一种面向金融时间序列的趋势特征挖掘算法研究

■ 谭　华　嘉兴学院经济学院

浙江省教育厅项目(Y200805517)

［摘　要］时序数据相似性挖掘是数据挖掘中的重要研究内容。本文根据金融事件序列自身特点，将股票中的时间序列转换为以价格变动率为变量的时间序列，对趋势特征提取、聚类算法进行改进，并给出新的相似度量标准，将时间序列的预测问题转化为频繁和有效特征集的发现问题，进而进行挖掘预测。实验结果表明，该方法能有效预测时间序列中的事件。

［关键词］金融时间序列　数据挖掘　相似性　趋势特征　聚类

一、引言

随着数据挖掘技术的发展，在时间序列中进行数据挖掘的研究也逐渐引起了许多学者的兴趣，其中一个研究热点就是从时间序列中发现相似的序列模式。将时间序列相似性研究应用于股票的预测，可以从历史数据中寻找与当前的股票相似的模式，因为人们相信历史会重现，所以可以用相似模式的历史数据来预测当前股票在未来的走势。

本文将股票中的时间序列转换为以价格变动率为变量的时间序列进行分析，并对趋势特征提取、聚类算法进行改进，将时间序列的预测问题转化为频繁和有效特征集的发现问题，进而进行挖掘预测，根据连续一段时间的涨跌情况判断市场趋势，以求能准确把握市场趋势，获取更大利润。

二、时间序列数据相似性模式挖掘

时序数据相似性模式挖掘的研究已有一些研究成果，对于时间序列相似性的研究主要集中在以下3个方面：(1)时间序列由时域转换到频域后研究。这种方法将时间序列从时域通过傅立叶变换或小波变换映射到频域，使用一个固定长度的滑动窗口在序列中移动，将窗口内的数据经过变换后，采用各种频率来代替原始数据；(2)在时域内研究。这种方法直接在时域内处理数据，主要技术包括数据平移、按比例调节数据幅值、平滑处理和时间弯曲等；(3)定性计算相似性。为了消除前面两种方法的缺点，人们提出了定性计算相似性的方法，这种方法是在时域内进行研究，但并不是逐点进行相似性计算，而是只考虑一些有意义的点，如平均值、峰值、斜率或趋势值等，这样将大大减少计算量。

三、趋势特征挖掘方法

常见的金融时间序列数据主要包括股票、期货、外汇、债券等金融产品的市场交易记录，记载这些交易的时间序列数据反映的是一个有众人参与的市场环境下相应交易品种的价格变动情况，市场参与者更关心自己的投入是赚还是赔以及赚和赔的程度有多大，具体商品价格是次要的，如果投入的本钱经过市场上一番交易之后能够增值，投资者的目的就达到了。投资者要的不是具体的商品，而是能从市场上得到比投入本钱更多的回报，至于投资品种、产品单价是多少并不重要。本文在对金融时间序列数据的分析中，以价格变动率$(x_i-x_{i-1})/x_{i-1}$作为研究切入点，正为涨，负为跌，而涨和跌是性质截然相反的市场走向。

由于股票时间序列含有很多噪声，两个极值点之间往往离的很近，有时只有2个时间单位，因此在进行特征提取前必须进行平滑处理，去除噪声，然后寻找转折点来对时间序列分段线性化。平滑处理技术很多，本文采用采用最简单的有限脉冲响应法(FIR)，具体算法如下：

给定时间序列$a_{raw}(n)$，则平滑过程为式（1）：

$$\hat{a}(n)=\sum_{i=0}^{N-1}a_{raw}(n-i+[N/2])\cdot c(i) \quad (1)$$

其中$a_{raw}(n)$是原始数据，$\hat{a}(n)$是清洗后的数据，$c(i)$是含N维系数的向量，N根据具体数据来定，$c(i)$是设计FIR的重点，由脉宽和精度来确定，可用Matlab信号处理工具箱中有关函数得到。

1. 趋势特征抽取算法

时间模式挖掘是在空间中寻找能表征和预测事件的区域，如果预测点之前的时间模式包含在这些区域当中，则预测该事件点的发生提供了一种决策方法。时间序列数据的特征提取是模式发现的前提条件。分段线性法是目前应用最为广泛的时间序列特征提取方法之一，该方法具有较高的滤除噪声和数据抽象能力，可以根据需要获得时间序列数据不同精度的抽象表示。由于以近似误差为目标函数将会使某些显著的趋势在拟合的过程中失去其原有的特征，因此本文提出相应趋势特征抽取算法（TFPA）。该算法从时间序列数据中提取显著特征，能更好地保留原始时间序列中的数据变化趋势。TFPA算法也是用直线段近似表示时间序列，每一直线段通常代表一种趋势特征。如果一个趋势特征是显著的，那么说明该趋势斜率很大，时间序列数据值发生了显著的变化，呈现出明显的趋势特征；或者该趋势持续时间长，具有一定的代表意义。本文为时序数据中的显著趋势特征作如下定义：

定义1 设l_i为时间序列数据分段线性表示的第i段直线段的斜率，$\triangle t_i$为该趋势线段持续的时间，如果$|l_i|$大于给定的阈值δ_l或者$\triangle t_i$大于给定的闽值δ_t，则认为该段特征是显著的。

算法1(TFPA)：

Input：T(1：n)；δ_l；δ_1：斜率差异阈值δ
Output：T(1：n)的趋势特征序列Seg_TS
Seg_TS=ϕ;
for(i=1;i<n;i++)
Create_Seg[T(i：i+1)];/ 生成初始分段/
for(i=l；i<n−1；i++){
if(($l_{i+1}-l_i<\delta$)& &($|l|<\delta_l$)& &($\triangle$ti $<\delta_t$))
Seg[i+1]=Merge(Seg[i]，Seg[i+11)；/合并非显著相邻分段/
else
Seg_TS=Seg_TS+Seg[i]；
}
return Seg_ TS；

其中T(i：j)为时间序列T(1：n)的位于时间t_i与t_j之间的子序列，δ为相关阈值，可以通过改变δ的大小控制TFPA算法对时间序列数据的近似精度。一段时间序列数据经过处理后会生成相应的特征序列F={f_i=(l_i，$t_{i,s}$，$t_{i,e}$)|1≤i≤M_i，M_i为特征数}，其中l_i为该特征的斜率，$t_{i,s}$，$t_{i,e}$分别为该特征的起止时间，且$t_{i,s}=t_{i+1,e}$。对所有的时间序列数据进行相同的处理将生成相应数量的特征序列数据(不同的时间序列产生的特征序列可能存在不等的特征数M_i)。

2. 趋势特征聚类

在生成特征序列的过程中将会得到大量的趋势特征，这些特征在预测过程中很难有效的处理，需对提取出的特征分组。本文采用基于划分的聚类方法对特征聚类，选取不同的k值进行实验，根据

预测精度确定合理的聚类数量。聚类的一个重要问题是计算聚类对象之间的相似度。一般欧氏距离对坐标值的变化以及坐标的偏移十分敏感，经常不能正确度量数据对象之间的相似性。本文采用带权值的欧氏距离作为相似度量标准，选取的变量为特征的斜率l_i和持续时间$\triangle t_i$（$\triangle t_i = t_{i,e} - t_{i,s}$），每个变量均可被赋予一个权值，以表示其所代表属性的重要性。

$$d(i,j)=\sqrt{\alpha\left(\frac{l_i-l_j}{2\sqrt{l_i^2+l_j^2}}\right)^2+(1-\alpha)\left(\frac{\Delta t_i-\Delta t_j}{2\sqrt{\Delta t_i^2+\Delta t_j^2}}\right)^2} \quad (2)$$

式(2)中，每一特征都除以相比较的两个特征的度量值平方和，可以消除比较基准不同所带来的影响。$\alpha(0\le\alpha\le1)$是斜率的权重，称之为聚类系数，$(1-\alpha)$是特征持续时间的权重。可以通过采用不同的聚类系数值决定二者在聚类过程中的相对重要程度。选择了相似度量标准后，可得趋势特征的聚类算法如算法2。

算法2

Input: 特征值$D=\{f_1,f_2,...,f_n\}$，聚类数量k

Output：聚类结果集$C=\{C_1,C_2,...,C_k\}$

(1)从D任意选择k个对象作为初始聚类中心$f^1,f^2,...,f^k$

(2)将数据对象$f_i(i=1,2,...,n)$赋予类$C_j, j\in\{1,2,...,k\}$当且仅当

$$d(f_i,f^p)\le d(f_i,f^j), p=1,2,...,k且j\ne p$$

(3)对类$C_j, j\in\{1,2,...,k\}$，计算新的聚类中心$f_*^j(f_*^j\in C_j)$，

$s.t.$对于$\forall f_p\in C_j$有

$$\sum_{f_i\in C_j, f_i\ne f_*^j} d(f_i,f_*^j)\le \sum_{f_i\in C_j, f_i\ne f_p} d(f_i,f_p)$$

(4)如果对于$\forall i=1,2,...,k, f_*^i=f_i$成立，结束，否则转（2）

经过聚类处理之后，时间序列数据转变为由若干类基本特征构成的特征序列。

3. 特征模式预测发现方法

对金融时间序列的趋势特征提取的目的是希望能在时段Y内发现能够预测出事件的特征模式。假定趋势特征集f为按时间偏序排列的某特征序列的子集，如对所有特征序列，f在时段Y内频繁出现，而Y之外出现的几率很小，则可以用该特征子集预测事件的发生。

对于所有特征序列，如果在目标事件发生前的时段Y内，s%的序列都包含特征集f，则f的支持度为s%。如果s%大于给定的阈值，则f是频繁的。定义了特征集f的支持度后就可以应用Apriori算法从经过变换得到的特征序列中发现频繁特征集，及其相应的支持度。如果在时间序列数据集中，c%的特征集f都出现在Y时间段内，则称f的置信度为c%。如果c%大干给定的阈值，则称f是有效的。

若频繁特征集f在预测时段Y内发生的次数为x_1，在Y时间段之外发生的次数x_2，则该特征集的置信度为$x_1/(x_1+x_2)$。如果置信度大于c%的某一特征集在目标事件发生之前的Y时间段内出现，则认为该特征集与目标事件是相关的，可以用来预测事件的发生。若某频繁和有效特征在预测时段出现，则目标事件发生的概率为该特征集的置信度。

四、应用研究

为验证提出方法的有效性，本文以上证A股中牧股份（600195）为例，选取2009年3月2日至2010年2月26日共240个工作日的时间序列数据进行实验研究，预测股价突变（突然上升、突然下降）事件。

沪深证券交易所于1996年12月16日之后设立价格涨跌幅度限制，涨跌幅限制在10%以内。据此本文首先将股票时间序列转换为（价格变动率×100）的时间序列，称其为股价变动时间序列，这样既能反映投资者感兴趣的参数，又能将价格曲线放在某一区间内。文中将曲线区间设置为[-15,15]，当|股价变动|>15时，都将其用±15代替。中牧股份（600195）股价时间序列转换得到的股价变动时间序列如图1所示。

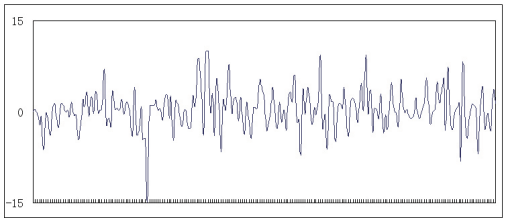

图1 中牧股份股价变动时间序列

本文将时序中2/3的数据用于生成预测模式，1/3的数据用于预测检验。取聚类系数 为0.8，支持度阈值为30%，置信度阈值为90%。应用本文所提出的方法，当Y=7天、k=30时，能够正确预测出77%的股价突变事件；预测事件为股价突变而实际没有突变的比例为4.6%，预测精度表示为正确预测的百分比。在Y=7天的条件下，对不同的聚类数量K值进行实验，实验结果如图2所示。

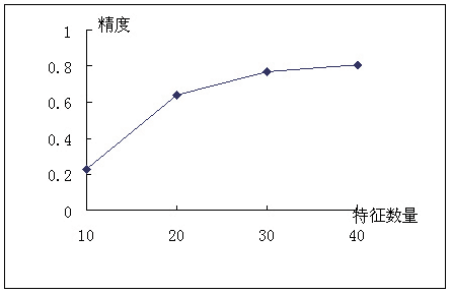

图2 特征数量变化曲线

当特征的聚类数量小于20时，随着聚类数量的增加预测精度增长较快，而当聚类数量大于20时，随着聚类数量的增加预测精度增长缓慢，并趋于平稳。在k=30时，对不同的Y值进行实验得到结果如图3所示。

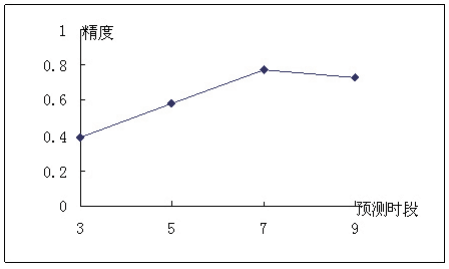

图3 预测时段与预测精度曲线

随着Y的增加，开始时预测精度逐渐提高，但当Y值增大到一定程度后预测精度反而下降。这是由于较小的Y值在预测时段内包含的特征较少，不足以产生可靠的预测，因此随着Y的增大，预测精度逐渐提高。当Y很大时，所有可以用于预测的趋势特征都已经包含到预测时段内了，而多余的数据反而对预测造成了噪声和污染，从而导致预测精度下降。

五、结论

本文根据金融时间序列自身特征，将股价时序转换为股价变动时序进行分析，对趋势特征提取、聚类算法进行改进，将时间序列的预测问题转化为频繁和有效特征集的发现问题，进而进行挖掘预测。实验结果表明，该方法能够有效地对股价时序数据中的股价突变进行有效预测。而要得到更精确的预测值，则需要对特征数量、预测时段的长度作进一步的调整。在后续研究中，不仅可以对方法本身进行改进，同时在应用上可以将其扩展到金融事件序列的其他方面，如汇率等。