金融时间序列挖掘综合模型①

Integrated Model of Financial Time Series Mining

朱　冲[1]　朱贤贵[2]　张向利[1]　(1.桂林电子科技大学 信息与通信学院 广西 桂林 541004;

2.中国人民银行娄底市中心支行 湖南 娄底 417000)

摘　要： 时间序列挖掘是数据挖掘的重要组成部分，本文通过对金融数据按地点划分，经过平滑、聚类处理，再对同一类别的各条金融序列分别发现其序列内频繁模式，综合一个得到同类别多条金融时间序列的复合挖掘模型。农业价格时序挖掘实践证明,该金融时间序列挖掘模型利用挖掘出来的知识对金融时间序列趋势进行了定性分析，能有效地指导用户的市场行为，辅助用户决策。

关键词： 金融时间序列 频繁模式 数据挖掘 模式发现

在金融市场中,信息连续地影响着市场价格变化，快速作出价格趋势判断是金融序列趋势预测的重点和难点，在实际应用中，经常需要发现不同金融时间序列间可能存在的关联关系，这种关系对于人们更彻底的认识各个金融时间序列的相互影响并据此做出合理的决策具有重要的参考价值。

目前,国内外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[1]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[2]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[3]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易确定,且计算复杂；文献[4]针对金融时间序列分析中注重快速作出趋势判断的特点,提出一种金融时间序列模式快速发现算法。

本文针对金融时间序列数据库信息,设计实现金融时序模式挖掘系统，该系统能找出价格随空间、时间分布的规律，有效地指导用户的市场行为，辅助用户决策。

1　金融时间序列挖掘分析

1.1 金融时间序列特点分析

非平稳性是金融时间序列其中最为显著的特征之一。随机过程的平稳性是指其统计特性不随时间而变化，一般是指广义平稳性，即随机过程的期望值和协方差与时间起点无关。但由于影响市场的政治、经济、文化环境等随时间的变迁，时间序列的一阶矩，协方差不可能维持不变，因而通常表现为明显的非平稳性。

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，具体而言，农产品价格时间序列的波动性一般比较小，但因为政策或特殊气候如暴雪、连续干旱等,也表现出随机波动特性，这种波动可以看作噪声信号。数据挖掘的目的是发现序列中隐含的一些本质规律，噪声的存在一方面会淡化规则，即降低规则的显著性，另一方面又可能提供一些“假规则”，从而严重影响挖掘的效果。因此在时间序列进行挖掘之前，有必要先对其进行预处理，尽可能消去一些随机噪声。

1.2 模型总体结构

金融时序挖掘的目标是从各个地区某段时间找出相同的或相似的变化模式，然而在没有先验知识的情况下，直接从庞大的产品价格数据库中发现相同或相似的变化模式是非常困难的，为了进行深度挖掘，必须先对数据分类，按市场、省份划分成一系列的时间序列，对这些时间序列进行平滑处理，再进行聚类分

① 基金项目:广西自然科学基金项目(桂科自 0832246);广西青年科学基金项目(桂科青 0832084);广西研究生教育创新计划资助项目(2008105950810M420)
收稿时间:2008-08-06

模式，时序挖掘复合模型结构如图 1 所示：

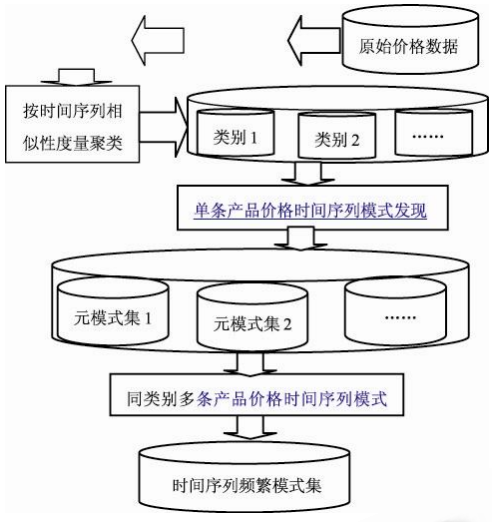

图 1　时序挖掘复合模型

处理步骤如下：

Step1.价格原始数据划分、平滑处理；

Step2.产品价格时间序列聚类；

Step3.单条产品价格时间序列模式发现；

Step4.同类别多条产品价格时间序列模式发现。

2　时间序列的平滑处理

由于在采集到的数据中经常会含有许多干扰数据，在进行静态模式挖掘之前必须对原始数据进行平滑处理，平滑数据方法有移动平均法和低通滤波器法等[5]，移动平均的计算方法很简单，以 m 阶移动平均为例来说明计算方法如下:

假设时间序列：s={ $A_{t1}, A_{t2}, A_{t3}, \cdots A_{tn}$ } ($t1 < t2 < t3 \cdots < tn$)，A_{ti} 是序列在第 ti 时刻的值，m 阶移动平均的计算方法是使用一个宽度为 m 的时间窗口从序列的起始点开始向结束点逐位移动，每移动一个时刻求一次窗口内 m 个值的平均值，即

$$\frac{A_{t1}+A_{t2}+A_{t3}+\cdots+A_{tm}}{m}, \frac{A_{t2}+A_{t3}+\cdots+A_{tm+1}}{m}, \quad (1)$$
$$\frac{A_{t3}+A_{t4}+\cdots+A_{tm+2}}{m}, \cdots$$

最后将这些计算结果按顺序排列，即可得到 m 阶移动平均序列 s'，一共有“tn -m+l”个值，可以用移动平均线序列来代替原始的时间序列 s'。移动平均可以降低数据集中的变化总量，因此用移动平均替代原时间序列可以减少不希望出现的波动。移动平均线比原始数据有一定的滞后，m 值越大滞后越多，同时 m 越大曲线越平滑越能反应时间序列的长期趋势。若选择 m 值太小,移动平均会丢失时间序列中的头尾数据，生成在原始数据中不会出现的循环或其它变化趋势，并且它可能受一些极端数据的影响。对于这些问题，通过采用适当的加权移动平均，可以消除数据中的循环、季节性和非规则的模式，而只保留趋势变化。

3　列相似性度量

由于相似性度量是相似序列搜索、序列聚类、分类、频繁模式、关联模式挖掘的基础，在时间序列挖掘中具有举足轻重的地位，人们从不同角度出发研究出许多度量方法[6]。

本文采用文献[7]提出的一种改进的相似度度量算法，算法思想：两个子序列曲线尽管它们的基线或振幅不同，但若具有相似的变化趋势，这两个子序列仍是相似的。采用如下方法：首先将序列的值标准化，然后再计算标准化后的序列间的距离，可以采用欧式距离或曼哈顿距离等。标准化方法采用 $\lambda(y_{i_k}) = (y_{i_k} - Ey_i)/Dy_i$，其中 Ey_i 是序列的均值，Dy_i 是序列的标准方差。这样可以使得标准后的序列值落在-1 与 1 之间，且序列的均值和方差分别为 0 和 1。

4　模式发现

在金融时间序列预测中，需要对单条金融时间序列频繁模式自动发现，还需要对多个金融时间序列进行分析。

4.1 时间序列频繁模式自动发现

给定支持度阈值 ξ，如果序列 a 在序列数据库中的支持数不低于 ξ，则称序列 a 为频繁序列。序列 a 在序列数据库 S 中的支持数为序列数据库 S 中包含 a 的序列个数，记为 Support(a)。对于给定的序列数据库和最小支持度阈值，时间序列频繁模式发现就是要找出序列数据库中满足最小支持度阈值的频繁序列中的最大序列，每一个这样的最大序列就是一个最大频繁模式，本文采取文献[1]提出的 Apriori 算法实现时间序列频繁模式自动发现。

4.2 多条时间序列模式发现

Tim Oates 等人提出了从多个数据流中搜索关联

模式的数据挖掘算法(MSDD)[8],其中,多数据流表示为严格同步的多个符号序列。Oates 等人给出了候选模式的产生和强关联模式的启发式搜索算法,但该算法要求数据序列必须是严格同步的。文献[9]提出了一种非同步多时间序列中频繁模式的发现算法，从多个时间序列中发现频繁结构模式，对不同序列间是否同步没有限制,并且能够发现多种结构形式的频繁模式,具有更大的灵活性和较低的计算复杂度。

Agrawal[10]给出了关于频繁模式的一个重要定理,即“任何频繁模式的子模式必定也是频繁的”。由该定理可以得到一个更为实用的推论,即“可以由已知频繁模式集产生更大长度的候选频繁模式”。因此，对于多时间序列的频繁模式可以由单时间序列中的频繁模式集合中产生。

为了进行多时间序列频繁模式发现，根据文献[9]，给出以下的定义：

定义 1.序列内频繁模式称为元模式。给定序列集合 S={$S_1, S_2 \cdots S_m$}，集合 C={$P_1, P_2 \cdots P_n$}为序列集合 S 中所有“元模式”集合。P_i 为“元模式”标识。

定义 2.模式 P={$P_i, P_{i+1} \cdots P_j$}为非空集合,且顺序不能调换。其中 P_i 为元模式。模式 P 的长度为集合 P 的大小|P|,即构成 P 的“元模式”的个数。

定义 3.称模式 P 在区间[t, t']的一次实现为最短实现，如果不存在任何子区间[u, u']$\subset$[t, t'],从中可以发现 P 的一次实现。以数组 Si(P)记录模式 P 的所有最短实现。Si(P)={[t, t']|[t, t']为模式 P 的一次最短实现}。

令 C_k 表示长度为 k 的候选频繁模式集，L_k 表示长度为 k 的频繁模式集，本文采用文献[9]的算法来实现模式发现。

5　实验结果

根据前面的金融时序挖掘综合模型，本文实现了一个以农产品价格数据库为基础，综合多种技术的农产品价格时序挖掘系统，测试中输入‘合肥’、‘包菜’，日期选择从 2004 年 1 月 1 日到 2008 年 1 月 1 日，图 2 展示了综合模型的入口及模型挖掘结果。

由图 2 可知，单条序列内时序模式发现成功地实现了对合肥、阜阳、广德、全椒的模式发现，比如合肥价格变化趋势为：1.06->1.05，多条时间序列模式发现则找出了不同地方多条金融时间序列之间的关系,如得出与合肥的包菜在 2004 年 1 月 1 日到 2008 年 1 月 1 日内价格最相似的有三个地区：阜阳、广德、全椒，由此得到的一个综合模型就能很好地找出产品

图 2　挖掘结果

价格随空间、时间分布的规律，有效地指导用户的市场行为，辅助用户决策。

参考文献

1 Agrawa L R, Ramakrishnan S. Fast algorithms for mining association rules in large databases. Proceedings of the Twentieth International Conference on Very Large Databases. Santiago: ACM Press, 1994:487－499.

2 Manila H, Toivonen H, Verkamo AI. Discovery frequent episodes in sequences. Proc of KDD95.1995.

3 Das G, Lin K, Mannila H, et al. Rule discovery from time series. Proceedings of Fourth Annual Conference on Knowledge Discovery and Data Mining. New York: AAAI Press, 1998: 16-22 Montreal: AAAI Press, 1995:210－215.

4 胡晓青,王波.基于数据挖掘的金融时序频繁模式的快速发现.上海理工大学学报,2006,28(4).

5 Kwok CO, Etzioni O, Weld DS. Scaling QuestionAnswering to the Web. ACM Trans. Information Systems, 2001, 19 (3):242－262.

6 史忠植.知识发现.北京:清华大学出版社,2002.

7 Whitehe SD. Auto-FAQ: An experiment in cyberspace leveraging. Proceedings of the Second International WWW Conference. volume1995:25－38.

8 王晓晔.时间序列数据挖掘中相似性和趋势预测的研究[博士学位论文].天津:天津大学, 2003.

9 黄河,黄轲,杭小树,熊范纶.时间序列中快速模式发现算法的研究.计算机工程与应用, 2003, 39(21):192－194.

10 Oyama S, Kokubo T, Ishida T. Domain-Specific Web Search with Keyword Spices. IEEE Trans. Knowledge and Data Eng, 2004, 16(1):17－27.