文章编号:1003－207(2011)05－0001－09

基于符号时间序列方法的金融收益分析与预测

徐 梅,黄 超

(天津大学管理与经济学部,天津 300072)

摘 要:引入符号时间序列分析方法从大尺度的角度分析收益变化的特征,提出了确定收益变化的主要模式并预测收益水平的方法。首先将收益序列转化为符号序列,由符号序列中不同的字代表不同的收益变化模式,根据符号序列直方图,可以确定收益变化的主要模式。然后,根据各收益变化模式的概率分布,在前几个时点收益水平确定的情况下,可以推知下一个或几个时点处于不同收益水平的概率,从而实现对收益水平的预测。对上证综指、深证成指以及上证工业股指数、上证商业股指数、上证地产股指数、上证公用事业股指数共六个股票指数的收益序列进行了实证分析,确定了各指数收益的主要变化模式,并基于主要变化模式进行了收益水平的预测,从而说明了该方法的有效性和可行性。

关键词:符号时间序列分析;直方图;收益;主要模式;预测

中图分类号:F224 文献标识码:A

1 前言

金融市场的资产收益往往受宏观经济形势、金融政策、公司财务状况、国际环境以及投资者心理承受能力等多方面的影响,变化极为复杂且难以预测。然而对资产收益的分析和预测不仅可为投资者的投资决策也可以为政府正确地制定各项宏观政策提供依据。因此,如何较为全面地把握资产收益的特征并进行预测是研究人员和市场分析人员的极为感兴趣的课题。

时间序列分析是资产收益分析与预测的重要方法。传统的时间序列分析采用线性模型和非线性模型对资产收益进行建模与预测,常用的线性模型包括自回归滑动平均(ARMA)模型、季节模型,非线性

模型则包括门限自回归模型(TAR)及其扩展模型、马尔科夫转换模型、非线性状态空间模型等[1]。近年来,随着分形市场理论的建立,又出现了一些基于该理论的收益分析方法,如 R/S 分析、多重分形分析等[2,3],特别是长记忆时间序列模型——分整自回归滑动平均(ARFIMA)模型[4]得到了广泛的应用。此外,一些非参数方法如混沌、神经网络、小波分析、支持向量回归机等在收益的分析与预测中也得到应用[5－7]。这些分析方法与模型,有的验证收益序列是否具有分形或混沌特征,有的着重刻画收益随时间变化的规律,有的预测未来具体的收益值。然而这对于准确、全面地把握收益变化的规律来说具有一定的局限性,需要利用新的方法从不同的角度来分析收益的特征和规律。

符号时间序列分析(symbolic time series analysis,STSA)起源于上世纪 90 年代中期,它是以非线性动力学的理论为基础,由符号动力学理论、混沌时间序列分析和信息理论发展起来的一种新的数据分析方法。将数据连续的状态空间划分为少量的离散胞元并对每个胞元分配不同的符号或数值,这是一个"粗粒化"的描述非线性系统的方法,从而将有许多可能值的数据序列变换为仅有几个互不相同值的符号序列。

最先使用 STSA 方法的是 Tang 等(1995,1997)[8,9],Daw 等(2003)对实验数据如何实现符号化分析给出了系统而全面综述[10]。目前已证实 STSA 方法可用于分叉检测、瞬态过程的特征描述、模型拟合、模式识别、受强噪声污染数据的确定性特征检测、分类、故障诊断等方面,应用领域包括天文学、地理学、医药生物、化工、机械、人工智能、控制通讯、数据挖掘等。研究证明利用 STSA 方法可以捕获非线性过程的大尺度特征,大大缩短计算时间,能

收稿日期:2010－10－15;修订日期:2011－07－03

基金项目:国家自然科学基金资助项目(70971097)

作者简介:徐梅(1968－),女(汉族),河北唐山人,天津大学管理与经济学部,博士,副教授,研究方向:金融波动分析、社会经济系统建模与预测.

够很好地抑制噪声，不用假设产生数据的模型是什么，也不用做出数据是否平稳等类似的假设。STSA 方法虽然在自然科学和工程领域得到了广泛的应用，但在金融和经济领域的应用起步较晚，国外有少量的有关 STSA 方法在金融、经济领域应用的文献[11−14]。文献[11]根据生物学中用 STSA 从整个基因序列中寻找特定基因片段的方法，对两个股票指数序列进行比较，采用相似程度的度量指标，分析两序列的相似性及联系，从而识别不同市场间相似的运作模式；文献[12]采用 STSA 及层次树的方法研究了不同汇率时间序列之间的协同运动关系及货币危机的传染性；文献[13]采用 STSA 及多维最小跨度树的方法将美国上市公司依据资产收益和交易量两种信息进行分类，从而确定股票市场结构；文献[14]研究了直方图时间序列的预测问题，直方图分析属于符号数据分析的领域，并指出由于高频金融时间序列的特殊性，直方图时间序列特别适合于描述其各个区间内的整体分布特性，直方图时间序列的预测结果会对交易者制定交易策略很有帮助。国内只查到一篇应用 STSA 方法研究金融、经济问题的文献[15]，该文用符号时间序列分析法，构造了上海板块股票网络的最小生成树和分层树，从上市公司之间关系的角度分析和解释了分层聚类结构的结果。

金融市场本质上是一个非线性系统，诸如混沌、分形等都是金融市场的非线性本质特征[16,17]，而 STSA 方法正适合于分析非线性动力学系统。本文的创新之处在于引入符号时间序列分析方法，提出以收益符号序列中不同的字代表不同的收益变化模式，根据符号序列直方图，揭示收益变化的特征并确定主要收益变化模式。在此基础上，提出根据前几个时点收益所处的区间或收益水平，推知下一个或几个时点收益处于各区间的概率的方法，从而实现对收益水平的预测。该方法从大尺度的角度分析资产收益的变化规律，是对已有的收益分析方法的扩展和补充。收益变化模式的概率分布、主要收益模式的确定以及对收益水平而不是具体收益值的预测，这些都是已有收益分析方法所无法得到的，相关的研究在国内外还未见报道。该方法对于从新的角度全方位地了解收益变化的规律，从而为市场监管和调控及投资决策提供依据具有重要意义。用该方法对上证综指、深证成指以及上证工业股、商业股、地产股、公用事业股共 6 个指数收益序列进行实证分析，确定主要变化模式，并做出基于主要变化模式的收益水平预测，用以验证该方法的可行性和有效性。

2 符号时间序列分析

在引入符号时间序列分析方法的同时，提出以收益符号序列中不同的字代表不同的收益变化模式，根据符号序列直方图确定主要收益变化模式。

2.1 时间序列符号化

时间序列符号化的基本思想是将原始时间序列（或由原始时间序列转换得到的序列，如连续时间点间的一阶差分序列）划分为有限个数的区间，每个区间分配不同的符号，每个原始数据根据落入区间的不同对应不同的符号[14]。

对于金融时间序列 $\{x_t, t=0,\cdots,N-1\}$，引入划分 $P=\{P_1,P_2,\cdots,P_{n-1}\}$ 将序列分割成 n 个互不重叠的区间，$n>1$。每个区间用一个符号 $S_i\in\{S_1,S_2,\cdots S_n\}$ 进行标记，如可用符号集 $\{0,1,2,\cdots,n-1\}$ 来标记，也可用任何其他的符号集来标记，则可将数据序列 $\{x_t, t=0,\cdots,N-1\}$ 转换为符号序列 $\{s_t, t=0,\cdots,N-1\}$，其中

$$s_t=\begin{cases}S_1 & x_t\leqslant P_1\\ S_i & P_{i-1}<x_i\leqslant P_i,\ i=2,\cdots,n-1\\ S_n & x>P_{n-1}\end{cases} \tag{1}$$

如果原序列 $\{x_t, t=0,\cdots,N-1\}$ 中不同时间点的数据值落入同一个区间，则被转换为同一符号。这些符号标记了序列值所属的区间，描述了序列的动态特征。可能的符号的个数 n 称为符号集的大小。最简单的情况二进制划分，有两个可能的符号 0、1，即 $n=2$。当 n 不断增大，原始时间序列的更多细节被包括进来，同时更多的噪声也被包括进来。极限的状态是，n 等于时间序列不同值的数目，此时符号化没有造成任何信息的损失，符号序列和原始时间序列包含相同的信息，二者是等价的，只是数据表示不同而已（如用符号代替具体的数值）。因此应根据符号化分析中需要保留多少原始序列的信息来确定符号集的大小 n。

对于金融序列 $\{x_t, t=0,\cdots,N-1\}$，可采用序列样本的均值、中值作为不同符号之间的划分，或将整个样本的数据范围分为相等大小的不同区间以此确定符号划分，也可依据使每个符号的出现是等概率的进行划分。这些针对原始序列的符号化方法称为静态符号化方法；当原始序列非平稳或原始数据随时间的变化比其绝对数值更重要时，通常针对原

始序列的一阶或高阶差分序列进行符号化,称为动态符号化方法。动态符号化方法对突发的强噪声不敏感,除针对的是原始序列的差分序列外,与静态符号化方法是一样的。数据符号化是一个"粗粒化"的过程,即只捕获大尺度的特征,从而降低噪声对统计算法的影响。

2.2　时间序列符号化分析

一旦原始时间序列转化为符号序列后,就要提取符号序列的特征量对其进行定量分析。

2.2.1　符号序列的编码

生成符号序列后,为了计算其统计量,通常需要对符号序列进行编码,以方便地实现符号序列的数值表示。对于符号集大小为 n,长度为 N 的符号时间序列,选择一个标准长度 L, L 个连续的符号组成一个字,称 L 为字长。每一个字被编码为十进制数,形成新的序列。设长度为 N 的符号时间序列为 $\{s_0,s_1,\cdots,s_{L-1},s_L,s_{L+1},\cdots,s_{N-1}\}$,其一般编码方法可概括为:

(1)依次按顺序取 L 个连续的符号数据组成一个字,即取第 0 到第 $L-1$ 个符号数据为第 0 个字,第 1 到第 L 个符号数据为第 1 个字,…,第 N-L 到第 N-1 个数据为第 N-L 个字,N-$L+1$ 个字构成一个新序列 $\{s_0s_1\cdots s_{L-1},s_1s_2\cdots s_L,s_2s_3\cdots s_{L+1},\cdots,s_{N-L}s_{N-L+1}\cdots s_{N-1}\}$。

(2)对新序列中每一个字进行编码,则可得十进制序列代码构成的编码序列 $\{S_0,S_1,S_2,\cdots,S_{N-L}\}$。其中 $S_i=s_{i+L-1}\times n^0+s_{i+L-2}\times n^1+\cdots+s_i\times n^{L-1}$, $i=0,1,\cdots,N-L$。

所有不同字的总数满足:

$$K=n^L \tag{2}$$

其中 n 为符号集大小,L 为字长,可见 S_i 是满足 $0\leqslant S_i\leqslant K$ 的任一整数。

例如,设符号集大小 $n=3$,有 3 个可能的符号 0、1、2,符号序列长度 $N=10$ 的符号序列为:2012110120,选取字长 $L=4$,则每个字构成的新序列为:{2012,0121,1211,2110,1101,1012,0120},十进制序列代码构成的编码序列为:{59,16,49,66,37,32,15},不同字的总数 $K=81$。可以通过分析编码序列中每一个字出现的相对频率,来揭示序列的动力学特性。对于给定的时间序列,选择适当的 n 与 L 值能更好地揭示其动力学特性。

在进行收益序列的分析时,本文提出以收益符号序列中不同的字代表不同的收益变化模式,某个字长为 L 的字对应的收益变化模式表明了连续 L 个时点收益所处的区间或收益水平。例如取字长 $L=4$,符号集大小 $n=3$,分别用三个符号 0、1、2 表示低、中、高三种收益水平,则字 1021 表示在四个连续的时间点上,收益的变化模式是中收益——低收益——高收益——中收益。在此基础上将收益符号序列中出现概率较大的字对应的变化模式称为主要收益变化模式。

2.2.2　字长的选取

编码时,需要先确定编码字长 L,即每个编码所包含的符号数据点个数。根据文献[10],可以利用改进 Shannon 熵选择合适的 L。改进 Shannon 熵定义为

$$H(L)=-\frac{1}{\log_2 N_{obs}}\sum_i P_{i,L}\log_2 P_{i,L} \tag{3}$$

其中 N_{obs} 是符号序列中出现的不同字的数量,即出现概率为非 0 的字的数量而不是所有可能字的总数,$N_{obs}\leqslant K$;i 为字的编号,即十进制序列代码,$P_{i,L}$ 是字长为 L 的第 i 个字出现的概率,且令 $0\log_2 0=0$。对于任意概率向量 $(P_{1,L},P_{2,L},\cdots,P_{K,L})$,$0\leqslant H(L)\leqslant 1$。$H(L)=1$,当且仅当 $(P_{1,L},P_{2,L},\cdots,P_{K,L})=(1/K,1/K,\cdots,1/K)$;$H(L)=0$,当且仅当对某个 i,$P_{i,L}=1$。可见,对于完全随机的过程产生的序列,所有字出现的概率相等,$H(L)=1$;不完全随机的序列,$H(L)$ 介于 0 和 1 之间;对于完全确定的过程产生的序列,所有的概率集中到某一字上,$H(L)=0$。

可见 $H(L)$ 越大,表明序列中每种变化模式出现的概率越接近,序列的随机性越大,变化规律越复杂;$H(L)$ 越小,表明序列中某些变化模式出现的概率更大,主要变化模式更明显,序列具有更确定性的结构,因而熵 $H(L)$ 是一个评价序列相对复杂性和随机性的指标。

令 L 从 1 开始增加,按式(3)计算的 $H(L)$ 值会逐渐达到一个最小值。当 $H(L)$ 达到最小值的时候,序列模式中的非随机部分最为明显,此时便于我们发现其中的确定性信息。因此可以认为 $H(L)$ 最小值对应的 L 值对于给定的序列和符号集大小 n 来说,是一个最优的选择。如果符号序列的字长 L 太小,会丢失某些重要的确定性信息,使得统计结果的实际意义不大;如果符号序列的字长 L 太大,则每个字出现的频率太小,无法得到可靠的统计结果。

2.2.3　符号序列直方图

对于由十进制序列代码构成的编码序列,可以用直方图表示编码序列中各个字出现的概率分布。

它以字的编号 $i(i=0,1,2,\cdots,n^L-1)$ 作为横坐标，即横坐标每个小区间的长度为 1，以编码序列中编号为 i 的字出现的相对频数作为纵坐标，这种符号统计量的表示方法称为符号序列直方图。符号序列编码后，某个字 i 出现的相对频数可由编码序列中字 i 出现的次数与编码序列长度之比计算得到。根据等概率划分原则可以断定，对真正的随机数据序列，每个字出现的相对频数是相等的，所有的直方图竖条的高度是相同的。反之，符号序列直方图中任何从等概率性的明显偏移就表示系统模式的某种程度的确定性。

收益符号序列直方图可以直观地表明每个字在编码序列中出现的相对频数，每个字表示某种确定的收益变化模式，因此收益符号序列直方图可以表明每种变化模式在序列中出现的相对频数。如果某字出现的相对频数较大，则该字所对应的变化模式是序列的主要变化模式，因此由收益符号序列直方图，可以确定收益序列的主要变化模式。

3 中国股票市场收益的符号化分析

本文首先选取上海证券交易所综合指数（简称上证综指）和深圳证券交易所得成分指数（简称深证成指）的每日收盘价作为样本序列，由这两个序列分别代表中国两个股票市场的价格。上证综指的样本时间段为：1991/07/15—2010/04/16，共 4594 个样本数据；深证成指的样本时间段为：1995/01/23—2010/04/16，共 3688 个样本数据。另外，为了进行行业间的对比，又选取了上证工业股、商业股、地产股和公用事业股指数作为样本序列，样本时间段为：1993/06/01—2010/04/16，共 4114 个样本数据。

设 t 时的价格为 P_t，收益定义为价格对数的一阶差分，即：

$$R_t = \log P_t - \log P_{t-1} \tag{4}$$

上证综指、深证成指的收益序列分别记为 $\{R1_t\}$、$\{R2_t\}$，上证工业股、商业股、地产股和上证公用事业股指数的收益序列分别记为 $\{R3_t\}$，$\{R4_t\}$，$\{R5_t\}$，$\{R6_t\}$。

3.1 收益序列的符号化

对于收益序列 $\{R_t\}$ 采用静态符号化方法，将其转化为符号序列 $\{s_t\}$。对任意符号集大小 n，依据使每个符号的出现是等概率的进行划分，即对于长度为 N 的收益序列 $\{R_t\}$，根据 $\{R_t\}$ 中数值的大小，位于某一数值区间的 N/n 个数据转化为同一符号。本文取符号集大小 $n=3$，分别用 0、1、2 表示 3 个数值区间所对应的符号，以 1/3 分位数和 2/3 分位数作为三个区间的划分，以便尽可能地使符号化后的符号序列中三个符号的出现是等概率的。即

$$s_t = \begin{cases} 0 & R_t \leqslant R_{1/3} \\ 1 & R_{1/3} < R_t \leqslant R_{2/3} \\ 2 & R_t > R_{2/3} \end{cases} \tag{5}$$

其中 $R_{1/3}$、$R_{2/3}$ 分别表示 $\{R_t\}$ 的 1/3 分位数和 2/3 分位数。可见符号 0、1、2 分别表示收益处于低、中、高三种不同的收益水平。这种对于收益序列的静态符号划分方法与对价格对数序列的动态划分方法是一致的。

6 个收益序列 $\{R1_t\}\sim\{R6_t\}$ 的最小值、1/3 分位数、2/3 分位数和最大值列于表 1 中，其所对应的符号序列分别记为 $\{s1_t\}\sim\{s6_t\}$。

表 1 收益序列的分位数

收益序列	最小值	1/3 分位数	2/3 分位数	最大值
$\{R1_t\}$	−0.1791	−0.005	0.0066	0.7192
$\{R2_t\}$	−0.1841	−0.0053	0.0066	0.2108
$\{R3_t\}$	−0.1966	−0.0054	0.0061	0.2745
$\{R4_t\}$	−0.1894	−0.0053	0.0064	0.2849
$\{R5_t\}$	−0.1474	−0.0072	0.007	0.2797
$\{R6_t\}$	−0.1907	−0.0055	0.0062	0.3371

3.2 收益符号序列字长 L 的确定及熵值分析

对于每一收益符号序列 $\{s1_t\}\sim\{s6_t\}$，令字长 L 从 1 开始增加，按式(3)计算改进 Shannon 熵值，其中 $P_{i,L}$ 可由字长为 L 的编码序列中字 i 出现的次数与编码序列长度之比计算得到。熵值计算结果如表 2 所示。

表 2 收益符号序列的改进 Shannon 熵值

收益符号序列 \ L	1	2	3	4
$\{s1_t\}$	0.9998	0.9970	0.9938	0.9882
$\{s2_t\}$	1.0000	0.9984	0.9963	0.9929
$\{s3_t\}$	1.0000	0.9988	0.9967	0.9930
$\{s4_t\}$	1.0000	0.9988	0.9975	0.9940
$\{s5_t\}$	1.0000	0.9990	0.9974	0.9951
$\{s6_t\}$	1.0000	0.9988	0.9974	0.9945

根据 2.2.2，改变字长 L，Shannon 熵最小值对应的 L 值对于给定的序列和符号集大小来说，是一个优化的选择。从表 2 中可以看出，随着 L 的不断增加，每个收益符号序列的 Shannon 熵值是不断下降的，当 $L=4$ 时，Shannon 熵仍然没有达到最小值。由于样本容量的限制，当 L 过大时，由于编码的个数过多，使得每个字出现的频数太小，按式(3)

计算 Shannon 熵时 $P_{i,L}$ 的误差过大,无法得到可靠的熵值,另外在后面的收益水平预测时也会增加计算误差,基于以上考虑,对于每个收益符号序列都取 $L=4$。

对比上证综指和深证成指收益符号序列$\{s1_t\}$和$\{s2_t\}$的熵值,对于所有的 L 取值,$\{s2_t\}$的熵值都大于$\{s1_t\}$的熵值,这说明深市收益序列中每种变化模式出现的概率更接近,收益序列的随机性更强;沪市收益序列中主要变化模式更明显,收益序列的确定性强于深市。

比较上证工业股、商业股、地产股和公用事业股指数的收益符号序列$\{s3_t\}$——$\{s6_t\}$与上证综指收益符号序列$\{s1_t\}$的熵值,发现$\{s3_t\}$——$\{s6_t\}$的熵值均大于$\{s1_t\}$的熵值,说明上证工业股、商业股、地产股和公用事业股指数收益序列中各种变化模式出现的随机性更大,而上证综指收益序列中某些变化模式出现的概率更大,主要变化模式更明显,序列的确定性更强。

最后,对比上证工业股、商业股、地产股和公用事业股指数收益符号序列$\{s3_t\}$——$\{s6_t\}$的熵值,发现总体差别不大,其中地产股指数$\{s5_t\}$的熵值最大,工业股指数$\{s3_t\}$的熵值最小。说明这几个指数收益序列中各种变化模式出现的随机性差别不大,相比之下地产股收益的随机性更大。

另外,随着 L 的增大,6 个指数收益符号序列熵值之间的差别是逐渐增大的,说明各个指数收益序列的复杂性或随机性的差别逐步显现出来。这里选取 L=4 也正能更好地体现各序列随机性之间的差别。

3.3　收益符号序列直方图

收益符号序列的改进 Shannon 熵可以从总体上反映序列中各种变化模式出现的概率情况,熵值越小,则某些变化模式出现的概率更大,但是究竟哪些变化模式是出现概率较大的主要变化模式,各种变化模式出现的概率分布情况如何,则需要通过收益符号序列直方图反映出来。

对于某一收益符号序列,在确定了合适的字长 L 值之后,对其进行编码,根据编码序列可以得到符号序列直方图。图 1—图 6 分别为上证综指、深证成指、上证工业股、商业股、上证地产股和公用事业股指数收益的符号序列直方图。图中,横坐标表示字的编号,为绘图方便起见,字的编号从 1 开始,由于字长 $L=4$,符号集大小 $n=3$,因此所有不同字的总数 $K=81$。这样横坐标 1 代表第 1 个字,其所对应的十进制序列代码为 0,对应的字为 0000;横坐标 2 代表第 2 个字,对应的十进制序列代码为 1,对应的字为 0001;…;81 代表第 81 个字,对应的十进制序列代码为 80,对应的字为 2222。纵坐标表示某个字在编码序列中出现的相对频数,可由该字在编码序列中出现的次数与编码序列长度之比计算得到。

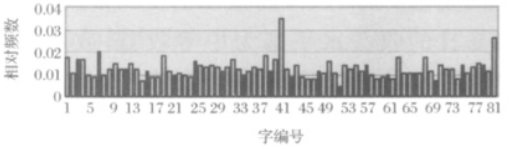

图 1　上证综指收益符号序列直方图

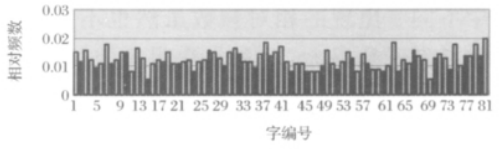

图 2　深证成指收益符号序列直方图

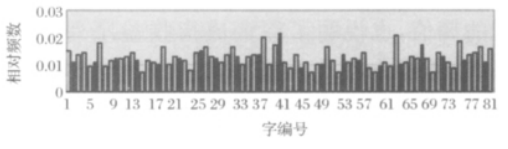

图 3　上证工业股指数收益符号序列直方图

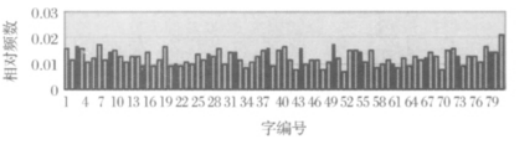

图 4　上证商业股指数收益符号序列直方图

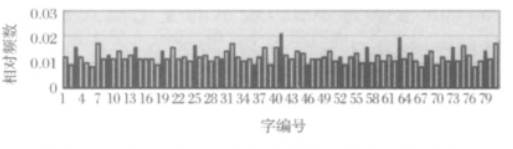

图 5　上证地产股指数收益符号序列直方图

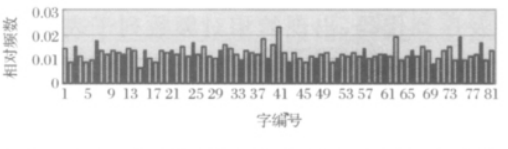

图 6　上证公用事业股指数收益符号序列直方图

3.4　收益序列的主要变化模式

由图 1——图 6 的收益符号序列直方图,可以确定各收益序列的主要变化模式。

由图 1 的上证综指收益符号序列直方图可见,

不同字出现的相对频数差别较大，其中相对频数最大的两个字分别是编号为 41 和 81 的字，它们对应的十进制序列代码为 40 和 80，分别代表字 1111 和字 2222，出现的频率分别为 0.0346 和 0.0264。因此，在所有的 81 种变化模式中，上证综指收益序列两个最主要的变化模式分别是字 1111 和字 2222 所对应的变化模式，字 1111 表示的变化模式是连续四个交易日的收益水平都为中等收益，收益值在 −0.005 和 0.0066 之间波动；字 2222 则表示连续四个交易日均为高收益，收益在大于 0.0066 的区间波动。

与图 1 相比，图 2 的深证成指收益符号序列直方图中不同字出现的相对频数虽然也不尽相同，但是总体差别不是很大，没有非常明显的相对频数较大的字，也即主要变化模式不是很明显，因此与上证综指相比，深证成指收益序列的随机性更强，这与表 2 所示的熵值是一致的。在表 2 中，当 $L=4$ 时，深证成指收益符号序列的熵值大于上证综指收益符号序列的熵值，也说明了深证成指收益序列的随机性更强，主要变化模式没有上证综指收益序列的主要变化模式明显。尽管如此，仍然可以取相对频数较大的字所对应的变化模式作为主要变化模式，它们是编号为 81、38、63 的字，对应的十进制序列代码分别为 80、37、62，出现的相对频数分别为 0.0195、0.0185、0.0185。81 号字为 2222，表示的变化模式是连续四个交易日都为高收益，收益在大于 0.0066 的区间波动；38 号字为 1101，表示的变化模式是第一、二、四个交易日为中等收益，收益在 −0.0053 和 0.0066 之间，第三个交易日为低收益，收益低于 −0.0053；63 号字为 2022，表示的变化模式是第一、第三、第四个交易日为高收益，收益高于 0.0066，第二个交易日为低收益，收益低于 −0.0053。

由图 3——图 6 同样可以确定上证工业股、商业股、地产股和公用事业股指数收益序列的主要变化模式。6 个指数收益序列的主要变化模式、对应的字及序列代码、出现的相对频数列于表 3 的前 5 列中。

由表 3 可见，6 个指数收益序列的主要变化模式大都集中在字 1111、2222 和 2022 所对应的变化模式上。除深证成指和上证商业股指数外，其他 4 个指数收益序列的主要变化模式都包括字 1111 所对应的模式；上证综指、深证成指、上证商业股指数收益序列的主要变化模式都包括字 2222 所对应的模式；除上证综指、上证商业股指数外，其他 4 个指数收益序列的主要变化模式都包括字 2022 所对应的模式。另外，上证工业股，地产股和公用事业股 3 个指数收益序列中出现频数最大的主要变化模式完全相同，都是字 1111 和 2022 所对应的变化模式。由此可见，这几个指数收益序列的主要变化模式具有较高的一致性。

4 基于主要变化模式的中国股票市场收益水平预测

在由收益符号序列直方图确定了每种收益变化模式出现的相对频数的基础上，本文提出根据前几个时点收益所处的区间或收益水平，推知下一个或几个时点收益处于各区间或不同水平的概率的方法，从而实现对收益水平的预测。

4.1 收益水平预测原理

符号序列直方图表明了每个字在序列中出现的相对频数，当样本量足够大时，符号序列直方图纵坐标表示的相对频数也可近似为某字在序列中出现的概率。设字长 $L=4$，符号集大小 $n=3$，分别用 0、1、2 表示 3 个数值区间所对应的符号，根据全概率公式有：

$$P(b_4 \mid b_1b_2b_3) = \frac{P(b_1b_2b_3b_4)}{P(b_1b_2b_3)} = \frac{P(b_1b_2b_3b_4)}{P(b_1b_2b_30)+P(b_1b_2b_31)+P(b_1b_2b_32)} \tag{6}$$

其中，$b_1b_2b_3b_4$ 表示任一字长为 4 的字，$P(b_1b_2b_3b_4)$ 表示字 $b_1b_2b_3b_4$ 在序列中出现的概率；$P(b_1b_2b_3)$ 表示字的前 3 位为 $b_1b_2b_3$ 的概率；$P(b_4 \mid b_1b_2b_3)$ 表示当一个字的前 3 位为 $b_1b_2b_3$ 时，第 4 位出现 b_4 的概率；$b_1b_2b_3i$ $(i=0,1,2)$ 表示前 3 位为 $b_1b_2b_3$ 且第 4 位的值为 i 的字，$P(b_1b_2b_3i)$ 表示字 $b_1b_2b_3i$ 在序列中出现的概率。可见，根据符号序列直方图中表明的每个字在序列中出现的概率及式(6)，可计算出当一个字的前 3 位确定的情况下，第 4 位出现某个符号的概率。

收益符号序列直方图表明了每种收益变化模式出现的相对频数，在字长 L 确定的情况下，某种收益变化模式表明了连续 L 个时点收益所处的区间或收益水平。例如，在分析日收益序列时，当字长 $L=4$ 时，某种变化模式可以反映出连续 4 个交易日收益所处的数值区间或收益水平，因此依式(6)可根据前 3 个交易日收益所处的数值区间或收益水平，计算出第 4 个交易日收益处于各个不同数值区间或取得不同收益水平的概率，以此实现由前 3 个交易

日的收益水平对下 1 个交易日收益水平的预测。

以上证综指收益序列为例,字 0000、0001、0002 出现的概率分别为 0.0176、0.0107 和 0.0166。根据式(6)可计算出当前 3 个交易日的收益连续为低收益(由 $b_1b_2b_3 = 000$ 表示)时,下一个交易日取得不同收益水平的概率,即 $P(0 \mid b_1b_2b_3) = 0.39$,$P(1 \mid b_1b_2b_3) = 0.24$,$P(2 \mid b_1b_2b_3) = 0.37$。由此可预测当前 3 个交易日收益均为低收益或收益值小于 −0.005 时,下一个交易日为低收益(收益值小于 −0.005)的概率是 0.39、中收益(收益值在 −0.005 和 0.0066 之间)的概率是 0.24、高收益(收益值大于 0.0066)的概率是 0.37。又如,前 3 个交易日均为高收益(由 $b_1b_2b_3 = 222$ 表示),要预测下一个交易日为低收益的概率,根据符号序列直方图中字 2220、2221、2222 出现的概率分别为 0.0139、0.0111 和 0.0264,由式(6)可得 $P(0 \mid b_1b_2b_3) = 0.27$,即下一个交易日为低收益(收益值小于 −0.005)的概率为 0.27。

同理,根据全概率公式可得式(7)。

$$P(b_3b_4 \mid b_1b_2) = \frac{P(b_1b_2b_3b_4)}{P(b_1b_2)} = \frac{P(b_1b_2b_3b_4)}{P(b_1b_200)+P(b_1b_201)+P(b_1b_202)+P(b_1b_210)+P(b_1b_211)+P(b_1b_212)+P(b_1b_220)+P(b_1b_221)+P(b_1b_222)} \tag{7}$$

根据式(7),可由前两个交易日的收益水平(由 b_1b_2 表示),预测后两个交易日出现各种收益水平(由 b_3b_4 表示)的概率。根据类似的全概率公式,还可以计算 $P(b_2b_3b_4 \mid b_1)$、$P(b_3 \mid b_1b_2)$、$P(b_2 \mid b_1)$、$P(b_2b_3 \mid b_1)$ 等,由此实现由前几日或一日的收益水平对后一日或几日不同收益水平出现概率的预测。

4.2　基于主要变化模式的收益水平预测

对于表 3 给出的各指数收益序列的主要变化模式,根据式(6),表 3 的最后一列给出了当前 3 个交易日的收益变化符合主要变化模式时,下一个交易日仍符合主要变化模式的概率。例如,字 1111 所对应的变化模式是上证综指收益序列的最主要变化模式,它表示连续 4 个交易日均为中等收益,则当前 3 个交易日均为中等收益时,可预测下一个交易日仍为中等收益的概率是 0.54。

表 3　各个指数收益序列主要变化模式分析

指数	序列代码	相对频数	字	主要变化模式	$P(b_4 \mid b_1b_2b_3)$
上证综指	40	0.0346	1111	连续四个交易日均为中等收益,收益值在 −0.005 和 0.0066 之间波动	$P(1 \mid 111) = 0.54$
	80	0.0264	2222	连续四个交易日均为高收益,收益值在大于 0.0066 的区间波动	$P(2 \mid 222) = 0.51$
深证成指	80	0.0195	2222	连续四个交易日均为高收益,收益值在大于 0.0066 的区间波动	$P(2 \mid 222) = 0.38$
	37	0.0185	1101	第一、二、四个交易日为中收益,收益值在 −0.0053 和 0.0066 之间,第三个交易日为低收益,收益值小于 −0.0053	$P(1 \mid 110) = 0.40$
	62	0.0185	2022	第一、三、四个交易日为高收益,收益值大于 0.0066,第二个交易日为低收益,收益值小于 −0.0053	$P(2 \mid 202) = 0.51$
上证工业股指数	40	0.0212	1111	连续四个交易日均为中等收益,收益值在 −0.0054 和 0.0061 之间波动	$P(1 \mid 111) = 0.43$
	62	0.0209	2022	第一、三、四个交易日为高收益,收益值大于 0.0061,第二个交易日为低收益,收益值小于 −0.0054	$P(2 \mid 202) = 0.51$
上证商业股指数	80	0.0207	2222	连续四个交易日均为高收益,收益值在大于 0.0064 的区间波动	$P(2 \mid 222) = 0.42$
	49	0.0173	1211	第一、三、四个交易日的收益为中等收益,收益值在 −0.0053 和 0.0064 之间,第二个交易日为高收益,收益值大于 0.0064	$P(1 \mid 121) = 0.44$
上证地产股指数	40	0.0212	1111	连续四个交易日为中等收益,收益值在 −0.0072 和 0.007 之间波动	$P(1 \mid 111) = 0.42$
	62	0.0192	2022	第一、三、四个交易日为高收益,收益值大于 0.007,第二个交易日为低收益,收益值小于 −0.0072	$P(2 \mid 202) = 0.45$
上证公用事业股指数	40	0.0231	1111	连续四个交易日均为中等收益,收益值在 −0.0055 和 0.0062 之间波动	$P(1 \mid 111) = 0.45$
	62	0.0192	2022	第一、三、四个交易日为高收益,收益值大于 0.0062,第二个交易日为低收益,收益值小于 −0.0055	$P(2 \mid 202) = 0.45$

由表 3 可见,某种收益变化模式出现的概率大,并不代表在已知前 3 个交易日收益变化符合该变化模式时,下一个交易日仍符合该变化模式的概率也大。例如,对深证成指收益符号序列来说,字 2222 比 2022 出现的概率大,但是 $P(2 \mid 222)$ 却比 $P(2 \mid 202)$ 小。另外,表 3 中 $P(s_4 \mid s_1s_2s_3)$ 的计算结果均大于 0.4,这说明在已知前 3 个交易日收益变化符合主要变化模式的情况下,下一个交易日仍符合该模

式的概率大于平均水平(由于符号集大小 $n=3$,因此平均水平所对应的概率为 1/3)。

5 结语

本文将符号时间序列分析方法引入金融资产收益的分析中,在符号集大小 n 和字长 L 确定的情况下,提出由收益符号序列的各个字代表各种收益变化模式,各种收益变化模式在收益序列中出现的概率则可由收益符号序列直方图反映出来。根据各种收益变化模式的概率分布,可以发现收益变化的规律、揭示收益变化的特征、确定收益的主要变化模式。在此基础上,提出在前几个时点收益所处的数值区间或收益水平确定的情况下,计算下一个时点收益处于各个不同数值区间或取得不同收益水平的概率的方法,以此实现由前几个时点的收益水平对下一个时点收益水平的预测,并由此推广到由前几个或一个时点的收益水平预测后一个或几个时点出现不同收益水平的概率。该方法的优点是计算简便、直观,与已有收益分析方法不同,它是从大尺度的角度反映收益变化的规律,实现对收益水平而不是具体收益值的预测,因此可以作为已有收益分析方法的扩展和补充。该方法的应用对于全方位地把握收益变化的规律,从而为市场监管和调控及投资决策提供依据具有重要意义。

对中国股票市场的上证综指、深证成指以及上证工业股、商业股、地产股、公用事业股等 6 个指数的日收益序列进行了实证分析,在符号集大小 $n=3$ 和字长 $L=4$ 的情况下,确定了各指数收益序列的主要变化模式,并计算出当前 3 个交易日的收益变化符合主要变化模式时,下一个交易日仍符合主要变化模式的概率。实证分析验证了该方法的可行性和有效性,并得出了有益的结论:不是所有的指数收益序列都有特别明显的主要变化模式,例如上证综指收益序列有两个特别明显的主要变化模式,而深证成指收益序列的主要变化模式则不是很明显,主要变化模式越明显,说明序列的确定性越强,反之,则说明序列的随机性更强;6 个指数收益序列的主要变化模式具有较高的一致性,大都集中在字 1111、2222 和 2022 所对应的变化模式上,即连续四个交易日均为中等收益、连续四个交易日均为高收益、第 2 个交易日为低收益其他 3 个交易日为高收益。

符号集大小 n 的确定是时间序列符号化的关键,n 越大,收益的划分区间越多,计算结果越详细,但随着 n 的增大,更多的噪声也会被包括进来从而部分地丧失符号化方法能很好地抑制噪声的优点,在实际应用中,应根据需要选择合适的 n;理论上,字长 L 的选取应根据改进 Shannon 熵的计算结果进行,但由于样本容量的限制,本文取 $L=4$,对于不同的 L 和 n 的取值,虽然计算结果不同,但都可以得出有益的结论;由于符号时间序列分析方法不用假设产生数据的模型是什么,也不用做出数据是否平稳等类似的假设,因此本文方法即可用于收益序列的分析也可用于其他金融时间序列(如波动序列)的分析中。本文是以日收盘价作为样本数据的,该方法也同样可以对高频和超高频数据进行分析。

参考文献:

[1] Tsay, R. S.. Analysis of Financial Time Series [M]. 北京:机械工业出版社, 2006.

[2] 张晓莉, 严广乐. 中国股票市场长期记忆特征的实证研究[J]. 系统工程学报, 2007, 22(2): 190－194.

[3] Wang, Y., Liu, L., Gu, R.. Analysis of efficiency for Shenzhen stock market based on multifractal detrended fluctuation analysis [J]. International Review of Financial Analysis, 2009, 18(5): 271－276.

[4] Bhardwaj, G., Swanson, N. R.. An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series [J]. Journal of Econometrics, 2006, 131(1): 539－578.

[5] 李红权, 马超群, 邹琳. 中国证券市场的混沌动力学特征研究[J]. 中国管理科学, 2005, 13(z1): 194－200.

[6] Connor, N. O., Madden, M. G.. A neural network approach to predicting stock exchange movements using external factors [J]. Knowledge-based System, 2006, 19(5): 371－378.

[7] 辛治运, 顾明. 基于最小二乘支持向量机的复杂金融时间序列预测[J]. 清华大学学报(自然科学版), 2008, 48(7): 1147－1149.

[8] Tang, X. Z., Tracy, E. R., Boozer, A. D., et al. Symbol sequence statistics in noisy chaotic signal reconstruction [J]. Physical Review E, 1995, 51(5): 3871－3889.

[9] Tang, X. Z., Tracy, E. R., Brown, R.. Symbol statistics and spatio-temporal systems [J]. Physica D, 1997, 102(3): 253－261.

[10] Daw, C., Finney, C., Tracy, E.. A review of symbolic analysis of experimental data [J]. Review of Scientific Instruments, 2003, 74(2): 916－930.

[11] Takuya, Y., Kodai, S., Taisei, K., et al. Symbolic analysis of indicator time series by quantitative sequence alignment [J]. Computational Statistics and Data Analysis, 2008, 53(2): 486－495.

[12] Juan, G. B., David, M. G., Wiston, A. R.. symbolic hierarchical analysis in currency markets: an application to contagion in currency crises [J]. Expert Systems with Applications, 2009, 36(4):7721－7728.

[13] Juan, G. B., Wiston, A. R.. Multidimensional minimal spanning tree: The dow Johns case [J]. Physica A, 2008, 387(21):5205－5210.

[14] Brida, J. G., Punzo, L. F.. Symbolic time series analysis and dynamic regimes [J]. Structural Change and Economic Dynamics, 2003, 14(2): 159－183.

[15] 张杰,陈晔君. 基于符号时间序列分析法的 A 股上海板块网络结构分析[J]. 科学技术与工程, 2010,10(5): 1184－1187.

[16] Wei, Y., Wang, P.. Forecasting volatility of SSEC in Chinese stock market using multifractal analysis [J]. Physica A: Statistical Mechanics and its Applications, 2008, 387(7): 1585－1592.

[17] Atin, D.,Pritha, D.. Chaotic analysis of the foreign exchange rates [J]. Applied Mathematics and Computation, 2007, 185(1): 388－396.

Analysis and Forecasting of Financial Returns Based on Symbolic Time Series Method

XU Mei　HUANG Chao

(School of Management and Economics, Tianjin University, Tianjin 300072, China)

Abstract: Symbolic time series analysis method is introduced into the analysis of the characteristics of return change from the angle of large scale. The method of determining the principal return change patterns and forecasting return levels is proposed. Firstly, the return series is transformed into symbolic series. Different words in symbolic series represent different return change pattern. The principal return change patterns can be determined according to symbolic series histogram. Then, on condition that the return levels of the former several time points are determined, according to the probability distribution of various return change patterns, the probability of different return levels of the next one or several time points can be deduced to realize the forecasting of return levels. The return series of six indexes which are Shanghai composite stock, Shenzhen component stock, Shanghai industrial stock, Shanghai commercial stock, Shanghai property stock and Shanghai utility stock are analyzed. The principal return change patterns of each index are determined and the return levels based on the principal return change patterns are forecasted to prove the effectiveness and feasibility of the method.

Key words: symbolic time series analysis; histogram; return; principal pattern; forecasting