越强。而一个持续性序列具有保持倾向的能力，即下一时刻的值相对于现在值的变化，更有可能与这一时刻相对于上一时刻值的变化一致。H的值越接近于1.0，序列保持倾向的能力也越强。大多数的经济和金融时间序列具有持续性，即H>0.5。

在时间序列的预测当中，我们首先需要解决的问题是我们想要研究的这个时间序列是否可以被预测。如果这个时间序列是随机的，一切的方法都是无效的。我们想要确定这些序列具有一定的可预测等级。我们知道一个具有很高H值的时间序列是具有很强的倾向性的，所以我们自然地认为这样的时间序列要比那些H值接近于0.5的时间序列更可能被预测。接下来，我们将要使用神经网络来测试这个假设。

神经网络是无参数的通用函数逼近，可以无假设地从数据中进行学习。在过去的十年里，神经网络预测模型已经被广泛应用于金融时间序列分析。神经网络可以被用来代替通用函数逼近，进行预测。在同样的条件下，一个时间序列如果比另外一个时间序列具有更小的预测误差，我们便说它更容易被预测。从1930年1月2日到2004年5月14日，我们研究每日的道琼斯指数，计算每1024交易日的Hurst指数。从当中选出30个具有最大的Hurst指数与30个Hurst指数接近于随机序列的周期，然后我们用这些数据来训练我们的神经网络。我们对比这两组数据的预测误差，发现他们的预测误差完全不同。这个研究是通过Matlab来实现的，这篇文章所有的Matlab程序生成的结果都可以从www.arches.uga.edu/ qianbo/research下载。

在这篇论文接下来的部分：第二部分将会详细描述Hurst指数，第三部分我们将利用蒙特卡洛模拟过程来构造一个类似让我们感兴趣的金融序列，第四部分描述了一个我们模拟生成的混乱序列用来验证根据样本顺序构造的模型，第五部分描述了神经网络和他们用来验证高Hurst指数的序列能够比低Hurst值的序列更加准确地被学习和预测。最后，这篇论文将在第六部分作出结论。

2．Hurst指数与R/S分析

Hurst指数能够通过重标极差分析(R/S)分析。对于一个时间序列，X = X1, X2, ⋯，Xn，R/S分析方法如下：

1) 计算平均值m

$$m = (X1 + X2 + \cdots + Xn)/n$$

2)计算均值调整序列Y

$$Yt = Xt - m \quad t = 1, 2, \cdots, n$$

3)计算累计偏离序列Z

$$Zt = Y1 + Y2 + \cdots + Yt \quad t = 1, 2, \cdots, n$$

4)计算序列范围R

$$Rt = max(Z1, Z2, \cdots, Zt) - min(Z1, Z2, \cdots, Zt)$$

$$T = 1, 2, \cdots, n$$

5)计算标准差序列S

$$St = sqrt(((X1 - u)^2 + \cdots + (Xt - u)^2)/t)$$

$$T = 1, 2, \cdots, n$$

在这里，u是从X1到Xt的平均值

6)计算重标极差序列(R/S)

$$(R/S)t = Rt/St \quad t = 1, 2, \cdots, n$$

我们把(R/S)t记为区间平均值[X1,Xt],[Xt+1,X2t]知道[X(m-1)t+1,Xmt]

其中m=floor(n/t)。事实上，为了计算所有的数据，t的值是可以整除n的。

Hurst发现(R/S)随着时间的增加具有指数增长的规律，研究表明

$$(R/S)t = c * t^H$$

在这里c是一个常数，H被称为Hurst指数。为了分析Hurst指数，我们画出(R/S)随着t变化的log图。这条回归直线的斜率便可以用来估计Hurst指数。当t<10时，(R/S)t是不准确的。所以我们将利用至少10个值来计算重标极差。图2.1便是一个R/S分析的例子。

在我们的实验中，我们计算了一个1024交易日周期的Hurst指数。我们用$t = 2^4, 2^5, \cdots, 2^10$来进行回归计算。在金融领域，采用对数变化率的计算每日收益的方法十分普遍。对于累积变化对应的累积收益率，这在R/S分析中是非常有意义的。图2.2展示的是道琼斯指数从1930年1月2日至2004年5月14日的日收益率。图2.3展示的是在这个期间内对应的Hurst指数。在这个期间内，Huest指数从0.4200至0.6804波动。我们同样也想知道什么样的Hurst指数能够满足我们的条件。

3.蒙特卡洛模拟

对于一个随机序列，Feller给出了一个预计的(R/S)t公式:

$$E((R/S)t) = (n * \pi/2)^{0.5} \quad (3.1)$$

然而，这是一个近似的关系并且只在t很大时才有效。Anis和Lloyd为了在t很小时克服这个误差提供了下面这个公式:

$$E((R/S)t) = (\Gamma(0.5 * (t-1))/(\sqrt{\pi} * \Gamma(0.5 * t))) * \Sigma_{r=1}^{t-1}\sqrt{(t-r)/r} \quad (3.2)$$

当t>300时，大多数的计算机都很那来计算gamma函数了。利用Sterling的方程，此公式还能够被近似为:

$$E((R/S)t) = (t * \pi/2)^{-0.50} * \Sigma_{r=1}^{t-1}\sqrt{(t-r)/r} \quad (3.3)$$

Peter又给出了公式3.2的一个修正(公式3.4)

$$E((R/S)t) = ((t-0.5)/t) * (t * \pi/2)^{-0.50} * \Sigma_{r=1}^{t-1}\sqrt{(t-r)/r} \quad (3.4)$$

我们计算了对应t=$2^4, 2^5, \cdots, 2^10$期望的(R/S)值并且在$\alpha = 0.05$的置信水平下作了平方回归。结果如表格3.1所示。

从表格3.1可以看出，对于Feller, Anis和Peter之间的公式有不少的差别。更进一步，他们的公式都是基于大量的数据点的计算。我们现在的数据量固定在1024个点。所以在我们这种情况下，随机序列的Hurst指数是怎样的?

幸运地，我们可以利用蒙特卡洛估计法来得到结果。我们生成了10000个高斯随机序列。每个序列都有1024个值。我们计算了每个序列的Hurst指数和平均值。我们希望这个平均值尽量地接近实际的值。然后重复这个过程10次。下面的表格3.2给出了这个模拟结果。

从表格3.2，我们可以看出在我们实验的情况下，可以计算出蒙特卡洛模拟的Hurst指数为0.5454，标准差为0.0485.这个结果是非常接近于Anis的公式的。通过以上的模拟，在95%置信度的情况下，Husrt指数处在0.5454+-1.96*0.0485的区间，即0.4503到0.6405之间。我们选择一个Hurst指数大于0.65的区间，希望找到这些区间一些不同于随机序列的规律。然而，从这些大样本(总共17651个周期)中选出的样本中，我们想知道是否存在或碰巧存在这些时期的真实模型。为了达到这个目的，我们进行了一个混乱测试。

4．混乱测试 为了测试是否存在一个真实的模型能满足Hurst指数大于0.65时期的样本，我们从中随机选择了10个样本。对于每一个样本，我们打乱这个序列，然后计算这个混乱序列的Hurst值。被打乱的序列与非随机序列的原始样本具有相同的分布。如果对于这些序列存在某些模型，在打乱顺序之后，这些模型将会被破坏并且计算的Hurst值也会接近于随机序列的Hurst值。在我们的实验当中，我们将每个样本打乱了500次，然后计算了平均的Hurst值。结果如表格4.1所示。

从表格4.1我们可以看出，在样本被打乱顺序之后，Hurst指数都非常接近于0.5454，这与我们的模拟随机序列一致。通过这个结果，我们可以得出在这些时期内必然存在某些模型使得时间序列不同于随机序列，并且会被颠倒顺序被破坏。我们希望这个模型能够被用来进行预测。神经网络作为一个通用方程的近似器，提供了一个强大的工具用来学习这个潜在的模型。当潜在的规则未知时，他们变得尤为有用。我们希望用神经网络来发现这个模型并从中受益。我们用神经网络来控制错误率在可控制的范围内。接下来，我们对比Hurst指数大于0.65的时期与处于0.54和0.55之间的时期的预测误差。

5.神经网络 在1943年，McClloch和Pitts提出了一个模拟神经估算模型。这项工作被普遍认为是人工神经网络研究的开端。Rosenblatt普及了感知器的概念并创造了很多感知器的学习规则。然而，在1969年，Minsky和Papert发现感知器无法解决一些非线性可分的问题。人们认识到多层感知器(MLP)能够模拟非线性可分方程，但是没人知道如何训练他们。神经网络的研究在1986年几乎停止了。在1986年，Rumelhart使用了反向传播算法来训练MLP，终于解决了这个长时间困扰人们的问题。从那以后，神经网络在很多领域重新得到了重视。神经网络开始在金融领域变得热门，并且在金融方面的神经网络研究投入资金排名第二。

一个神经网络是一系列简单的相互联系的处理节点。每个节点计算加权输入，然后输出其传递函数对其他节点的值。前馈反向传播网络是最广泛使用的网络范式。利用反向传播算法训练算法，神经网络将调整权重，使它减少所观察到的输出之间的平方差（误差）和他们的目标值。反向传播算法使用梯度下降法寻找局部极小值误差曲面。它对每一个权重计算平方差的偏导数。偏导数（梯度）的相反数给出了使误差减小的方向。这个方向被称为最速下降方向。标准反向传播算法调整权重沿最速下降方向。虽然沿最速下降方向的偏差减少的最快，它通常收敛缓慢，并且可能收敛于局部最小值并振荡。因此，许多反向传播变种算法被发明，他们通过优化方向和步长来提高性能。比如说几个有名的，我们有反向传播动量，共轭梯度，准牛顿以及LM。经过训练，我们可以使用这个网络来预测给定的不可见的输入。在神经网络预测中，第一步是数据的准备和预处理。经过训练后，我们可以用神经网络来预测给定的不可见的输入。

5.1 数据的准备和预处理 对于道琼斯日收益率数据，从1930年1月2日至2004年5月14日，我们计算每1024交易日的Hurst指数。在总共17651的时期中，有65个时期的Hurst指数大于0.65,1152个时期的Hurst指数处于0.54至0.55之间。图5.3展示了所有周期的条形图。

我们从Hurst指数大于0.65的样本中随机选择了30个周期，并从Hurst指数处于0.54至0.55之间的也选择了30个周期。两组总共60个样本作为我们的原始数据集。

给定一个时间序列x1,x2,⋯,xi,xi+1，基于x1,x2,⋯,xi我们应该如何构造一个向量X1用来预测xi+1呢？Taken的理论告诉我们，如果我们有合适的d和t，我们可以通过延时向量Xi=(xi,xi+t,xi+2t,⋯,xi+(d-1)t)来重构一个潜在的系统。这里d被称为嵌入维度，t叫做分离。使用自动同步信息和假最近邻方法，我们可以估算d和t。我们用TSTOOL包对我们的60个数据集来运行自动同步信息和假最近邻方法。建议使用自动同步信息将所有数据集的分离信息的方法。这符合我们的直觉，因为我们没有理由使用分离的值。至于嵌入维数，我们的数据集建议从3到5来取。我们将在后面进行检验。

在构造好了时间延迟向量Xi和目标值xi+1，我们将输入Xi和输出xi+1规范化到平均值为0，标准差为1的分布。我们没有必要将输出规范化至一个有限的区间里，比如说-0.85到0.85，以避免出现饱和的现象。因为我们在输出层使用了一个线性的转换方程。

我们使用了一个通常用来解决过拟合问题的神经网络。我们将数据集分成了三部分以用来进行训练，生效和测试。训练的数据集通过误差的反向传播来调整权重。生效集是通过判定

此集合中的均方差开始增加时，停止训练。神经网络的预测性能是通过测试集来评定的。我们一般使用前60%的数据来进行训练，接下来的20

5.2神经网络的构造 尽管神经网络是通用函数的近似器，我们仍然需要注意一下他们的构型。我们到底需要多少层？我们应该使用哪一种算法？实际上，到目前为止还没有直接的证据表明多隐含层的神经网络比单隐含层的神经网络更有优势，大多数在构建过程中都只使用了一层隐含层。因此，在我们的研究中将使用单隐含层的结构。对于学习算法，我们测试了Levenberg-Marquardt， 共轭梯度法，以及包含动量算法的反向传播算法。我们发现Levenberg-Marquardt 比其它的算法具有非常明显的优势。我们因此选择了Lvenberg-Marquardt算法，隐含层选用了Sigmoid传递函数，输出层则选用了一个线性函数。现在，我们需要确定隐藏的维数和隐含层的节点数。有一个探索式的规则用来确定隐含层的节点数，即神经网络总的自由度需要等于1.5倍的总数据量的平方根。根据这一条规则，我们得到以下的方程：(输入节点数+1)*(隐含层节点数)+(隐含层节点数+1)*(输出节点数)=1.5*sqrt(数据总数) (5.2.1) 方程得到隐含层节点数的解是10.类似的，我们发现维度为4和5的网络隐含层的节点数为8和7.对于每一个维度，根据建议的隐含层节点数，我们测试了5种神经网络构型。例如，对于3维度的8,9,10,11,12的隐含层节点数的神经网络。我们从每一组中(Hurst指数大于0.65的组和Hurst指数处于0.54和0.55之间的组)随机选择了5个时期用来训练神经网络。每一个神经网络被训练100次，然后最小的规范误差均方根NRMSE被记录下来。NRMSE被定义为：

$$NRMSE = \frac{\sqrt{\Sigma_i (O_i - T_i)^2}}{\sqrt{\Sigma_i (T_i - \overline{T})}} \quad (5.2.2)$$

在5.2.2中，O是输出的值，T是目标值。NRMSE给出了一个对于平均预测的性能比较方法。如果我们经常使用这个平均值去预测，NRMSE将会变成1.当NRMSE的值为0时表明所有的预测都是正确的。

表格5.1-5.3给出了对于不同的神经网络的训练结果。 从表格5.1到5.3我们可以看出，对于每一个维度，不同的隐含层的节点数，NRMSE的差异非常小。对于3,4,5维度的网络，拥有最小平均NRMSE值的隐含层节点数是8,8,6.然后我们利用8,8,6的隐含层节点数的神经网络来对3,4,5维度的数据进行预测。每一个神经网络将会被训练100次，最小的NRMSE将会被记录下来。最后NRMSE将会是维度3的最小值。表格5.4给出了我们最初60个样本分成两组后的NRMSE值。

我们将两组不同的学生测试作为零假设，即两个组的平均值相同。再计算之后发现，t的统计量是7.369，p的值是7.0290e-010.这表示两个平均值是完全不同的，或者说他们相同的概率为0.这个结论证明了具有高Hurst值的时间序列更容易被预测精准。

6.结论 在这篇论文当中，我们分析了从1930年1月2日至2004年5月14日道琼斯股票所有1024交易日周期的Hurst指数。我们发现具有高Hurst值的数据比那些Hurst值接近于随机序列的数据更加容易被预测精确。这表明了股票市场并不是在所有时刻都是随机的。有一些时期会有

很强的走势倾向模型，这些模型能够被神经网络学习并用于预测。

自从Hurst指数提供了一个评价可预测性的方法，在预测之前，我们能够使用这个值来指导我们进行数据的筛选。我们可以选择具有高Hurst值的模型来进行预测。更进一步，我们可以只关心具有高Hurst值的时期。这样能够大大地节约物力财力，指导我们更有效率低进行预测。

(total 5292 words)