第 36 卷 第 3 期 计 算 机 工 程 2010 年 2 月
Vol.36 No.3 Computer Engineering February 2010
·软件技术与数据库· 文章编号：1000—3428(2010)03—0083—03 文献标识码：A 中图分类号：TP311.12

多粒度时间下的近似周期挖掘研究

姜 华[1]，孟志青[2]，周克江[1]，肖建华[1]
(1. 湖南第一师范学院信息科学与工程系，长沙 410205；2. 浙江工业大学经贸管理学院，杭州 310023)

摘 要：研究时态数据库中多粒度时间下的近似周期的挖掘问题。在多粒度时间、多粒度时间格式的基础上引入多粒度时间间隔的定义以及相关性质，构造多粒度近似周期模型，提出一个基于 SOM 聚类的多粒度近似周期的挖掘算法。利用高频股票数据 580000 宝钢 JBT1 进行实验，证明了该算法的有效性。

关键词：数据挖掘；自组织映射网络；多粒度时间；近似周期

Study of Approximate Periodicity Mining with Multi-granularity Time

JIANG Hua[1], MENG Zhi-qing[2], ZHOU Ke-jiang[1], XIAO Jian-hua[1]
(1. Dept. of Information Science and Engineering, Hunan First Normal University, Changsha 410205;
2. College of Business and Administration, Zhejiang University of Technology, Hangzhou 310023)

【Abstract】This paper discusses a mining problem of approximate periodicity with multi-granularity time in the temporal database. It introduces the concepts and properties of the multi-granularity time interval on the basis of multi-granularity time and multi-granularity time format. It constructs multi-granularity approximate periodic pattern. It proposes an mining algorithm based on self-organizing map to find multi-granularity approximate periodic pattern. Results obtained from experiments on high frequency stock market data of 580000 Bao Steel JBT1 demonstrate that the proposed algorithm is efficient.

【Key words】 data mining; Self-Organizing Map(SOM) network; multi-granularity time; approximate periodicity

1 概述

周期行为在现实世界中普遍存在，然而，现实生活中很多周期往往都不是精确的，而存在近似周期[1]。近似周期挖掘是当前时态周期挖掘比较新的研究领域。

对于单一时间粒度下的周期挖掘，近年来已经有了大量的研究[2-3]。实际上，人们往往习惯于用多粒度时间来表示时间信息，很多有规律的事件都发生在基于多粒度时间的周期上。

为此，文献[4]对多粒度时间、多粒度时间格式给出了严格的数学表示和性质证明，文献[5]研究了多粒度时间下的部分周期模式挖掘。以多粒度时间的形式构成周期，越来越多潜在的周期规律被发现。因此，把近似周期和多粒度时间结合起来考虑是一个很有意义的研究方向，更贴近现实世界。

文献[4]中提出的多粒度时间和文献[1]提出的近似周期模型并没有能力描述多粒度近似周期，因此本文提出多粒度时间间隔严格的数学表示，构造了一个多粒度近似周期模型。文献[1]提出的近似周期挖掘算法也不适合挖掘多粒度近似周期，所以本文给出了一个基于 SOM 网络自组织聚类挖掘多粒度近似周期的算法。

2 模型构造

在文献[4]中已经给出了多粒度时间格式、多粒度时间的精确的数学定义，为了讨论方便，这里不再详细叙述。本文在此基础上提出构造多粒度近似周期模型所需的多粒度时间的一些定义，并证明相关性质。

2.1 多粒度时间的一些定义和性质

定义多粒度时间之间的几种关系：

设 $(\mu_1,\mu_2,\cdots,\mu_n)$ 是多粒度时间格式，n 元组 $(x_1,x_2,\cdots,x_n)$，$(x_1',x_2',\cdots,x_n')$ 为 $(\mu_1,\mu_2,\cdots,\mu_n)$ 的多粒度时间。

定义 1(大于关系) 若满足：

$x_1=x_1',x_2',\cdots,x_i=x_i',x_{i+1}>x_{i+1}',1\leqslant i<n$

则称多粒度时间 $(x_1,x_2,\cdots,x_n)>(x_1',x_2',\cdots,x_n')$。

定义 2(等于关系) 若满足：

$x_1=x_1',x_1',\cdots,x_i=x_i',x_{i+1}',\cdots,x_n=x_n',1\leqslant i\leqslant n$

则称多粒度时间 $(x_1,x_2,\cdots,x_n)=(x_1',x_2',\cdots,x_n')$。

定义 3(小于关系) 若满足：

$x_1=x_1',x_2',\cdots,x_i=x_i',x_{i+1}<x_{i+1}',1\leqslant i<n$

则称多粒度时间 $(x_1,x_2,\cdots,x_n)<(x_1',x_2',\cdots,x_n')$。

定义 4(大于等于关系) 若满足定义 1 或定义 2，则称多粒度时间 $(x_1,x_2,\cdots,x_n)\geqslant(x_1',x_2',\cdots,x_n')$。

定义 5(小于等于关系) 若满足定义 2 或定义 3，则称多

基金项目：湖南省自然科学基金资助项目(09JJ6093)；湖南省教育厅重点课题基金资助项目(2008(288))；湖南省教育厅科技处基金资助项目(08C017)；湖南第一师范学院校级课题基金资助项目(XYS08N03)

作者简介：姜 华(1980－)，女，讲师、硕士，主研方向：数据挖掘，神经网络；孟志青，教授；周克江，副教授；肖建华，教授

收稿日期：2009-11-06 **E-mail：**jianghua_clo1@126.com

粒度时间 $(x_1,x_2,\cdots,x_n)\leqslant(x_1',x_2',\cdots,x_n')$。

定义 6(多粒度时间间隔) 设 n 元组 $(x_1,x_2,\cdots,x_n)$，$(x_1',x_2',\cdots,x_n')$ 为多粒度时间格式 $(\mu_1,\mu_2,\cdots,\mu_n)$ 的多粒度时间，其中，$(x_1,x_2,\cdots,x_n)<(x_1',x_2',\cdots,x_n')$，$Relen\ (\mu_{i-1}(t))^{\mu_i}=m_i$ $(2\leqslant i\leqslant n)$，令 $d_i=x_i'-x_i(1\leqslant i\leqslant n)$，若 $d_i<0$，则 $d_{i-1}=d_{i-1}-1$，$d_i=m_i+d_i$，记 $D=(d_1,d_2,\cdots,d_n)$，称 D 为 $(x_1,x_2,\cdots,x_n)$ 到 $(x_1',x_2',\cdots,x_n')$ 之间的多粒度时间间隔，从当前多粒度时间 $(x_1,x_2,\cdots,x_n)$ 开始紧后的多粒度时间间隔 D 处的多粒度时间为 $(x_1',x_2',\cdots,x_n')$。

性质 1 $(\mu_1,\mu_2,\cdots,\mu_n)$ 是多粒度时间格式，$Relen\ (\mu_{i-1}(t))^{\mu_i}=m_i(2\leqslant i\leqslant n)$，$n$ 元组 $(x_1,x_2,\cdots,x_n)$，$(x_1',x_2',\cdots,x_n')$ 为 $(\mu_1,\mu_2,\cdots,\mu_n)$ 的多粒度时间。若 $(d_1,d_2,\cdots,d_n)$ 为 $(x_1,x_2,\cdots,x_n)$ 到 $(x_1',x_2',\cdots,x_n')$ 之间的多粒度时间间隔，则 $(d_1,d_2,\cdots,d_n)$ 是 $(\mu_1,\mu_2,\cdots,\mu_n)$ 的多粒度时间。

证明：$(\mu_1,\mu_2,\cdots,\mu_n)$ 是多粒度时间格式，$Relen\ (\mu_{i-1}(t))^{\mu_i}=m_i(2\leqslant i\leqslant n)$，根据定义 6 可知 $d_i=x_i'-x_i\in\{0,1,\cdots,m\}$，所以 $(d_1,d_2,\cdots,d_n)$ 是 $(\mu_1,\mu_2,\cdots,\mu_n)$ 的多粒度时间。

设 $(x_1,x_2,\cdots,x_n)$，$(x_1',x_2',\cdots,x_n')$，$(d_1,d_2,\cdots,d_n)$ 是多粒度时间格式 $(\mu_1,\mu_2,\cdots,\mu_n)$ 的多粒度时间，如果 $(x_1,x_2,\cdots,x_n)<(x_1',x_2',\cdots,x_n')$，且 $(d_1,d_2,\cdots,d_n)$ 是 $(x_1,x_2,\cdots,x_n)$ 到 $(x_1',x_2',\cdots,x_n')$ 之间的多粒度时间间隔，那么定义如下运算：

定义 7 加运算(简记为＋)：

$(x_1,x_2,\cdots,x_n)+(d_1,d_2,\cdots,d_n)=(x_1',x_2',\cdots,x_n')$

定义 8 减运算(简记为-)：

$(x_1',x_2',\cdots,x_n')-(d_1,d_2,\cdots,d_n)=(x_1,x_2,\cdots,x_n)$

2.2 多粒度近似周期模型

构造一个多粒度时间下的近似周期模型：

设 $(\mu_1,\mu_2,\cdots,\mu_n)$ 是多粒度时间格式，$Relen\ (\mu_{i-1}(t))^{\mu_i}=m_i$ $(2\leqslant i\leqslant n)$，$G=(g_1,g_2,\cdots,g_n)$，$G_k=(g_{k1},g_{k2},\cdots,g_{kn})$，$D=(d_1,d_2,\cdots,d_n)$，$D_k=(d_{k1},d_{k2},\cdots,d_{kn})$ 是 $(\mu_1,\mu_2,\cdots,\mu_n)$ 的多粒度时间。

定义 9 符号 (A_i,e,G,D) 表示属性/特征 A_i 在多粒度时间 G 处发生 e 值的事件(记为 (A_i,e,G))，从当前多粒度时间 G 开始紧后的多粒度时间间隔 D 处事件 $(A_i,e,G+D)$ 重复发生。

定义 10 符号 (A_i,e,D) 表示在时间段 $[T,T']$ 中存在多粒度时间 G_k 使得在每次事件 (A_i,e,G_k) 发生后紧后的多粒度时间间隔 D 处事件 (A_i,e,G_k,D) 重复发生的事件。记

$$Sum(A_i,e,D)=\sum_{k=1}^{n}E((A_i,e,G_k,D))$$

表示 (A_i,e,D) 在时间段 $[T,T']$ 覆盖的所有多粒度时间 G_k，使得事件 (A_i,e,G_k,D) 重复发生的次数，其中，$T\leqslant G_k\leqslant T'$，若 $i<j$，则 $G_i<G_j$，n 表示在时间段 $[T,T']$ 中所覆盖的多粒度时间的个数。若事件 (A_i,e,G_k,D) 发生，记 $E(A_i,e,G_k,D)=1$；否则 $E(A_i,e,G_k,D)=0$。

定义 11 设给定波动值 $\sup[D]>\inf[D]$ ($\inf[D]$ 可以取 0)，定义一种多粒度近似周期模式的表示：

$$P=((A_i,e):[\inf(D),\sup(D)])=\{(A_i,e,D_k)\,|\,\inf(D)\leqslant D_k\leqslant\sup(D)\}$$

其中，P 表示事件 (A_i,e,D_k) 每隔多粒度时间间隔 D_k 的重复发生所有的事件构成的集合；D_k 可以在一定范围 $[\inf[D],\sup[D])]$ 内波动，称 $[\inf[D],\sup[D])]$ 为属性 A_i 出现状态 e (表示为 (A_i,e))的多粒度近似周期。$\sup[D]-\inf[D]$ 称为多粒度近似周期波动阈值或近似精度。当 $\sup[D]-\inf[D]=0$ 时，表示多粒度精确周期。

定义 12 设多粒度近似周期模式 $P=((A_i,e):[\inf(D),\sup(D)])$，模式 P 的支持度和置信度分别为

$$support((A_i,e):[\inf(D),\sup(D)])=\mathrm{Sum}((A_i,e):[\inf(D),\sup(D)])/n \quad (1)$$

$$confidence((A_i,e):[\inf(D),\sup(D)])=\frac{\mathrm{Sum}((A_i,e):[\inf(D),\sup(D)])}{\mathrm{Sum}((A_i,e))} \quad (2)$$

其中，$\mathrm{Sum}((A_i,e):[\inf(D),\sup(D)])$ 是时间段 $[T,T']$ 中满足模式 P 的多粒度时间 G 的个数。记：

$$\mathrm{Sum}((A_i,e):[\inf(D),\sup(D)])=\sum_{k=1}^{n}E(A_i,e,G_k,D_j)$$

其中，D_j：(1) $\inf(D)\leqslant D_j\leqslant\sup(D)$；(2)若存在多粒度时间 G_k，使得 $(A_i,e,G_k,D_{j1}),(A_i,e,G_k,D_{j2}),\cdots,(A_i,e,G_k,D_{js})$ 均发生，其中，$\inf(D)\leqslant D_{j1},D_{j2},\cdots,D_{js}\leqslant\sup(D)$，则取 $D_j=\min\{D_{j1},D_{j2},\cdots,D_{js}\}$。$n$ 表示在时间段 $[T,T']$ 中所覆盖的多粒度时间的个数。式(2)中 $\mathrm{Sum}((A_i,e))=\sum_{k=1}^{n}E(A_i,e,G_k)$，表示在时间段 $[T,T']$ 中所有多粒度时间 G_k 使得事件 (A_i,e,G_k) 的重复发生次数。

定义 13 模式 $P=((A_i,e):[\inf(D),\sup(D)])$，模式 $P'=((A_i,e):[\inf(D'),\sup(D')])$，若 $\inf(D)\leqslant\inf(D')$，且 $\sup(D)\geqslant\sup(D')$，称模式 P 覆盖模式 P'。

性质 2 对于时间段 $[T,T']$ 中的对象 A，设模式 $P=((A_i,e):[\inf(D),\sup(D)])$，模式 $P'=((A_i,e):[\inf(D'),\sup(D')])$，若模式 P 覆盖模式 P'，则 $support(P)\geqslant support(P')$，$confidence(P)\geqslant confidence(P')$。

证明：证明参考文献[1]。

3 多粒度近似周期模式挖掘算法

3.1 输入特征向量的获取

给定对象(如某支股票)和时间段 $[T,T']$，选定多粒度时间格式 $u=(\mu_1,\mu_2,\cdots,\mu_n)$，其中，$Relen\ (\mu_{i-1}(t))^{\mu_i}=m_i(2\leqslant i\leqslant n)$，在时间段 $[T,T']$ 中所覆盖的多粒度时间的个数为 n，对所有的属性状态进行编码。设 $G=(g_1,g_2,\cdots,g_n)$ 代表事件发生的时间，对于每个多粒度时间 G，计算 (A_i,e,D)，$D\leqslant L$ (其中，$D=(t_1,t_2,\cdots,t_n)$，L 是多粒度周期长度上限或阈值)，那么得到 SOM 网络的输入向量为 $(A_i,e_j,t_1,t_2,\cdots,t_n)$。

3.2 基于SOM聚类挖掘多粒度近似周期

挖掘算法输入为：(1)给定对象(如某支股票)，时间段 $[T,T']$ 和选多粒度时间格式 $u=(\mu_1,\mu_2,\cdots,\mu_n)$；(2)多粒度周期长度阈值 L，支持度阈值 $\underline{s}$，置信度阈值 $\underline{c}$ 和近似精度。要找到在 $[1,2,\cdots,L]$ 范围内符合 $\underline{s}$，$\underline{c}$ 和近似精度的多粒度近似周期模式 P，算法如下：

(1)数据准备，根据 3.1 节获取 SOM 网络的输入向量。

(2)初始化 SOM 网络。

(3)根据输入的各分量在聚类划分时重要性的大小，对输入向量各分量分配不同的权重值，即将输入向量 $(A_i,e_j,t_1,t_2,\cdots,t_n)$ 表示成 $(\alpha A_i,\beta e_j,\gamma t_1,\gamma t_2,\cdots,\gamma t_n)$，其中，$\alpha$，$\beta$，

γ 为权重值($\alpha >> \beta >> \gamma$)，得到输入向量:

$$\boldsymbol{x}(t) = (x_1(t), x_2(t), \cdots, x_{n+2}(t))$$

(4)按如下方式计算距离:

$$d_j = \sum_{i=1}^{2}(w_{ij}(t) - x_i(t))^2 + (\sum_{i=3}^{n+2}\frac{1}{1\times m_2 \times \cdots \times m_{i-2}}(w_{ij}(t) - x_i(t)))^2$$

找出距离最小的输出节点作为获胜节点。

(5)调整获胜节点及其邻域内的节点权值:

$$w_j(t+1) = \begin{cases} w_j(t) + \eta(t)[x(t) - w_j(t)] & j \in N_C \\ w_j(t) & j \notin N_C \end{cases}$$

(6)降低学习率 $\eta(t)$ 和邻域函数 $N_C(t)$，转(4)，直到满足条件(达到指定的学习次数)。

(7)聚类结束，输出节点的个数即为类的个数。

(8)二次聚类。分析所有获胜神经元 i 和 k 之间的权值距离，计算类之间的差别明显程度:

$$L = \sqrt{\sum_{j=1}^{n+2}(w_{ij} - w_{kj})^2}$$

若 $L <$ 常数 C，则判定为同类。对于类对应的每个输入向量 $\boldsymbol{x}_i(t) = (x_{i1}(t), x_{i2}(t), \cdots, x_{in+2}(t))$，如果 $x_{i1}(t) = x_{j1}(t)$ 且 $x_{i2}(t) = x_{j2}(t)$，其中，$i \neq j$，则认为该类达到了聚类目的，记录最小多粒度时间和最大多粒度时间，得到 $((A_t, e):[\inf(D), \sup(D)])$。

(9)计算支持度 s、置信度 c 和近似精度；并输出满足给定的近似精度、$s \geqslant \underline{s}$、$c \geqslant \underline{c}$ 的模式。

4 实验及结果分析

对 580000 宝钢 JTB1 股票 2005 年 9 月 1 日~2005 年 9 月 2 日的高频交易数据进行了实验。设定多粒度时间格式为(分,10 秒), 买一价属性编码化状态规则编码为 5 种状态: (1)设 $SX_n^1 = x_n - x_{n-1}$，若 $SX_n^1 \in (-\infty, -5]$，意指买一价大幅下跌，置状态值为 1；(2) $SX_n^1 \in (-5, -1]$，意指买一价小幅下跌，置状态值为 2；(3) $SX_n^1 \in (-1, 1]$，意指买一价正常波动，置状态值为 3；(4) $SX_n^1 \in (1, 5]$，意指买一价小幅上涨，置状态值为 4；(5) $SX_n^1 \in (5, +\infty)$，意指买一价大幅上涨，置状态值为 5。设置阈值 $\underline{s} = 10\%$, $\underline{c} = 50\%$, $L = (10, 0)$，对多粒度近似周期的近似精度分别取(0,0), (0,2), (0,4), (1,0)。

在实验中发现了许多满足要求的模式的多粒度近似周期，其中周期短、置信度高、满足模式覆盖的多粒度近似周期是最有意义的，列出部分实验结果于表 1 中，对应的图如图 1 和图 2 所示。

表 1 多粒度高频数据实验结果

股票名称	时间段	近似周期	支持度	置信度	近似精度 sup(p)-inf(p)
580000 宝钢 JTB1	2005 0901 09:30:0 - 2005 0901 11:30:0	((买一价,3):[(5,4),(6,0)]	30.769 231	79.051 383	(0,2)
		((买一价,3):[(5,4),(6,2)]	35.538 462	91.304 348	(0,4)
		((买一价,3):[(5,4),(6,4)]	37.538 462	96.442 688	(1,0)
		((买一价,3):[(8,3),(8,5)]	29.846 154	76.679 842	(0,2)
		((买一价,3):[(8,1),(8,5)]	35.538 462	91.304 350	(0,4)
		((买一价,3):[(8,1),(9,1)]	38.000 000	97.628 460	(1,0)
	2005 0901 09:30:0 - 2005 0902 11:30:0	((买一价,3):[(3,5),(4,1)]	33.737 990	78.194 610	(0,2)
		((买一价,3):[(3,3),(4,1)]	39.757 210	92.145 370	(0,4)
		((买一价,3):[(3,2),(4,2)]	41.679 310	96.600 230	(1,0)

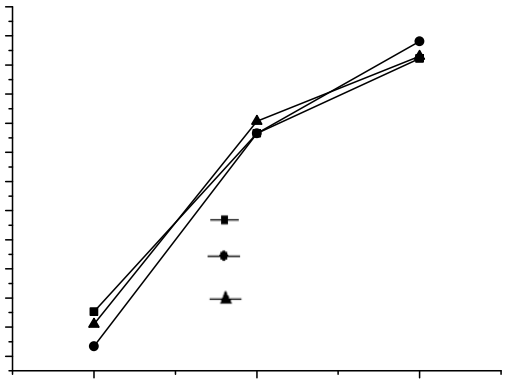

图 1 近似精度与模式的置信度的关系

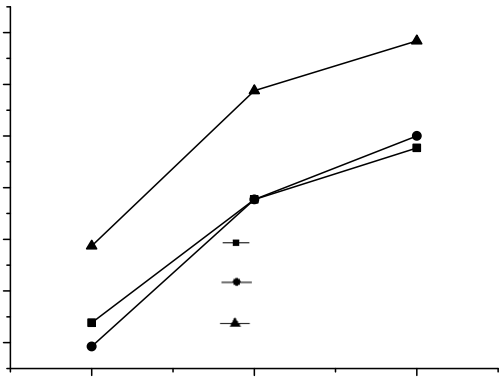

图 2 近似精度与模式的支持度的关系

从表 1、图 1 和图 2 可以看出:

(1)在时间段 2005 年 9 月 1 日 9 点 30 分-11 点 30 分，股票 580000 宝钢 JTB1 正常波动的近似周期为 6 min 左右，具体说来，580000 宝钢 JTB1 在某个多粒度时间呈正常波动状态,那么相隔 6 分 40 秒-7 分又出现该状态的可能性为 75.494 07%,相隔 6 分 20 秒-7 分出现该状态的可能性为 91.304 35%，而相隔 6 分 10 秒-7 分 10 秒的可能性高达 96.442 688%。在 8 分钟-9 分钟之间也存在类似的近似周期规律。

(2)多粒度近似周期规律满足性质 2，即对于相同对象相同的时间段，若模式 A 覆盖模式 B，那么 A 的置信度和支持度均高于 B。例如，股票 580000 宝钢 JTB1 在相同的时间段 2005 年 9 月 1 日 9 点 30 分-11 点 30 分，模式((买一价,3):[(5,4),(6,4)])覆盖模式((买一价,3): [(5,4),(6,0)])，显然模式((买一价,3):[(5,4),(6,4)])的置信度和支持度高于模式((买一价,3):[(5,4),(6,0)])。

(3)相同对象，若时间段不同，则得到的多粒度近似周期规律不同,这说明随时间的变化周期特性也会随之发生变化。

5 结束语

本文提出了多粒度近似周期模型和相应的挖掘算法，通过对股票的高频交易数据进行挖掘，发现了一些有意义的多粒度近似周期模式。

本文仅研究在给定多粒度时间格式下的多粒度近似周期挖掘问题，还存在如何在挖掘中自动选取时间粒度、如何自动选取多粒度时间格式、如何提高多粒度近似周期挖掘的算法效率等问题值得研究。

(下转第 88 页)