基于时间依赖的改进样本熵分析股票时间序列

于文静,余 洁,徐凌宇
(上海大学 计算机工程与科学学院,上海 200444)

摘 要: 样本熵是一个度量时间序列复杂度的非线性方法,广泛应用于各领域。然而,研究表明熵值的大小并不总是和时间序列的复杂性相关。为了解决这个问题,提出了多尺度熵,用来度量不同尺度下的时间序列的复杂度。但是,考虑到这种方法并没有解决样本熵在度量时间序列复杂度的问题,提出了基于时间依赖的改进样本熵,并将其用在股票收盘价和成交量时间序列上,研究它们对应的复杂度关系。同时,结合多尺度的方法,衡量不同尺度下股票收盘价时间序列和成交量时间序列的复杂性。实验结果表明,从收盘价时间序列和成交量时间序列的复杂度变化上能够揭示一定的股票的发展规律。另外,收盘价序列在不同的尺度上能够保持一致性,而成交量序列在不同的尺度上熵值变化则有不同的趋势,且股票类型越接近,熵值变化曲线也越接近。

关键词: 样本熵; 时间依赖; 多尺度熵; 股票时间序列
中图分类号: TP39 **文献标识码:** A **文章编号:** 1673-629X(2019)03-0060-04
doi: 10.3969/j.issn.1673-629X.2019.03.012

Analysis of Stock Time Series Based on Time Dependent Modified Sample Entropy

YU Wen-jing, YU Jie, XU Ling-yu
(School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: Sample entropy is a nonlinear method to measure the complexity of time series and widely applied in various fields. However, studies have shown that the entropy is not always related to the complexity of time series. To solve this problem, multi-scale entropy is proposed to measure the complexity of time series over different scales. However, considering that this method does not solve the problem of sample entropy in measuring the complexity of time series, a modified sample entropy based on time dependent is proposed and applied in the stock closing price and volume time series to study their corresponding complexity relations. At the same time, combined with multi-scale method, the complexity of closing time series and volume time series is measured over different scales. The experiment shows that the complexity of the closing price time series and volume time series can reveal a certain rule of stock development. In addition, the closing price sequence can maintain consistency on different scales, while the entropy value changes of the volume sequence at different scales have different trends, and the closer the stock type is, the closer the entropy change curve is.

Key words: sample entropy; time dependent; multi-scale entropy; stock time series

0 引 言

股票市场是一个比较集中体现经济水平的地方,当企业赢利或者倒退时,直接表现在股价的波动上[1]。同时,在股市中成交量是股票市场的原动力,成交量的变化最能反映股市的大趋势[2]。研究股票价格和成交量的关系可以预测股票市场的走势,具有重要意义。但是,股票市场是一个非常复杂的系统,易受到国家政策、经济形势及个人投资心理等因素的影响[3]。研究股票市场收盘价和成交量时间序列的复杂性可以揭示股票市场运行的内在机制,具有广阔前景[4]。

样本熵是一种应用广泛的非线性方法,用来度量时间序列的复杂度[5-7]。样本熵的算法过程很简单,被定义为序列中 m 点数据段模式互相相似的情况下,$m+1$ 点数据段模式依然互相相似的条件概率负平均

收稿日期: 2018-04-16 **修回日期:** 2018-08-21 **网络出版时间:** 2018-12-20
基金项目: 科技部重点研发计划(2016YFC1401902)
作者简介: 于文静(1990-),女,硕士研究生,研究方向为不确定信息挖掘; 余 洁,副教授,研究方向为网络个性化搜索、用户兴趣建模、基于语义的网络交互计算等; 徐凌宇,教授,研究方向为基于 Web 的远程软件服务技术、网络多源信息融合技术、大规模数据挖掘、数字地球技术。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20181219.1542.062.html

自然对数的精确值[8]。样本熵的计算模型可以表示为: $S=-\ln\left(\frac{P_{m+1}}{P_m}\right)$,也可以表示为 $S=\ln P_m-\ln P_{m+1}$。由此可知,样本熵取决于函数的一步之差,反映的是根据这个时间序列历史数据,下一个新点的不确定性[9]。换句话讲,样本熵实际上衡量的是新信息的产生率。

在研究患病系统输出的时间序列和健康动力系统输出的时间序列的复杂度时发现,具有很多规则信号的患病时间序列的熵值反而比自由运行的健康时间序列的复杂度高。Costa 等认为产生这个问题的原因可能是传统的熵研究方法没有考虑到复杂系统的输出的多尺度特性,因此提出了多尺度熵[10]用来度量不同时间尺度下的样本熵值。随后这种从多个尺度上研究时间序列复杂度的方法在各领域得到了广泛应用[11-13]。

但是,这种方法并未从根本上解决样本熵在度量时间序列复杂度上存在的问题。从样本熵的定义可看出,样本熵取决于函数的一步之差,因此,该方法没有考虑与结构相关的特征。而多尺度熵的方法实际上是改变时间序列的结构,并未从熵方法中解决这一问题。另一方面,由条件概率模型可知,样本熵在计算复杂度时,并没有考虑两个模式下向量的自相似概率的具体值。实际上也就是没有考虑时间序列结构的复杂性。一般来说时间序列自相似程度越高,就意味着时间序列中有很多相似的向量,时间序列的向量模式的种类就少,时间序列相对来说构成不复杂。但样本熵中单纯的条件概率模型,使得不同的时间序列,两种模式下向量的自相似概率也不同,但是样本熵却可能相同。

针对上述两个问题,提出了基于时间依赖的改进样本熵(time dependent modified entropy,TD_entropy)。考虑到时间序列中的向量具有时间属性,因此在向量相似性匹配时,引入了时间衰减函数[14-15],根据时间序列中相似向量之间的时间距离,给出相似性大小。在最后求熵值的模型中,考虑了 $m+1$ 维向量的自相似概率。结合多尺度方法,提出了基于时间依赖的多尺度改进样本熵,从时间序列多个尺度下度量时间序列的复杂性。

1　方　法

1.1　基于时间依赖的改进样本熵方法

改进样本熵的定义和样本熵类似,改进样本熵参数用 N,m,r,k 表示。其中,N 为序列长度,m 为嵌入维数,r 为相似容限,k 为时间衰减系数,具体算法如下:

对于一条长度为 N 的时间序列 $\{u(j):1\leqslant j\leqslant N\}$,组成 $N-m+1$ 个向量 $x_m(i)$,$x_m(i)=\{u(i+k):0\leqslant k\leqslant m-1\}$,其中 $\{i\mid 1\leqslant i\leqslant N-m+1\}$,也就是从 $u(i)$ 到 $u(i+m-1)$ 的 m 个数据点。两向量之间距离被定义为它们对应点距离差的最大值,$d[x(i),x(j)]=\max\{|u(i+k)-u(j+k)|:0\leqslant k\leqslant m-1\}$。如果两个向量 $x_m(j)$ 和 $x_m(i)$ 的距离在 r 范围内,则根据两个向量之间的时间距离,根据时间衰减函数,得到两个向量相似的概率 $\mu_{i,j}^m=\mathrm{e}^{-k\times|i-j|}$。将向量 $x_m(i)$ 和其他向量 $x_m(j)$ 之间的平均相似性记为: $B_i^m(r,k)=\frac{\sum_{i=1}^{N-m}u_{i,j}^m}{N-m-1}$。因此 m 维向量自相似性记为: $B^m(r,k)=\frac{\sum_{i=1}^{N-m}B_i^m}{N-m}$。同样道理,将维数 m 加 1,得到 $B^{m+1}(r,k)$,改进样本熵表示为: $\mathrm{TD\_entropy}(m,r,k)=\lim_{N\to\infty}-\ln\left[\frac{B^{m+1}(r,k)}{B^m(r,k)}\times B^{m+1}(r,k)\right]$,也可以近似表示为: $\mathrm{TD\_entropy}(N,m,r,k)=-\ln\left[\frac{B^{m+1}(r,k)}{B^m(r,k)}\times B^{m+1}(r,k)\right]$。

根据 Pincus 建议,实验中将 m 设为 2,r 为 $0.2\times$ SD,SD 为时间序列的标准差; 同时将衰减系数 k 设为 0.01。

1.2　基于时间依赖多尺度改进样本熵方法

基于时间依赖的多尺度改进样本熵算法,实际上分为两步。首先将原始时间序列进行粗粒化,然后对粗粒化后的时间序列求改进样本熵的值,最后得到不同尺度下的改进样本熵值[16]。

根据尺度因子 τ,构造粗粒化序列 $\{y^{(\tau)}\}$。$y_j^{(\tau)}=1/\tau\sum_{i=(j-1)\tau+1}^{j\tau}x_i$,其中 $1\leqslant j\leqslant N/\tau$。当尺度为 1 时,时间序列 $\{y^{(1)}\}$ 就是原始的时间序列。每一个粗粒化的时间序列的长度等于原始时间序列的长度除以 τ。然后求不同粒度下的基于时间依赖的改进样本熵的值,得到多尺度熵。

2　实验数据

为了研究股票收盘价时间序列和成交量时间序列之间的相应关系,选用股票市场白酒板块的六只股票。这六只股票的收盘价和成交量时间序列的时间跨度为 2010 年 1 月 4 日到 2016 年 12 月 31 日,共 7 年数据(数据来源自网易财经),如表 1 所示。

图 1 是酒鬼酒这只股票 7 年的收盘价序列和成交量序列。从图中可以看出,股票的收盘价的总体趋势和成交量的波动趋势在很大程度上是相似的,收盘价的波峰对应成交量的波峰。在 300 点之前,股价波动持续增长,成交量也相伴而升,是市场继续看好的表现。300 ~ 600 点时,股价依然波动上涨,但是成交量上涨的趋势却不明显,说明升势难于维持。700 点左右时,股价大跌,但成交量上升,说明股票形势不看好。

表 1　六只股票代码及名称

股票代码	股票名称
000596	古井贡酒
000799	酒鬼酒
600199	金种子酒
600519	贵州茅台
600559	老白干酒
600809	山西汾酒

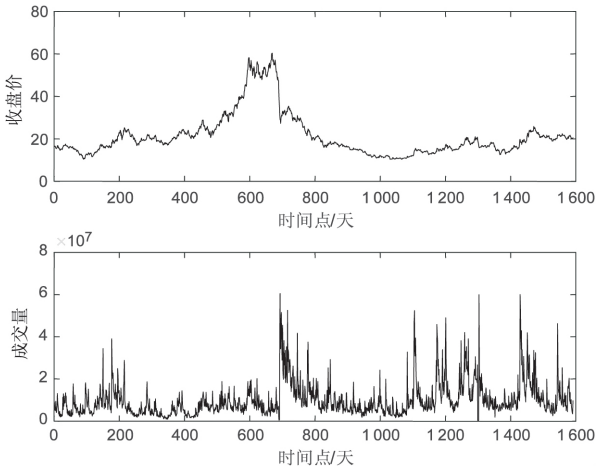

图 1　酒鬼酒股票收盘价和成交量 7 年的时间序列

3　实验过程及结论

3.1　基于改进样本熵的时间序列复杂度研究

图 2 和图 3 是这 6 只股票每一年的收盘价和成交量的基于时间依赖的改进样本熵的熵值序列。

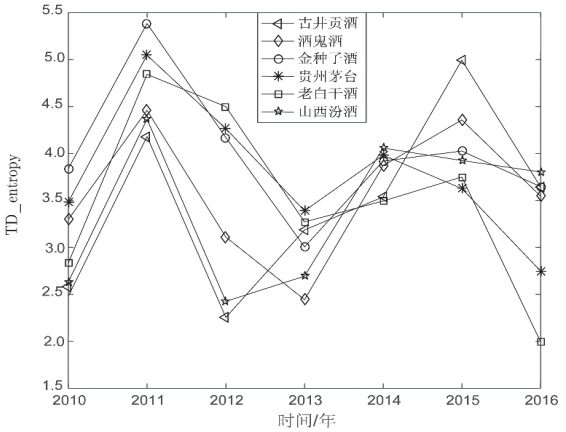

图 2　6 只股票 7 年的收盘价时间序列的复杂度

从图 2 和图 3 中可以看出，不同白酒的每一年的收盘价复杂度和成交量复杂度的变化趋势差不多是一致的。收盘价序列的复杂度在 2011 和 2015 年时熵值最大，产生这一现象的原因是在这两年时间里白酒板块的股票收盘价正常波动变化，多种因素影响其股价变化，股票收盘价序列规则性较小，复杂度相对来说比较大；而在 2013 年熵值几乎处于波谷，导致这一现象的原因可能是 2012 年底酒鬼酒的塑化剂事件对白酒板块股价的影响。塑化剂事件后白酒板块的股票价格受到不同程度的影响，股价跌的程度也不一样，但是总体趋势几乎都是下跌的。从复杂度的角度来讲，这一时期股价波动的规则性比较明显，收盘价序列的复杂度相应来讲就比较小。

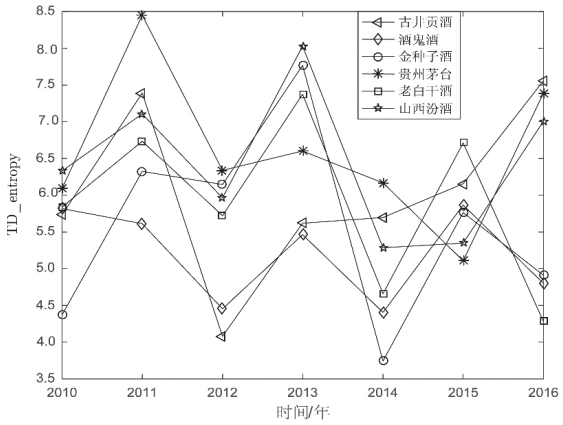

图 3　6 只股票 7 年的成交量时间序列的复杂度

从图 3 的成交量时间序列的复杂度来看，这几只股票在 2011 年成交量的复杂度也达到了一个波峰。因为这段时间股票市场是属于上升期的，股价的增长促进成交量的波动，经济形势被看好，股票市场比较活跃，所以成交量时间序列比较复杂。而 2013 年股票的收盘价复杂度相应的小，成交量的复杂度却相对其他年份来说比较大。这是因为股价发生大跌，或者在经济形势不被看好的情况下，股票市场就会有大量股票被抛售以减少损失，在这种情况下，股票成交量时间序列的复杂度就相对的高。

此外，从这 6 只股票收盘价序列复杂度趋势来看，酒鬼酒和古井贡酒以及山西汾酒序列相对其他序列更加相近；从成交量复杂度变化趋势上看，除了酒鬼酒和贵州茅台以及古井贡酒，其他股票的成交量复杂度趋势变化曲线很相近。

3.2　基于多尺度改进样本熵的时间序列复杂度研究

图 4 和图 5 展示了 6 只股票收盘价时间序列和成交量时间序列在不同尺度下的熵值序列。从图 4 可以看出，这 6 只股票的收盘价序列的复杂度在不同尺度下能够保持一定的一致性。贵州茅台收盘价序列在不同尺度下的熵值都是最大的；而山西汾酒收盘价序列的复杂度在各个尺度下都是最小的。同时可以看出在小尺度下这 6 只股票收盘价序列的熵值呈现出变小的趋势，当尺度增大时，大部分序列的熵值呈现出增大的趋势，除了山西汾酒收盘价序列。

图 5 是这 6 只股票成交量序列在不同尺度下的熵值曲线。从图中可以看出，随着尺度的变化，这 6 只股票的成交量复杂度变化是不同的，且每只股票成交量复杂度的大小关系也随着尺度在变化。但是，老白干酒、酒鬼酒和金种子酒收盘价序列在不同尺度下的熵

值大小及趋势都是很相似的。而很明显的,贵州茅台、山西汾酒以及古井贡酒这三只高端白酒的熵值大小,和其他普通白酒股票相差很大。其中,山西汾酒和古井贡酒的成交量序列在不同尺度下的熵值更加接近,也就意味着复杂度相似,且在大部分尺度上能保持一致性。而贵州茅台的成交量序列在不同尺度下的熵值序列和其他股票的熵值序列明显不同,且熵值随着尺度的增加属于下降的趋势。

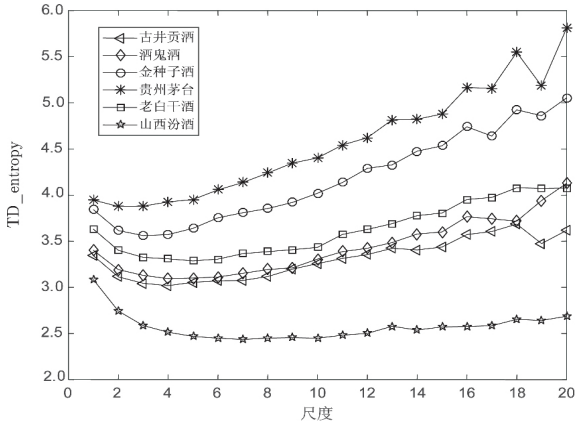

图 4　6 只股票收盘价序列在不同尺度下的熵值序列

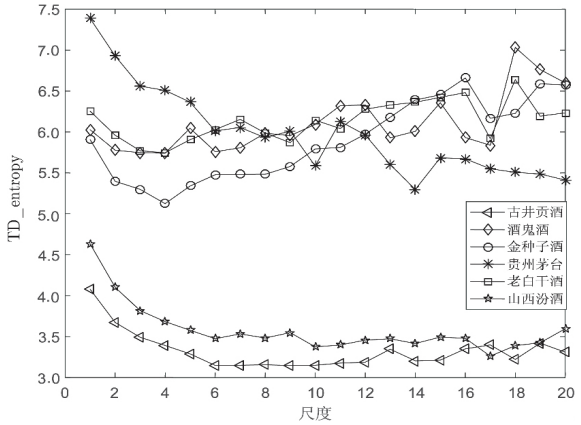

图 5　6 只股票成交量序列在不同尺度下的熵值序列

4　结束语

股票时间序列复杂度的研究具有重要意义,但是传统的熵研究方法在度量时间序列复杂度时存在一些问题。为了更好地研究股票时间序列的复杂度,提出了基于时间依赖的改进样本熵,同时结合多尺度的方法,提出了基于时间依赖的多尺度改进样本熵。

股票市场中的收盘价和成交量是股市中非常重要的要素,且两者存在一定的关系。但是,股票市场是非常复杂的系统,因此,为了更好地研究股票市场的演化,根据收盘价和成交量时间序列的复杂度变化情况,揭示股票的发展形势。结果表明,研究收盘价序列和成交量序列的复杂度变化情况具有一定的意义,通过两者的复杂度变化情况,在一定程度上能够看出股票的发展情况。收盘价序列在不同尺度下熵值能保持一定的一致性; 而收盘价序列随着尺度的变化有不同的趋势。

参考文献:

[1] 王　怡. 上市公司的网络舆情事件演化与股价变动的关联性研究 [D]. 南京: 南京理工大学,2017.

[2] 闫树熙. 上证股市价格与成交量关系的实证研究 [J]. 河南科学,2017,35(2):340-344.

[3] XU Mengjia,SHANG Pengjian,HUANG Jingjing. Modified generalized sample entropy and surrogate data analysis for stock markets [J]. Communications in Nonlinear Science & Numerical Simulation,2016,35:17-24.

[4] 张　群. 金融市场复杂性与金融泡沫现象研究 [D]. 广州: 华南理工大学,2016.

[5] 赵利民,朱晓军. 基于局部均值分解与样本熵的脑电信号特征提取与分类 [J]. 计算机工程,2017,43(2):299-303.

[6] 孙曙光,于　晗,杜太行,等. 基于振动信号样本熵和相关向量机的万能式断路器分合闸故障诊断 [J]. 电工技术学报,2017,32(7):20-30.

[7] 成　娟,陈　勋,彭　虎. 基于样本熵的肌电信号起始点检测研究 [J]. 电子学报,2016,44(2):479-484.

[8] COSTA M,PENG C K,GOLDBERGER A L,et al. Multi-scale entropy analysis of human gait dynamics [J]. Physica A: Statistical Mechanics & Its Applications,2003,330(1-2):53-60.

[9] COSTA M,GOLDBERGER A L,PENG C K. Multiscale entropy analysis of complex physiologic time series [J]. Physical Review Letters,2007,89(6):068102.

[10] COSTA M,HEALEY J A. Multiscale entropy analysis of complex heart rate dynamics: discrimination of age and heart failure effects [J]. Computers in Cardiology,2003,30:705-708.

[11] 介　丹. 脑电信号的多尺度熵分析方法研究 [D]. 太原: 太原理工大学,2017.

[12] WU S D,WU C W,LEE K Y,et al. Modified multiscale entropy for short-term time series analysis [J]. Physica A: Statistical Mechanics & Its Applications,2013,392(23):5865-5873.

[13] WU S D,WU C W,LIN S G,et al. Analysis of complex time series using refined composite multiscale entropy [J]. Physics Letters A,2014,378(20):1369-1374.

[14] 仲兆满,胡　云,李存华,等. 微博中特定用户的相似用户发现方法 [J]. 计算机学报,2016,39(4):765-779.

[15] 王晨旭,管晓宏,秦　涛,等. 微博消息传播中意见领袖影响力建模研究 [J]. 软件学报,2015,26(6):1473-1485.

[16] HUMEAUHEURTIER A. The multiscale entropy algorithm and its variants: a review [J]. Entropy,2016,17(5):3110-3123.