文章编号:1007 - 6735(2006)04 - 0381 - 05

基于数据挖掘的金融时序频繁模式的快速发现

胡晓青, 王 波
(上海理工大学 管理学院, 上海 200093)

摘要: 针对金融时间序列分析中注重快速作出趋势判断的特点,利用数据挖掘的思想和工具,提出一种金融时间序列模式快速发现算法. 与传统的预测算法相比较,该算法对数据的分布和平稳性等方面的要求不高,不基于任何假设,能够非常快速地发现时间序列中的频繁模式,经过模式匹配后,可以用于金融时间序列的分析与预测. 以实际汇率数据为例,证明了该算法的有效性.

关键词: 数据挖掘; 时间序列; 频繁模式; 关联规则; 金融
中图分类号: TP 301.6; TP 18 **文献标识码**: A

Fast discovering frequent patterns in financial time series data based on data mining

HU Xiao-qing, WANG Bo
(*College of Management, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: The financial time series analysis pays attention to the fast prediction of trend. An algorithm for fast discovering frequent pattern in financial time series data is proposed based on data mining. Comparing with traditional methods, the algorithm requires no special demand on the distribution and stability of raw data, is not based on any hypothesis, and can fast discover the frequent patterns from time series data. The discovered frequent patterns can be used for prediction of financial trend by matching patterns. In addition, the experiments on exchange rate data demonstrate the utility of this approach.

Key words: *data mining; time series; frequent pattern; association rule; finance*

金融时间序列是经济与金融领域中最重要的数据类型,目前一般的预测方法着重于从全局上构造线性或非线性模型. 如自回归滑动平均(autoregressive moving average, ARMA)模型,它要求时间序列是平稳的,或者通过差分后是平稳的,并要求 ARMA 模型所产生的时间序列与时间观测序列的误差互不相关且呈正态分布[1]. 对于大多数的实际金融系统中的时间序列,平稳性假设、互不相关性和正态性并不是很容易满足. 而且,由于金融市场是一个非常复杂的非线性系统,所能观测到的仅仅是其演化的数据,对其系统结构和参数了解很少,要建立一个合理的全局性的数学模型比较困难,因此,用传统方法建立的模型也许在局部或短期内的预测精度较高,但是全局效果并不尽如人意. 而采用数据挖掘来对金融时间序列进行分析,以数据为驱动,可不基于任何假设,由计算机自动快速地发现一些隐藏的、有

收稿日期: 2005 - 10 - 10
基金项目: 上海市重点学科建设资助项目(T0502)
作者简介: 胡晓青(1982 -),女,硕士研究生.

价值的规律.

时间序列分析的基本假设——影响过去的和现在的活动模式的因素,将以基本相同的方式对未来产生影响.本文以此为基础,以汇率为例,针对金融时间序列数据的特点,尝试用一种新的符号化方法,结合数据挖掘方法和工具,挖掘汇率变化中的频繁模式.这些模式通常能刻画出汇市的波动性,如小幅震荡上涨,急剧下跌后小幅回升,暴涨后暴跌等.且通过模式匹配,可以对其走势作出分析与预测.采用这种方法来做预测,以趋势的快速准确判断为研究重点.因为在现实中预测一种前瞻性,即趋势的把握,对市场人士的参考意义要远大于一个准确的数字.

1 相关研究与基本思路

金融时间序列数据为非平稳序列,但其发展变化有其内在的发展规律,比较常见的方法就是通过确定其变化模式来做预测.而且对金融的时间序列分析,重视趋势的变化点,对趋势的变化信息感兴趣.如股票,价格由升变为降时抛出,由降变为升时买进,因此行情的转折点是最关注的信息.

针对时间序列的数据挖掘方面的研究已经有了不少报道.文献[2]提出了一个事物型数据库中频繁项目集的启发式搜索策略——Apriori 算法;文献[3]将 Apriori 算法引入事件序列频繁模式的分析中,提出事件序列中的频繁模式发现算法;文献[4]又进一步将该策略发展到时间序列中,提出了时间序列关联规则的分析,并提出一种固定窗口分割时间序列的方法,其不足在于窗口大小不容易确定,且计算复杂;文献[5]提出基于线性化分段的符号化方法,该方法能较好地对原始数据进行滤波及平滑处理,但这种自底向上的分段方法会过滤掉一些在金融分析中的重要信号,如趋势变化的转折点——峰、谷,处理金融时间序列数据效果并不是很好.

本文通过一种适合金融时间序列数据特点的算法,来提取数据中隐含的频繁模式.基本思路,是对原始数据进行分段线性化处理,既约简了数据、消除噪声、实现分割,又保持数据的原有结构特性和趋势变化点;进而对得到的线性分段进行聚类分析,将具有共同特征的、相类似的分段分成一类,以类标志符来表示,得到离散的符号序列;再对这个符号序列进行关联分析,挖掘出其中的频繁模式;最后通过模式匹配,来实现走势的分析与预测.

2 基于数据挖掘的金融时序频繁模式的快速发现算法

2.1 分段线性化

分段线性化是一种经典的时间序列表示方法,目的是为了实现数据的抽象、约简、除噪声和形态的分割.其一个很重要的优点是具有很好的形态表达和分割能力.经过处理,时间序列及变化形态被很自然地分割成形态各异的线性分段,每一分段都简洁、直观地表达了时间序列在该时间段内的变化特征,并且不同分段在形态上相对独立.因此,可以认为每个分段都能够代表一个相对独立的变化模式.

本文采用"自顶向下"的分段方法,用每个分段拟合直线的最大误差来控制分段数目.方法描述如图1所示.将时间序列 S 分成两个线性部分:上升部分 A_1 和下降部分 A_2.用 A_1 和 A_2 来表示 S,近似误差会比较大.为了进一步得到误差更小的分段,继续把 A_1 分成 B_{11}, B_{12};把 A_2 分成 B_{21}, B_{22}得到 S 更精确的表示.继续这个过程,一直到所有的线性分段的最大偏差都小于给定偏差,最终得到 S 的分段线性化表示.

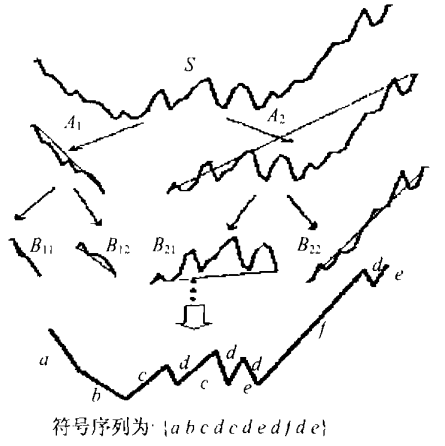

图1 时间序列的分段线性化及符号化

Fig. 1 Piecewise linear representation and symbolism of time series data

2.2 聚类分析

2.2.1 数据标准化

经过线性化分段后,将原始时间序列转变为由很多个线性分段近似表示.如果给每个线性分段赋予一个类标志符,则随着时间序列长度的增加,类标志数量会急剧增加,对以后的挖掘造成很大的负担.

在这些线性分段中,许多具有相同的特征,细微的差别,可以通过聚类分析,以线性分段的斜率和时间长度为属性特征值,将具有相同变化模式的线性分段归为同一类,并赋以相同的类标志符(如图 1).

但是,由于空间坐标和属性特征之间、各个属性特征之间的取值单位不同,具体数值可能相差悬殊. 这样,绝对值较大的属性特征其影响可能会湮没绝对值较小的属性特征,使后者应有的作用得不到反映. 为消除取值单位的影响,确保各属性在分析中的地位相同,必须对坐标值和属性特征进行标准化处理. 其公式为

$$u_i = \frac{v_i - v_{\min}}{v_{\max} - v_{\min}} \times 100$$

式中　u_i ——原坐标植

v_i ——标准化后坐标值

标准化后, u_i 介于 0～100 之间,各属性特征之间、属性特征和空间坐标之间具有横向的可比性.

2.2.2　聚类分析

线段的聚类比较复杂,但是将线性分段投影到向量空间,所有线段都从原点出发,以线段长度为向量的模,线段的方向为向量的方向,可以把线段聚类问题简化为平面点聚类问题. 通过聚类分析,也可以清理一些噪声数据,排除一些孤立点对整体数据的影响. 聚类后将离散的时间序列表示为符号序列.

本文采用逐步聚类的 B_k-均值算法, B_k 代表该算法在数据集上分割并计算后输出的聚类的数量. 算法在初始化时,随机地给定 B_k 个初始分类,然后通过不断地迭代改变分类,使每一次改进之后的分类方案都较前一次好,直到类内成员不再变化. 而所谓好的标准,是同一类内的记录越近越好,而类之间尽可能地区分. 类中心为类内所有成员的质心点.

该算法对存在孤立点的情况下鲁棒性较好,而且聚类结果与输入数据顺序无关,在数据平移与正交变换后不会影响聚类结果. 算法的难点在于 k 值的选择,即聚类个数的确定[6],通常可以通过选择若干个 k 值实验来确定.

2.3　频繁模式的发现

一个项目集是一个非空的项目集合. 一个序列模式(sequence)是有序的若干项目集组成的队列. 不失一般性,将项目集映射到一个连续的整数集. 将一个项目集 i 定义为 $(i_1 i_2 \cdots i_m)$,其中 i_j 是一个项目. 定义一个序列 S 为 $\langle s_1 s_2 \cdots s_n \rangle$,其中 s_j 是一个项目集. 序列的长度是序列中的项目集的个数,一个长度为 k 的序列称为 k 序列.

一个序列模式的支持度,定义为支持该序列模式的项目集数与总项目集数之比,即出现该模式的概率. 一个序列模式的置信度,定义为支持该 m 序列模式的项目集 $(i_1 i_2 \cdots i_m)$ 数和支持 $m-1$ 序列模式的项目集 $(i_1 i_2 \cdots i_{m-1})$ 数之比,即出现 $m-1$ 序列模式后出现 m 序列模式的概率. 称满足最小支持度限制的一个序列是一个大序列,若大序列满足置信要求,则判为频繁模式.

2.3.1　频繁模式发掘算法描述

AprioriAll 算法是一种基于 Apriori 的改进算法,它的一般结构都要遍历数据多次,在每一遍中都利用一个大序列的种子集合作为开始,利用种子来产生新的潜在大序列,称为候选序列(candidate sequence). 在遍历数据期间,需发现这些候选序列的支持度,在遍历结束时,要确定哪些候选序列确实是大的,这些大的候选者作为下一遍历的种子.

定义 L_k 为所有大 k 序列的集合,而 B_k 定义为候选 k 序列的集合.

算法描述为:

L_1 = (大 1 序列);
for $(k = 2; L_{k-1} \neq \emptyset; k++)$ do
　B_k = 从 L_{k-1} 中产生新的候选者
　for 在数据库中每一个序列模式 c do
　　包含于 c 中的 B_k 内的所有候选者计数增 1
　　$L_k = B_k$ 中满足最小支持度的候选者
　　输出 L_k 中满足置信度要求的序列模式
　end
end

2.3.2　Apriori 候选者的产生

该算法中,候选序列由前一遍的大序列产生,然后再遍历数据库测试其支持度. 本文对这一部分进行了一些改进,以提高候选者产生的效率.

首先,进行 L_{k-1} 与 L_2 的连接运算,替代原算法中的 L_{k-1} 与 L_{k-1} 的连接. 改进后 L_{k-1} 与 L_2 连接算法为:

insert into B_k
select $p.\text{litemset}_1, p.\text{litemset}_2, \cdots, p.\text{litemset}_{k-1}$,
　　$q.\text{litemset}_2$
from $L_{k-1}\ p, L_2, q$
where $p.\text{litemset}_{k-2} = q.\text{litemset}_1$

改进后,少了在遍历 L_{k-1} q 序列和遍历 q 序列中与 p 序列中相同项目的过程,只需遍历 L_2 序列. 而且可以将 L_2 用二维表格表示,按下标来直接查

询,省略遍历 L_2 序列的过程,进一步优化算法. 下一步,删除 B_k 中不在 L_{k-1} 里的 $(k-1)$ 子序列,得到在一个 L_k 中序列的超集. 这与删除 B_k 中不满足支持度要求的子序列是等价的.

通过上述步骤,可以发掘出满足支持度和置信度要求的子序列,即频繁模式. 这些频繁模式可以用来建立模式库.

2.4 模式的匹配与预测

通过上述步骤,可以从历史数据中得到大量有用的序列模式,建立模式库. 将近些年的测试数据先通过分段线性化和聚类处理,形成符号序列.

2.4.1 模式的匹配算法

设符号序列 W 由线性分段 $E_1\ E_2\ \cdots E_m$ 组成,$L_{(u,v)}$ 表示序列模式库,则匹配算法为

$W = \{E_1, E_2, \cdots, E_m\}$

$L_{(u,v)} = \{$序列模式库$\}$

$B_k = \{$第 k 个序列模式的置信度$, k = 1, \cdots, u\}$

$S_k = \{$第 k 个序列模式的支持度$, k = 1, \cdots, u\}$

for ($i = 1; i <=$ 模式库中模式的数目 $u; i++$)

/ *搜索模式库 */

if 从 E_m 到 $E_{(m-v+2)}$ 和 $L_{(i,v-1)}$ 到 $L_{(i,1)}$ 完全匹配

匹配成功,预测 $E_{m+1} = L(i, v)$,置信度为 c_i,支持度为 s_i;

end

end

2.4.2 预测分析

在实际预测过程中,所取的测试数据的起止点不一定在趋势的转折点上,因此一般是取最近的一个转折点来作为止点,倒数第二个线性分段为符号序列的最后一个分段,用来保证模式的完整性. 根据这种方法做预测,结果并不一定是惟一的,可以结合置信度、支持度的高低来对决策者提供支持.

3 实例分析

选取 1989 年 7 月到 2005 年 3 月间欧元/美元的汇率收盘价,共 4 148 组数据,以 2002 年前 3 305 组数据为对象挖掘频繁序列模式,建立模式库;以其后 843 组数据作为测试数据. 图 2 为分段线性化后部分效果图,其中最大允许偏差为 0.01. 3 305 组原始数据约简为 1 049 个线性分段表示.

然后对这些线性分段以 B_k-均值算法进行聚类,共分成 7 类,聚类中心由表 1 所示. 不同的线段类型代表不同的发展趋势,如快速上涨,缓慢下跌等.

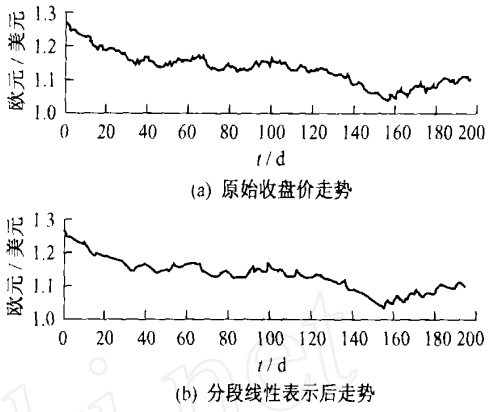

图 2 欧元/美元汇率时间序列分段线性表示

Fig. 2 Piecewise linear representation of daily price of EURO/USD exchange rate

表 1 B_k-均值聚类产生的聚类中心

Tab. 1 Clustering centers of B_k- means cluster

类别	时间长度/d	汇率变化
1	1.713 2	0.013 209
2	3.872 9	-0.024 59
3	4.375	0.024 854
4	17.267	-0.022 95
5	9.021 7	0.027 459
6	9.023 3	-0.026 3
7	1.490 9	-0.007 19

对符号序列进行频繁模式挖掘,设定最小支持度为 3,最小置信度为 50 %,得到几十种频繁模式. 表 2 所示为部分频繁模式示例.

表 2 频繁模式示例

Tab. 2 Some examples of frequent patterns

模式	C/ %	S(频度)
(2,1,7)	57	50
(7,5,7)	56	15
(1,4,1,7)	100	3
(1,7,3,7)	59	20
(6,3,2,1)	67	4
(2,1,7,3,7,1)	70	7
(1,2,1,7,1,7,1,7)	100	3
(2,1,2,1,2,1,7,3,7)	75	3
…	…	…

图 3 为符号模式对应的图形模式,这些模式反映了汇市波动的特征. 以 2002 年 1 月到 2005 年 3 月的欧元/美元的日收盘价为测试数据,在对其进行模式匹配后可以发现:

a. 如果以置信度大于 0,支持度大于等于 1 来建立模式库,即只要出现过的模式都记录在模式库中,则测试数据中出现的模式 95 %以上都能在序列模式库中得到匹配. 这充分说明了,在汇率市场时间序列中绝大多数模式,都曾经在历史上出现过,是历史模式的再现,也印证了在时间序列分析中的基本假设 ——影响过去的和现在的活动模式的因素,将以基本相同的方式对未来产生影响.

b. 如果以置信度大于 50 %,支持度大于等于 3 来建立模式库,并结合预测结果中置信度和支持度的高低,预测的准确率最高可以达到 67 %. 以图 3 和表 2 的结果为例,如果当前的趋势变化中出现了 (2,1,7,3,7) 的模式,则可预测下一个变化趋势有 70 %的概率为符号 1 代表的线段模式.

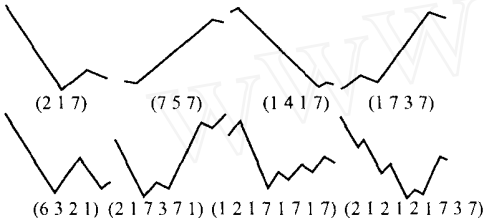

图 3　表 2 频繁模式代表的走势变化图形示意

Fig. 3　Graphic example of frequent patterns in Tab. 2

由于影响汇率的因素众多,各因素之间又彼此相互影响,汇价的预测非常困难,一般的汇率预测算法准确率在 50 %左右. 相对而言,这也证明了此算法的有效性.

4　结束语

根据金融时间序列数据的特点,结合数据挖掘方法,提出了一个基于聚类分析和关联分析的金融时间序列模式发现算法. 结果表明,该算法由于处理时间几乎与序列长度成线性关系,具有良好的伸展性,复杂度也很小. 通过调整不同的参数(如线性分段最大允许偏差,聚类的数目,频繁模式的最小支持度和最小置信度等),可以实现不同要求的频繁模式的快速挖掘. 因此算法的难点也在于这些参数的选择上,通常情况下,可以针对用户的问题域,选择若干个不同的参数进行实验. 同时,该算法还可以推广到其他时间序列的模式中.

参考文献:

[1] 何书元. 应用时间序列分析[M]. 北京:北京大学出版社,2003:148 - 276.

[2] AGRAWAL R, RAMAKRISHNAN S. Fast algorithms for mining association rules in large databases[A]. Proceedings of the Twentieth International Conference on Very Large Databases[C], Santiago: ACM Press, 1994: 487 - 499.

[3] MANILA H, TOIVONEN H, VERKAMO A I. Discovery frequent episodes in sequences [A]. Proc of KDD - 95[C]. Montreal: AAAI Press, 1995: 210 - 215.

[4] DAS G, LIN K, MANNILA H, et al. Rule discovery from time series[A]. Proceedings of Fourth Annual Conference on Knowledge Discovery and Data Mining[C]. New York: AAAI Press, 1998: 16 - 22.

[5] 李斌,谭立湘,章劲松,等. 面向数据挖掘的时间序列符号化方法研究[J]. 电路与系统学报,2000:9 - 14.

[6] 邵峰晶,于忠清. 数据挖掘原理与算法[M]. 北京:中国水利水电出版社,2003:197 - 224.

(下期发表论文摘要预告)

具阶段结构和 Holling II类功能反应的捕食系统研究

徐长永,　王美娟,　周艳丽

(上海理工大学 理学院, 上海　200093)

摘要: 讨论了一类捕食者具 Holling II类功能反应,捕食者和食饵均具阶段结构的非自治捕食者-食饵系统,并且成年捕食者只对成年食饵捕食. 运用 Liapunov 函数方法,得到了该系统一致持续生存的充分条件. 讨论了周期系统存在惟一、全局渐近稳定周期解的条件. 对更具普遍意义的概周期现象,得出了概周期正解惟一存在且全局渐近稳定的充分条件.