文章编号:1005－0523(2007)05－0081－04

一种基于 Rough 集的序列粒模型的构建

凌仕勇,谢剑猛

(华东交通大学 现代教育技术中心,江西 南昌 330013)

摘要:均对粒模型的构建问题,提出了基于 Rough 集的一种粒表示方法:文章首先对于时间序列作了粒的描述,采用 SAX 符号表示不仅因为它的适用于粒的描述,更重要的是为后面的逻辑推理打下良好的基础.然后用基于 SAX 距离的相似性找出相近的模式,利用 Rough 集的规则发现算法找出关联规则.通过实验证明,本文中所提出的时间序列挖掘方法以及时序粒挖掘的算法基本上可满足实际时序挖掘的需要.

关 键 词:粒计算,SAX,时间序列挖掘,Rough 集

中图分类号:TP311 **文献标识码**:A

1 引言

Pawlak 在不分明关系和 Rough 隶属函数的基础上,探讨了知识粒的结构和粒度问题.Peters 等人[1]使用不分明关系将实数划分为多个子区间,提出了两个信息粒之间的邻近关系以及包含关系的度量.J.T.Yao 和 Y.Y.Yao[2]使用粒计算模型来学习分类规则,用粒网格来表示学习所得的分类知识,提出了粒之间关联粒网格的一个算法.

目前,序列挖掘的应用已涉及到诸多领域,也出现了很多中描述序列模式的模型,但是,用粒的方法来定义和描述序列模式还鲜有报道.Anders 将时间信息系统 TIS 转换为信息系统 IS[3],进而用 Rough 集理论进行时间序列的挖掘.Curtis E.Dyreson 用粒的观点对时间序列进行了描述,并在 SQL 语言中实现了信息系统中不同粒度的转换[4].

粒计算的研究引起了广泛的关注,但还处于开始和发展阶段.本文提出了一种新的方法来挖掘时间序列:首先用粒对时间序列进行时序粒的构建;然后在构建好的模型上,采用时序相似性作为量度,用基于 Rough 集理论的挖掘方法得出时间系列蕴涵的规律.本文在这些理论和算法的基础上已成功地开发出了一个基于 Rough 集理论的粒计算时序挖掘平台.

2 基于 Rough 集的序列粒的构建

2.1 时间序列的粒表示

考虑从时间序列中抽取高质量的模式,粒子的表示不是一个简单数据预处理,一旦表示后,时间序列必须首先适合决策序列所需要的详细水平,然后定义这个粒的表示作为原始数据抽象的步骤,以获取最适合的决策目标的信息.在这里,我们目标定位于,通过有意义的符号来表达原始的时间序列,每一个符号表达了在一点时间里获取的行为.这样,时间序列就被抽象成为一序列符号——一些多维向量,这些多维向量的每一个描述了在一定周期内,监控参数所表达的某一刻状态.

在这里,定义一种时间序列的符号表达方式:SAX[5](Symbol Aggregate Approximation).首先定义一些用到的符号:

时间序列:$C=C_1,\Lambda,C_n$

时间序列的 PAA(Piece Aggregate Approximation):$\bar{C}=\bar{C}_1,\Lambda,\bar{C}_w$

收稿日期:2007－04－20

基金来源:江西省自然科学基金(0411035)

作者简介:凌仕勇(1975－),男,江西九江人,硕士,讲师,研究方向为粒计算,Rough 集理论及应用.

时间序列的符号表示：$\bar{C}=\bar{C}_1,\Lambda,\bar{C}_w$

w：代表时间序列C的PAA段数目

α：符号数(例如对于alphabet＝{a，b，c}，α＝3)

考虑长度为n的时间序列 $C=C_1,\Lambda,C_n$，通过向量 $\bar{C}=\bar{C}_1,\Lambda,\bar{C}_w$ 在w维空间中表示，则 $\bar{C}$ 的第I个元素 $\bar{C}_i=\frac{w}{n}\sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i}$.即，为了把时间序列从n维缩小到w维，数据被分成等宽的长为w的段.

在时间序列变换为PAA之前，还应对数据进行归一化处理，使得各变量的均值为0，标准方差为1，进而消除由于不同特征因子量纲不同和数量级不同所带来的影响.

当已经把时间序列表示成了PAA方式后，可用一种更进一步的转换以获取更好的离散化表达.因为归一化后的时间序列的累计概率分布呈现给我们的是一个高斯分布.为此，可定义“割点”以描述在高斯曲线下的α等宽区域.

定义1 割点 割点是一个排序后的列表 $B=\beta,\Lambda\beta_{\alpha-1}$，高斯曲线下的区域 β_i 到 β_{i+1} 是 $1/\alpha$，(α是割点数，$\beta_0=-\infty$，$\beta_\alpha=+\infty$).

为此，可通过建立一个统计表作为查询表如下：(β＝10，α＝3，…10)

表1 割点查询表

β \ α	3	4	5	6	7	8	9	10
β1	−0.43	−0.67	−0.84	−0.97	−1.07	−1.15	−1.22	−1.28
β2	0.43	0	−0.25	−0.43	−0.57	−0.67	−0.76	−0.84
β3		0.67	0.25	0	−0.18	−0.32	−0.43	−0.52
β4			0.84	0.43	0.18	0	−0.14	−0.25
β5				0.97	0.57	0.32	0.14	0
β6					1.07	0.67	0.43	0.25
β7						1.15	0.76	0.52
β8							1.22	0.84
β9								1.28

例如，当α＝3时，$\beta_{i+1}-\beta_i=1/\alpha=0.333$(i＝1，2)，查标准正态分布表当累计概率＝1−0.333＝0.667时 β＝0.43，根据对称性，有 β＝{−0.43，0.43}.

当α＝4时，$\beta_{i+1}-\beta_i=1/\alpha=0.25$(i＝1，2，3)，查标准正态分布表当累计概率＝1−0.25＝0.75时 β＝0.67，当累计概率＝1−0.25＊2＝0.5时 β＝0，根据对称性，有 β＝{−0.67，0，0.67}.

一旦求得了这些割点，可用以下方式来离散化一个时间序列：首先，计算出时间序列的PAA，所有在最小割点以下的PAA数映射成字符‘a’，接下来大于等于最小割点的数映射成字符‘b’，……．为此，可定义如下：

定义2 词SAX 长度为n的时间序列C可通过词 $C=C_1,\Lambda,C_w$ 定义如下：让alpha代表字母表的第i个元素，如alpha＝a，alpha＝b.然后映射PAA的 $\bar{C}$ 到词C，有 $C_i=alpha_i$(当且仅当 $\beta_{j-1}\leqslant\bar{C}_i<\beta_j$)，于是最终获得了最终需要的粒子符号表达式SAX.

2.2 时间序列的相似性度量

类似于上述定义，设 $B=\beta,\Lambda\beta_{j-1}$ 是排序后的断点列表，α是符号数(也即是上面谈到的割点数)，序列 $C=C_1,\Lambda,C_n$ 可转换为词 $C=C_1,\Lambda,C_w$，且 $C=\alpha_j$ 当且仅当 $\beta_{j-1}\leqslant\bar{C}_i<\beta_j$.

我们定义原始时间序列Q和C，以及它们转换后的词 $\hat{Q}$ 和 $\hat{C}$，其最小距离 mindist($\hat{Q}$，$\hat{C}$)定义如下[6,7]：

$$\text{mindist}(\hat{Q},\hat{C})=\begin{cases}0 & if\ |i-j|\leqslant 1\\ \beta_{\max(i,j)-1}-\beta_{\min(i,j)} & otherwise\end{cases}$$

对于p维时间序列 $C=((C_{1,1},\Lambda,C_{1,p},\Lambda,(C_{n,1},\Lambda,C_{n,p}))$，其转换后的词 $\hat{C}=((\hat{C}_{1,1},\Lambda,\hat{C}_{1,p},\Lambda,(\hat{C}_{n,1},\Lambda,\hat{C}_{n,p}))$，类似有 $\text{mindist}(\hat{Q},\hat{C})=\sqrt{\frac{n}{w}}\sqrt{\sum_{j=1}^{p}(d(\hat{q}_{i,j},\hat{C}_{i,j}}^2$

3 实验分析

在构建好时序粒模型和时序粒的算法基础上，

采用 Rough 集作为工具，实现了所有提出的时序挖掘算法，并完成了一个挖掘平台．本文以文献[8]提供的美国各项经济指标作为实验数据集，从中提取六项指标：①国民生产总值②居民收入和支配量③城市消费水平④工业生产指数⑤居民失业率；⑥联邦基金利率．试图找出以上 6 种时间序列之间的某种关系，其实现方法是首先利用 SAX 来表示这些时间序列的粒表示，然后利用模式发现算法找出子序列的模式集，最后利用 Rough 集的规则发现算法找出其中蕴涵的规律．

3.1　用 SAX 来表示时间序列

对于时间序列①，选取 n＝4，四组数据用一个 SAX 符号表示；alpha＝4，序列粒表示的粒度为 4．图 1 是当 n＝4，alpha＝4 时的 SAX 粒符号表示．

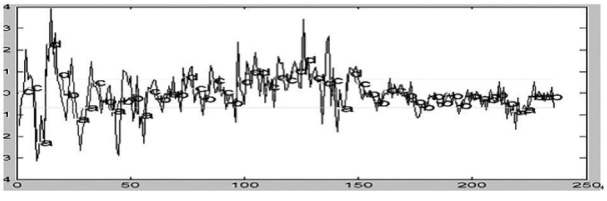

图 1　SAX 粒符号表示(n＝4，alpha＝4)

3.2　时序模式发现

时序模式发现是建立在对时间序列的粒解释之后的一个必要步骤．在时间序列的粒解释阶段，首先将时间序列进行分段，求出了每段的 PAA，进而用 SAX 进行时间序列的粒表达．如图 2 所示，假设有一个长 T＝1000 的时间序列，我们将它用 SAX 粒表达为 1000 个粒符号 ab…a…b．在模式发现阶段，需要给定一个观察窗口 w，这样，通过随观察窗口的滑动移动，我们就可以得到观察后长为 T－w 的子序列，每个子序列的长度为我们设定的观察窗口 w．

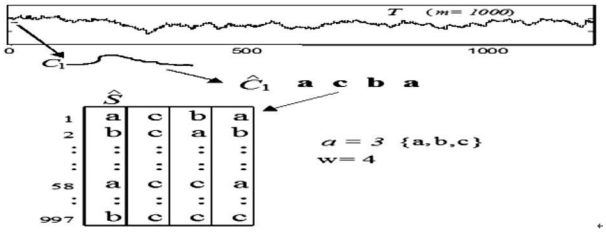

图 2　时序模式发现

3.3　时序规则的提取

下面是本文提出的时间序列挖掘的流程，有如下的步骤：

第一步：原始时序数据库的预处理——不完全数据的处理，缺失数据的处理，数据的归一化处理，数据粒表示(SAX)等等．

第二步：将预处理后的数据经过时序模式发现找出关联的子序列，得到时序数据的模式发现结果．

第三步：将模式发现的结果进行编码，得到用 Rough 集可处理的信息表．实际上，在本文的算法中，在第二步的输出结果中已将模式发现结果编码成可用 Rough 集处理的格式．

第四步：用 Rough 集数据约简和规则提取算法得出其中蕴涵的联系．

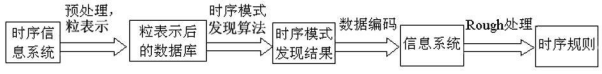

图 3　时序挖掘流程

采用本系统的挖掘平台，当 alpha＝4 时用迭代挖掘算法得到的最先规则和最终规则分别见图 4(a)－(b)

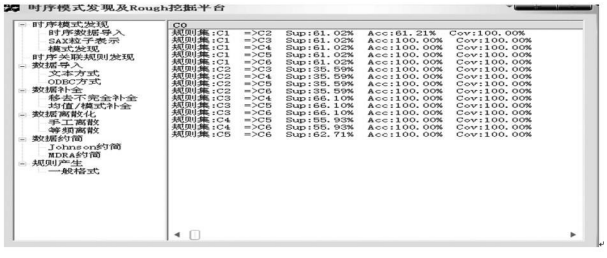

图 4(a)　两个序列时的规则(alpha＝4)

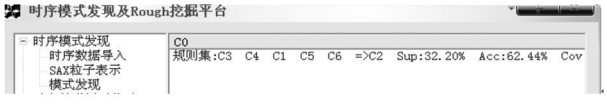

图 4(b)　六个序列时的规则(alpha＝4)

4　总结

粒计算的研究涉及两个基本问题的研究：粒结构的定义及在粒上的运算．本文针对时间序列挖掘的问题，用粒的观点对时间序列进行了粒结构的定义，提出了一套挖掘时序规则的方法，并实现了一个挖掘实验平台．本文在比较各种时间序列的表示方法基础上，采用用一种新的粒表现形式 SAX 来表示时间序列．一方面它符合时间序列的表现形式(高斯概率分布)，另一方面它的符号表示方法为后面的 Rough 集逻辑推理奠定了基础．然后在构建合适的粒模型基础上，利用粒形式的距离量度来匹配发现相近的模式，用 Rough 集的规则发现迭代找出多维时间序列之间的某种联系．

以上的所有工作都通过一些实验进行了验证，实验证明，本文中所提出的时间序列挖掘方法以及时序粒挖掘的算法基本上可满足实际时序挖掘的需

要.但是,粒计算的研究和实现是一个不断探索前进的过程,在时序规则的挖掘上,本文提出的挖掘方法尽管证明是一种成功的方法,但本文的某些细节有待提高.例如,在粒子构建上,粒度大小的选择现在是依靠人为指定的,如何根据后续规则来调整粒度大小,以得到更好的时序规则,将对本文粒运算的算法提出更高的要求.

参考文献:

[1] Peters J·F·, Skowron A·, Suraj Z·, Borkowsi M·and Rzasa W·, Measures of Inclusion and Closeness of Infromation Granules: A Rough Set Approach[C]·Proceedings of the Thrid International Conference on RSCTC '2002, October 14 — 16, 2002, Malvern, PA, USA, Springer, 300—307.

[2] Yao J·T·and Yao Y·Y·, Induction of Classfication Rules by Granular Computing[C]·Proceedings of the Third International Conference on RSCTC '2002, October 14—16, 2002, Malvern, PA, USA, Springer, 331—338.

[3] Adners T·B·, Mining Time Series Using Rough Sets—A Case Study[J]·Proceeding of first Europe Symposium, PKDD '97, 361—158, 1997.

[4] Dyreson, C·E·, William S·Evans, Hong Lin, and R·T·Snodgrass·Efficiently Supporting Temporal Granularities [R]·September 3, 1998.

[5] J·Lin, E·Keogh, S·Lonardi, and B·Chiu, A symbolic Representation of Time Series, with Implications for Streaming Algorithms [C]·in Proc·of the ACM Workshop on Research Issues in Data Mining and Knowledge Discovery(DMKD '03), San Diego, CA, 2003, 2—11.

[6] J·Buhler and M·Tompa, Finding motifs using random projections[J]·Journal of Computational Biology, vol·9(2), 225—242, 2002.

[7] M·Vlachos, G·Kollios, and G·Gunopulos, Discovering Similar Multidimensional Trajectories[C]·in Proc·of the 18th International Conference on Data Engineering (ICDE '02), San Jose, CA, February 2002, 673—684.

[8] http://www·economagic·com/popular·htm

A Model of a Temporal Granules Based on Rough Set

LING Shi-yong, XIE Jian-meng

(Center of Modern Education and Technoloyy, East China Jiaotong Univ; Nanchang 330013, China)

Abstract: For the problem of granular model 's representation, the paper proposes a method of representing granules based on Rough Set·Firstly it makes a granular description regarding the time series·Using the SAX symbolic representation is due to its suitability and good foundation for the later logical reasoning·Then it discovers the pattern with defined temporal granules based on the SAX distance similarity·Finally it uses the excavation algorithm to discover the relative rule·Experiment proves the method of mining time series and the algorithm of temporal granules may meet the need of actual use·

Key words: granular computing; SAX; mining time series; rough set