维普资讯 http://www.cqvip.com

发现模糊状态演化模式*)

Finding Fuzzy States Evolution Patterns

张保稳　何华灿　冯红伟

（西北工业大学计算机系　西安710072）

Abstract　Temporal data mining is one of the important branches of data mining. Current researches on time series in temporal data mining mostly are similarity research. In this paper we propose a way to find fuzzy states evolution patterns and then to extract rules from time series. Experiments show that the patterns and rules obtained are meaningful to predict the tendency of time series.

Keywords　Temporal data mining, Similarity search, Fuzzy states evolution patterns, States evolution rules, Time-delay embedding

0　引言

时间序列(Time Series)是指按时间顺序排列的一组数据。对时序数据进行分析，从中获取生成这些数据的系统的相关信息从而完成对系统的模型构造和对系统的未来的行为做出预测，具有重要的价值和意义。

数据挖掘(也称为数据库中的知识发现)，是指从数据中提取模式的过程，这些模式是有效的、新颖的、潜在可用的和易于理解的[1]。当前数据挖掘领域对时间序列进行的研究主要限于时间序列的相似性研究，即从同一时序或者不同时序中发现相似模式[9]。对于如何从一个时间序列中提取知识的问题，国内外尚很少见。本文首先从系统论的角度对时间序列问题进行了分析，然后将模糊性引入到时序处理中，提出了从时间序列中进行频繁状态演化模式挖掘的问题，然后论证并给出了其挖掘算法，最后在实际应用中对上述理论进行了分析和验证。

1　问题的背景

从系统论的观点出发，时间序列可以看作为系统输出的一部分，而系统内部的动力学机制是未知的或所知信息是有限的。本质上来说，时间序列问题就是对可以获得的部分的系统输出数据进行分析，提取其蕴含的系统特征，构造对应的等价系统从而完成对该系统的功能刻画，并依据相应的模型完成对系统未来行为的预测的过程。

时间序列分析领域当前的研究现状是：由于实际应用中时间序列具有不规则、混沌等非线性特征，使得预测系统未来的全部行为已经不可能，对系统行为的精确预测效果也难以令人满意[2]。这使得我们不得不转向对系统的关键行为进行预测和建模。在解决时间序列问题的思路上，由原来的应用概率论、随机过程等纯数学的方法，逐渐转变为引入模式识别、机器学习等人工智能技术和数学手段相结合的方法。

为了解释过去和预测未来，我们需要去发现那些隐藏在现象背后的规律。对于一个系统而言，如果存在确定性的方程，理论上讲我们可以求解它们并用于预测未来的输出。如果这些方程我们并不知道，我们就必须从过去的数据中发现主宰系统演化的规律。

在本文中，我们假定生成时序的动力学系统在其状态空间中的状态演化情况已知，并且该系统不是完全随机的，即系统的演化存在确定性的规律。我们尝试从状态演化的历史数据中发现这些规律，这就是我们提出的状态演化模式挖掘问题。

2　状态演化模式挖掘

给定系统S，设状态量为d个，其t_i时刻的状态向量为$s_i \in R$，其中状态空间R为d维实数空间，已知S在R中的状态演化序列$S=s_1s_2\cdots s_n$。

定义1(状态演化片段)　给定状态演化序列$S=s_1s_2\cdots s_n$，我们定义s上的任意一个连续子序列$s_js_{j+1}\cdots s_{j+m-1}$为状态演化片段。子序列的长度为状态演化片段的长度。一个状态演化片段对应着系统在状态空间内演化的一段连续的轨迹。

定义2(状态演化片段集合)　给定状态演化序列$S=s_1s_2\cdots s_n$和状态演化片段长度m，则状态演化片段集合$S_m=\{s_js_{j+1}\cdots s_{j+m-1}|j=1,\cdots n-m+1\}$。显然$|S_m|=n-m+1$。

定义3(状态演化模式)　给定状态演化序列$S=s_1s_2\cdots s_n$，状态$s_i\in R^d$，长度为m的状态演化模式为$P=a_1a_2\cdots a_m$，其中$a_i\in R^d$，状态演化序列S对模式P的支持度定义为状态演化片段集合S_m对P的支持度$\sigma(S_m,P)$。

我们的挖掘目标是从S中发现规律性出现的状态演化片段，即从中发现状态演化模式，这些模式应该具有一定的支持度。类比关联规则中的支持度定义，我们可以直观地将状态演化片断集合Sm中模式P出现的频度作为状态序列S对模式P的支持度。但是由于状态向量s一般情况下是连续性的量值数据，我们需要对其进行离散化处理。比较常用的离散化方法便是对状态空间中已经出现的状态数据进行聚类。为了保证后面挖掘结果的价值，我们希望经过离散化后的数据能够尽可能多地保留原有数据中的信息。但一般的聚类方法属于硬分类，被聚类的对象被严格区分出来，一个对象只能划分到一个类中，经过处理后原来数据中所蕴含的信息损失较多。这里我们采用具有一定柔性特征的模糊聚类方法。在模糊聚类中，一个被聚类对象属于多个类，但隶属度不同。模糊化的分类结果比明确的分类结果中包含更多的信息，这种方法在模式识别中已经得到广泛的应用[7]。这样，原来的状态演化模式就相应地转变为模糊状态演化模式。

定义4(模糊状态演化模式)　给定状态演化序列$S=s_1s_2\cdots s_n$，其中状态$s_i\in R^d$，设对已出现的状态集$\{s_1,s_2,\cdots,s_n\}$进

*)本文得到国家教委博士点基金(98069923)资助。张保稳　博士生，主要从事人工智能、时间序列、数据挖掘的研究；何华灿　博士生导师，主要从事人工智能、泛逻辑理论的研究。冯红伟　博士生，从事数据库、数据仓库、数据挖掘研究。

维普资讯 http://www.cqvip.com

行模糊聚类处理后得到模糊集合 $FS=\{cs_1,cs_2,\cdots,cs_k\}$，则长度为 m 的状态演化模式为 $P=\alpha_1\alpha_2\cdots\alpha_m$，其中 $\alpha_i\in FS$，定义状态演化片段 $s_is_{i+1}\cdots s_{i+m-1}$对模式 P 的支持为 $s(j,P)=\min\{\mu\alpha_i(s_{i+i-1})|(i=1,\cdots,m)\}$。

定义5(模糊状态演化模式的支持度) 定义状态演化序列 S 对模式 P 的支持度为状态演化片段集合 S_m 对 P 的支持度。$\sigma(S,P)=\sigma(S_m,P)=(\sum_{j=1}^{n-m+1}s(j,P)/(n-m+1)$，其中 s(j,P)为状态演化片段 $s_is_{i+1}\cdots s_{i+m-1}$对模式 P 的支持，$j=1,\cdots,n-m+1$。

定义6(频繁模糊状态演化模式) i 频繁模糊模式集合 $L_i=\{P_i|\sigma(S,P_i)>\sigma_{min}\}$。

$L=\bigcup L_i$ 成为频繁模糊状态演化模式集合。其中 σ_{min}为最小支持度。

定义7(状态演化规则) 在得到状态演化模式后，我们就可以从中发现状态演化的规则。对于任意状态演化模式 P，我们可以得到下列形式的状态演化规则。

(1)规则 $A\Rightarrow B$，AB=P，其中 A，B 均为模式，AB 为 A 和 B 的连接串。

(2)$\sigma(S,AB)>\sigma_{min}$，$\psi(S,A\Rightarrow B)>\psi_{min}$，其中 $\psi(S,A\Rightarrow B)=\sigma(S,AB)/\sigma(S,A)$为规则的可信度，$\psi_{min}$为最小可信度。

由上可知，状态演化规则挖掘实质就是从状态演化序列中挖掘出满足以上条件的规则，进行状态演化规则挖掘的关键是从中发现频繁模糊状态演化模式集合。

3 状态演化的频繁模糊模式集的生成

类比关联规则挖掘中频繁项集的生成算法，我们发现长度为 k 的频繁模糊状态演化模式的候选集可以从长度 k－1 的频繁模糊状态演化模式集中生成出来。

定理1 任意长度为 k 的频繁模糊模式集 L_k 都是集合 C_k 的子集，其中 C_k 由下列方式生成：$C_k=\{P_k|P_k=AP_{k-2}B$，其中 $AP_{k-2}\in L_{k-1}$，$P_{k-2}B\in L_{k-1}$，$A\in FS$，$B\in FS\}$。

证明：设 $P_k\in L_k$，且 $P_k=AP_{k-2}B$，其中 $A\in FS$，$B\in FS$。根据支持度的定义和频繁模糊序列模式的定义，有 $\sigma(S,P_k)>\sigma_{min}$。即

$$\sigma(S,P_k)=(\sum_{j=1}^{n-k+1}s(j,P)/(n-k+1)>\sigma_{min}$$

又因为对于任一窗口子序列 $s_i=s_is_{i+1}\cdots s_{i+k-1}$，其对模式 P 的支持度为 $s(j,P)=\min\{U_{j,i}|(i=1,\cdots,k)\}$，对模式 $P'=AP_{k-2}$的支持度为 $s(j,P')=\min\{U_{j,i}|(i=1,\cdots,k-1)\}=\min(\{U_{j,i}|(i=1,\cdots,k-1)\}\cup\{1\})\geqslant s(j,P)$。同时又因为 $n\gg k$，故：

$$\sigma(S,P')=\frac{\sum_{j=1}^{n-k+2}s(j,P')}{(n-k+2)}\approx\frac{\sum_{j=1}^{n-k+1}s(j,P')}{(n-k+1)}\geqslant\frac{\sum_{j=1}^{n-k+1}s(j,P)}{(n-k+1)}>\sigma_{min}$$

所以 $P'\in L_{k-1}$，即 $AP_{k-2}\in L_{k-1}$。同理，$P_{k-2}B\in L_{k-1}$。由此可知，候选的长度为 k 的模糊序列模式可以由长度为 k－1的模糊序列模式集合生成。问题得证。

在 Apriori 算法中，候选频繁项集在通过连接生成后，还需要对其进行剪枝处理。设 $C'_k=L_{k-1}*L_{k-1}$，对于 C'_k 中的项集 X，如果 X 的任何一个长度为 k－1的子集 Y 是非频繁项集，即 $Y\notin L_{k-1}$，则将 X 从 C'_k 中去除。经过这种处理后最后才得到 C_k。对于频繁模糊序列模式集而言，同样设 $C'_k=L_{k-1}*L_{k-1}$，对于任意 $P_k\in C'_k$，设 $P_k=AP_{k-2}B$，由于模糊序列模式的连续性和序关系的存在，则 P_k 对应着 AP_{k-2}和 $P_{k-2}B$ 两个长度为 k－1的模式，又根据 * 操作的含义，知已经有 $AP_{k-2}\in L_{k-1}$，$P_{k-2}B\in L_{k-1}$。所以没有必要再对 C'_k 进行剪枝，$C_k=C'_k$。

于是，我们得出如下状态演化的频繁模糊模式集的生成算法。且有，当状态演化序列的长度为 n 时，算法的复杂度为 $\sum_{i=1}^{k}|C_i|n+\sum_{i=1}^{k-1}|L_i|^2$。

输入：模糊集合 FS，状态演化序列 S，σ_{min}
输出：频繁模糊模式集 L
过程：
$C_1=FS$;
$L_1=\{P_1|P_1\in C_1$，且 $\sigma(S,P_1)>\sigma_{min}\}$;
$k=2$;
While $L_{k-1}\neq$Null Do {
$C_k=\{P_k|P_k=AP_{k-2}B$，
其中 $AP_{k-2}\in L_{k-1}$，$P_{k-2}B\in L_{k-1}$，$A\in FS$，$B\in FS\}$;
$L_k=\{P|P\in C_k$，且 $\sigma(S,P_k)>\sigma_{min}\}$;
$k=k+1$;
};
$L=\bigcup_k L_k$。

4 对单一时序的预处理

状态演化规则挖掘是在状态变量全部已知的条件下提出的。然而在实际应用中，我们面对的常常是单一的时序。在这种情况下，我们可以首先对时序进行基于 Taken's 定理的时间延迟嵌入，重构状态空间，然后在重构的状态空间进行状态演化规则挖掘。

时间延迟嵌入理论是20世纪80年代基于微分拓扑和动力学系统的一些思想提出的，是时间分析研究的一次突破。可以用于辨识由确定性系统产生的时序数据，并抽取蕴含在观察数据下的系统几何特征。Takens 定理证明，在给定条件下，一个未知的系统的状态空间可以按一种特定的方式重建[8]。如果嵌入得当，Takens 定理保证了重建后的动力学系统和原动力学系统具有拓扑意义上的相似性，因此动力学不变特征量也相同。这样，对于一个给定的时序，通过时间延迟重建技术可以重构一个和原时序系统状态空间拓扑同构的状态空间。

一般地，给定一个时序 $s=\{x_1,x_2,\cdots,x_n\}$，d 为嵌入维数，τ 为延迟时间，则由时序 s 重构出来的子时序 $s_i=(x_i,x_{i-\tau},\cdots,x_{i-(d-1)\tau})$，生成的子时序集合为 $W(s)=\{s_i|i=(d-1)\tau+1,\cdots,n\}$，重构后的序列为 $s'=s_is_{i+1}\cdots s_n$。

用时间延迟嵌入理论来进行时序的预处理，能够从数学意义上严格地保证经过变换后系统的动力学规律等价性。这种等价的一个充分条件就是 Takens 定理，它要求嵌入维数至少比原系统维数的2倍多1，即 $d\geqslant 2Q+1$。实践中，d 的选择具有后验性，可以通过尝试来确定。而且，即使不满足此条件，重新生成的子时序集合中仍然能够提取原系统的有用信息。另外，延迟时间 τ 的确定原则上并不重要[6]，在我们的算法中，可以取 $\tau=1$，这时延迟处理退化为常用的滑窗技术；$\tau\neq 1$ 时的情形我们以后将专文另作讨论。相应的一个模糊聚类概念将对应着时序延迟窗口内的一类演化趋势。模糊序列模式则对应着时序在各类演化趋势之间的转变模式。

5 实验

上面已经给出了算法的理论复杂度，这里限于篇幅，在实验情况的分析中，对算法性能部分的分析我们不再讨论。下面是我们的实验情况。

我们选取美国联邦基金每日利率作为实验数据对象。数据内容为从2001年2月28日往前的1000个数据点。数据来源为互联网站 http://www.economagic.com/，该网站专门提供各类经济时间序列为科学研究和商业应用服务。我们试图发现基金利率的起伏的规律性。为此，我们首先对数据进行了预处理，得到了利率的增量序列，然后对增量序列进行挖掘。我

维普资讯 http://www.cqvip.com

们发现在给定的参数下算法能够发现一些有用的规则。为了对实验结果进行某种程度的验证，我们在模糊聚类数目为20，支持度为0.035时发现的规则中选取了两条规则，它们对应着在该实验参数下所能发现的最长的模糊序列模式。(这是因为，一般地，序列模式越长，预测性能越低。)我们对2001年3月1日至6月15日之间的数据进行了验证。发现规则1可以得到一定的验证，验证情况如图，对于规则2未发现相应的验证情形。

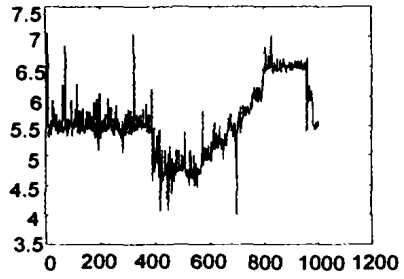

图1 联邦基金日利率图

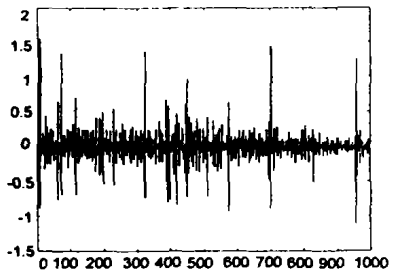

图2 联邦基金日利率增量图

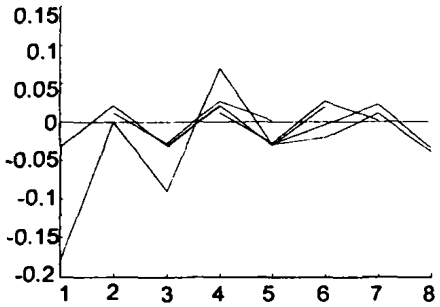

图3 标准数据和预测数据匹配图

表1 对应最长状态演化模式的两条规则

规则	规则前件	规则后件	前件支持度	AB支持度	可信度
1	18 2 18 2	19	0.0377	0.0351	93.1%
2	19 18 2 18	2	0.0390	0.0353	90.51%

图中的实线为2001年5月24日起8日内的增量序列，虚线为拼凑后的模糊模式。我们可以看到在前7日的模式和原数据之间匹配效果不太良好的情况下，规则仍然能对第8日的曲线走势作出一定的预测。我们可以发现在第8天，增量由0线上方下行到0线下方，变为负数，这意味着联邦利率即将下降。

6 相关工作

当前时态数据挖掘中对时间序列的处理主要是从数据库角度出发，研究相似性搜索。在如何从时间序列中发现知识，进行深层的数据挖掘的研究方面，工作并不多见。

Michael T. Rosenstein 等人提出了一种从时间序列中发现概念的方法[3]。这里概念是预测意义上的，概念就是模式的预测内容。在这个过程中，Rosenstein 利用了时间序列数据背后动力学系统的性质，首先对时间序列进行了延迟嵌入，然后对延迟后的数据进行了一种动态聚类。实验表明，通过这种方式形成的数据分类可以很好地对应物理意义上的概念。

M. Gas 等人提出了一种从时间序列中发现规则的方法[4]。这里他采用了常用的滑窗技术对时序进行预处理。然后对形成的窗口向量集合进行了聚类。再用这些类对原来的时序进行重构。这样就完成了对时序进行离散化和符号化的过程。接下来，对于重构后的时序进行规则发现。但是 M. Gas 等并没有对他们的工作给出合理的理论解释，而且他们发现的规则形式过于简单。

Heikki Mannila 等人在对无线通讯网络故障管理数据库进行处理时提出了从事件序列进行模式发现的问题[5]。给定一个输入的事件序列、一类偏序事件集合模板，事件模式发现就是从序列中发现满足频度阈值的符合偏序模板的模式。其中的串行模板和我们算法中的模式相似，但是他们的算法并不适用于数值型的时间序列。

结论 本文首先从系统论的观点出发，提出了从时间序列中进行状态演化模式挖掘的问题，然后我们将模糊性引入到问题的解决方法中，提出了一种模糊状态演化模式的生成算法并给出了其关键部分的证明。最后我们对算法进行了实现。实验表明采用这种方法挖掘出的规则可以从时序中挖掘出状态演化模式，产生一些有价值的规则。用它们可以进行时序演化趋势的预测。

当前我们对规则的评价主要是人工根据领域知识进行的。然而一般地，算法产生的规则数目是大量的。结合模糊状态演化模式挖掘的情况，提出具有一定普适性的标准来度量生成规则的价值，这具有重要的意义。这是我们以后的研究内容。

参考文献

1 Agrawal R, Mamnila H, Srikant R, et al. Fast Discovery of Association Rules. In: Fayyad M, Piatetshy-Shapiro G, Smyth P, eds. Advances in Knowledge Discovery and Data Mining, Menlo Park, California: AAAI/MIT Press, 1996. 307～328

2 Weigend A S, Gershenfeld N A. Time Series Prediction: Forecasting the Future and Understanding the Past. eds. Reading, MA: Addison-Welsley, 1993

3 Rosenstein M T, Cohen P R. Concepts from Time Series. In: Proc. of the Fifteenth National Conf. on Artificial Intelligence, 739～745

4 Das G, et al. Rule discovery from time series. In: Proc. of the Fourth Intl. Conf. on Knowledge Discovery and Data Mining, 1998

5 Mannila H, Toivonen H, Verkamo A I. Discovering frequent episodes in sequences. In: Proc. First Intl. Conf. on Knowledge Discovery and Data Mining (KDD-95), Montreal, Quebec, Canada. AAAI Press, Menlo Park, California. 1995. 210～215

6 Liebet W, Schuster H G. Proper choice of the time delay for the analysis of chaotic time series, Phys. Lett. A, 142, 1988. 107～111

7 边肇祺，张学工. 模式识别. 北京：清华大学出版社，1999，12：273～283

8 Takens F. Detecting strange attractors in turbulence. In: Proc. of Dynamical Systems and Turbulence, Warwick, 1980. 366～381

9 Roddick J F, Spiliopoulou M. A Biblograph of Temporal, Spatial and Spatial-Temporal Data Mining Research. In Proc of ACM SIGKDD, 1999, 1(issue 1): 34～38