时间序列表示进展及比较研究:时间序列挖掘建模环境[1]

李俊奎，王元珍，刘城成，曹忠升
华中科技大学数据库与多媒体研究所(430074)
email:jkltk2000@126.com

摘　要: 时间序列表示是时间序列挖掘的一个基础和关键问题。对当前出现的各种典型的时间序列表示方法进行了综述，对各自的特点从多个角度进行了比较研究。结果说明，大部分时间序列表示方法将时间序列降维，且都与应用领域紧密相关，在实际构建系统时仍需对各种表示方法按照实际需求进行转化和改造。

关键词: 数据挖掘 时间序列 表示 建模

1．引言

时间序列是一种重要的高维数据类型，它是按照时间顺序观察所得到的一串数据。时间序列的应用日益广泛，其涉及天文、地理、生物、物理、化学等自然科学领域，图像识别、语音处理、声纳技术、遥感技术、机械工程等工程技术领域，以及市场经济、金融分析、人口统计、地震检测等社会经济领域，当前对于时间序列挖掘的研究正得到越来越多的重视[18][15]。

本文的背景是国家发展与改革委员会“安全智能整合平台开发及产业化”项目，该项目的一个重要目标是以人工智能、数理统计等先进的数据挖掘技术为基础，满足用户的智能化知识发现和趋势分析的需求，为用户战略决策提供服务。在项目进行中，构建时间序列挖掘子系统时，我们面临首先必须解决的一个重要而基础的问题是对时间序列进行建模表示。经过大量深入细致的相关研究和文献查阅，我们发现当前对于时间序列的表示问题虽然取得了部分进展，但是如果需要将研究成果应用于实际系统构建过程，仍然需要深入考察以及对各种表示方法进行实际转化和处理。

本文综述了当前时间序列挖掘研究领域中出现的各种时间序列的建模表示方法，指出在构建实际系统时，这些表示方法都存在各自的问题。并且指出时间序列表示方法是一种与应用领域和应用需求相关的方法，为实际的时间序列表示和建模提供了参考。

本文的其余部分如下组织：第 2 节讨论时间序列表示的相关背景；第 3 节对当前已经出现的时间序列的表示方法进行综述，陈述各自的不足；第 4 节对各种时间序列表示方法进行比较；第 5 节总结全文，并指出未来的进一步工作。

2．时间序列表示的相关背景

为了说明的严谨，我们首先给出时间序列的相关定义。

定义 1 时间序列(Time Series) 时间序列$T=t_1,t_2,...,t_n$是一串有序的n实数变量。

定义 2 时间序列长度(Length of Time Series) 对于有限长时间序列$T=t_1,t_2,...,t_n$，T的长度为组成T的实数个数，记为$|T|$，即$|T|=n$。对于无限长时间序列，T的长度定义为$|T|=\infty$。

无限长时间序列一般在数据流的建模中使用，有限长时间序列则在时间序列数据库中使用。

定义 3 时间序列区段(子序列) (Segment, Subsequence) 给定长度为n的时间序列T，T

[1]本课题得到国家发展与改革委员会“安全智能整合平台开发及产业化”项目(项目编号[2005]538 号)资助

的序列区段C_s是在T中从点t_s开始，数量为$w(1 \le w \le n)$个连续位置点所组成的T的一个抽样，即

$$C_s = t_s, t_{s+1}, ..., t_{s+w-1},$$

其中$1 \le s \le n-w+1$。

时间序列区段一般通过在T上给定一个滑动窗口(窗口大小为w) 获得。

实际生活中的时间序列都具有很长的长度，一般对于长度为n的时间序列可以看作是n维向量空间中的一个点，为了便于查找，可以首先利用 R tree，R* tree, k-d tree， skyline index[9]等多维索引机制将这些点索引，然后查找则首先在索引上进行。不幸的是，由于目前现存的多维索引方式普遍仅对 8-10 维空间中的点比较有效，对于更高维的数据索引则将会导致索引性能的急剧下降，即引发所谓的“维度灾难”(*Dimension Curse*)问题[4][16]。

为了解决维度灾难问题，一般的做法是对时间序列数据进行降维处理，然后再对降维后的数据进行索引，对原始数据进行降维处理后，在索引空间的查找可能出现两类问题[25]：

(1)漏查(False Dismissal)：在原始数据(T)中两点距离小于给定的阈值ε，但是在索引空间($\overline{T}$)中的该两点距离却大于ε，从而对索引空间的点查询时发生漏查，即

$$\begin{aligned} &\exists t_i, t_j \in T, \overline{t_i}, \overline{t_j} \in \overline{T}, \varepsilon > 0, \\ &D(t_i, t_j) < \varepsilon \Rightarrow D_{index}(\overline{t_i}, \overline{t_j}) \ge \varepsilon \end{aligned} \quad (1)$$

(2)错查(False Positive)：在索引空间($\overline{T}$)中的两点距离小于给定的阈值ε，但是在原始数据(T)中该两点距离却大于ε，从而对索引空间的点查询的结果中出现错查，即

$$\begin{aligned} &\exists t_i, t_j \in T, \overline{t_i}, \overline{t_j} \in \overline{T}, \varepsilon > 0, \\ &D_{index}(\overline{t_i}, \overline{t_j}) < \varepsilon \Rightarrow D(t_i, t_j) \ge \varepsilon \end{aligned} \quad (2)$$

对于错查问题，可以通过针对索引空间中的查询结果再次到原始数据空间中查询，剔除其中$D(t_i, t_j) \ge \varepsilon$的点来解决，由于在索引空间中查询时已经剔除了大量不符合条件的点，只保留了原时间序列数据集合中一个很小的子集，所以再次在原始数据空间中查询时的耗费是可以接受的。漏查问题则决定了是否能够对时间序列进行有效的相似性查找，为了能够解决这个问题，Faloutsos 等人在文献[7]中给出了降维下界定理**(Lower Bounding)**，即：

$$D_{index}(T, C) \le D(T, C). \quad (3)$$

于是很多时间序列表示方法都侧重于首先对时间序列进行降维处理。

3 各种典型的时间序列表示方法

在研究初期提出对时间序列进行 *DFT*(Discrete Fourier Transform, 离散傅立叶)变换[1][13]，然后用 *DFT* 的前k个系数作为原时间序列的表示，其底层的理论依据是数字信号处理领域的 *Parseval* 定理，该定理保证了时间序列数据的 *DFT* 变换前几个系数中保存了序列中大部分能量。在实际应用中， *DFT* 变换对于自然产生的时间序列信号较为适合，但是对于其他来源的时间序列数据则效果不佳。

随后出现了 *DWT* (Discrete Wavelet Transform，离散小波变换)[4][23]，其中研究最多的是 *Haar* 小波变换，并用 *Haar* 变换的系数作为时间序列的表示。 *Haar* 小波变换的基函数不平滑，对时间序列的表示近似类似于梯形的结构，导致其对于一段很短的时间序列数据进行变换都会产生大量的系数。

SVD(Singular Value Decomposition, 单值分解)变换技术[22], 采用 *KL* 分解的技术来实现时间序列数据的降维处理。这种技术的劣势在于其依赖于数据，由于使用数据集来产生新的基向量，因此数据项的任何改变都需要重新进行计算。

PAA(Piecewise Aggregate Approximation, 滑动平均聚集近似)方法[16][24], 在时间序列上滑动一个大小固定的滑动窗口，并计算滑动窗口中数据的均值作为整个窗口内数据的表示。这种方法利用了时间序列在短期内数据变化不大的特性，能够在一定程度上实现时间序列的有效降维。但这种方法具有如下的不足：1)滑动窗口的大小是一个关键的因素，实际中需要根据时间序列数据仔细选取；2)利用求均值的方法对时间序列进行平滑处理，可能会丢失时间序列中的极值等特征信息；3) *PAA* 没有考虑到时间序列数据的随着时间推进对于未来数据大小的参考价值越来越大的性质，它对每段时间序列都同等对待[29]。

APCA(Adaptive Aggregate Constant Approximation，自适应平均聚集常量近似)方法[17]，将时间序列分段成为一系列变长的子段，每一子段用该子段中数据均值以及时间点右端值组成的二元组表示。相比较 *PAA* 方法，它克服了要求滑动窗口大小唯一的限制。但它具有如下不足：1)同等条件下，它需要 2 倍于 *PAA* 方法的存储量；2)在进行查询时需要对查询串采取同样长度的分解；3)对长度为 n 的时间序列完成 N 段精确的 *APCA* 表示，需要 $O(Nn^2)$ 时间复杂度，优化以后仍需 $O(n\log(n))$，在时间序列流的情形下则不适合。

PLA(Piecewise Linear Approximation, 分段线性近似)[7][26]方法首先将时间序列分段，然后利用线性拟合函数进行近似表示。这种方法的缺陷主要是计算复杂性较高，对长度为 n 的时间序列完成 N 段最优的分段拟合需要 $O(n^2N)$ 的时间复杂度。虽然有很多的优化方法[20]，但是仍难以满足实用的要求。

文献[28]中提出一种维约简的方法 *PRA*(Piecewise Re gression Approximation, 逐段回归近似)的方法，该方法对时间序列滑动窗口中的数据计算回归系数，并利用回归系数作为窗口内数据的表示。它的结果对均值平稳的独立噪声干扰不敏感，较 *PAA* 方法在进行相似性查找时准确度更高。但该方法的表示结果难以为人所直观理解。

文献[27]中提出一种基于向量空间 $\{1,t,...,t^n\}$ 的多项式拟合的方法，利用拟合的系数作为时间序列表示，这种方法过于复杂，在实际应用中求解将会出现困难。

普通的时间序列都是离散的数值型数据，研究过程中也出现了将数值型数据量化形成字符串形式，然后采用常用字符串匹配操作[2][3](如 Hash，Markov 模型，Suffix Tree)等相对成熟技术的方法。

文献[19]中提出一种 Clipper Data 方法，将时间序列进行“过零”量化成一串对应的二进制数串，利用二进制的算术运算来完成比较。这种方法对时间序列的数值量化过于简单，虽然采用了二进制压缩技术，但是实际中仍会出现大量冗余数据。

文献[11]中提出一种 *SAX* (Symbolic Aggregate approXimation，符号聚集近似)的时间序列表示方法，该方法首先采用 *PAA* 对时间序列进行时间轴向分段，然后根据正态分布，对数据轴向量化，量化后的结果利用查表方法确定相应段的字符串，从而完成时间序列的符号化。该方法被应用于时间序列的模式抽取和可视化中[10][12]。不同于一般的等值量化方法，它假设数据呈现正态分布，采用的是不等值量化方法。但是这种方法有如下不足：1)采用 *PAA* 对时间序列进行分段，无法克服 *PAA* 方法的缺陷；2)对数据的随机分布具有硬性规定；3)实际应用中字符串编码规则仍然需要仔细确定。

在时间序列相似问题的研究过程中，一些研究人员独辟蹊径，他们不是从整个时间序

列出发，而是选取时间序列中的若干特殊的数据点，从而完成时间序列的表示。

文献[6]中提出一种称为 Landmarks 的时间序列表示方法，将时间序列中的极值点、弯曲点识别出来，并在这些点上建立模型，从而完成原时间序列的近似表示。

文献[21]中提出一种类似的 Important point 的时间序列表示方法，定义时间序列的局部最大点和局部最小点，并用这些点的相邻连线作为时间序列的表示。

这些取若干点的方法的问题主要是:1)需要被选中的点能够突出序列中数据的主要特征；2)选取的点的数量、范围以及特征要求比较严格。

4 各种时间序列表示的比较

在这一节，我们对典型的时间序列表示方法进行比较。

由于数据挖掘过程是一个需要用户参与的不断交互的过程，而且时间序列数据经常以流序列的形式存在，所以结合我们的项目需求，我们从以下方面进行：

(1) 是否进行时域-频域变换；
(2) 是否有效降维；
(3) 是否线性计算复杂度；
(4) 是否符号化；
(5) 能否处理变长序列；
(6) 能否动态插入/删除；
(7) 结果用户是否容易理解；
(8) 局部特征是否保持。

比较结果如表 1 所示。

表 1 时间序列表示方法比较(Y:Yes, N:No)

表示方法	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
DFT	Y	Y	N	N	Y	N	N	N
DWT	Y	Y	N	N	N	N	N	N
SVD	Y	Y	N	N	Y	N	N	N
PAA	N	Y	Y	N	Y	Y	Y	N
APCA	N	Y	N	N	Y	Y	Y	N
PLA	N	Y	N	N	Y	Y	Y	N
PRA	N	Y	Y	N	Y	Y	N	N
多项式拟合	N	Y	N	N	Y	N	N	N
Clipper Data	N	N	Y	Y	Y	Y	Y	N
SAX	N	Y	Y	Y	Y	Y	Y	N
Landmarks	N	Y	Y	N	Y	Y	Y	Y
Important point	N	Y	Y	N	Y	Y	Y	Y

5 结论

总结以上出现的时间序列不同表示方法，可以得出结论如下：1)由于实际中不同应用对于时间序列数据的关注角度不同，对于时间序列的表示的方法也不同，时间序列表示与应用相关；2)大部分时间序列的表示方法都围绕降维展开，千方百计地降维而不至于大量地丢失原有数据的信息和特征。

本文中我们对当前出现的比较典型的时间序列表示方法进行了综述和比较研究，进一步的工作是这些方法进行实际转化和改造，并应用到我们的项目中，并且对于新的时间序列表示方法，特别是时间序列流环境下的时间序列表示方法进行探索和研究。

参考文献

[1] Agrawal, R., Faloutsos, C. and Swami, A. Efficient similarity search in sequence in sequence databases. In: Proc. 4th International Conference on Foundations of Data Organization and Algorithms, Chicago, Illinois, USA, 1993. 69~84.

[2] Agrawal, R., Psaila, E. L. Wimmers and M. Zait. Querying shapes of histories. In Proc. of the 21st International Conference on Very Large Databases(VLDB), 1995. 502~514.

[3] Agrawal, R., K. I. Lin, H.S. Sawhney, et al. Fast similarity search in the presence of noise, scaling , and translation in time-series databases. In: Proc. of the 21st International Conference on Very Large Databases(VLDB), 1995. 490~500.

[4] Chakrabarti, K., Ortega-Binderberger. M. Porkaew, et al. Similar shape retrieval in MARS. In: Proc. of IEEE International Conference on Multimedia and Expo.2000.

[5] Chan, K. and Fu, A. W. Efficient time series matching by wavelets. In: Proc. 15th IEEE International Conference on Data Engineering (ICDE). Sydney, Australia, 1999.126~133.

[6] Chang-Shing Perng, Haixun Wang, and Sylvia R. Zhang, et al. Landmarks: A New Model for Similarity-Based Pattern Querying in Time Series Databases. In: Proc. of 16th International Conference of Data Engineering(ICDE), San Diego, USA.2000.

[7] Faloutsos, C., Ranganathan, M., Manolopoulos, Y. Fast subsequence matching in time-series databases. In: Proc. of ACM SIGMOD Conference, Mineapolis, 1994. 419~429.

[8] H. Shatkay, H. and S. Zdonik. Approximate queries and representations for large data sequences. In: Proc. 12th International Conference on Data Engineering(ICDE), 1996. 536~545.

[9] Li Quanzhong, Vega Lopez I.F., Moon B. Skyline Index for Time Series Data. IEEE Trans. on Knowledge and Data Mining, 16(6), 2004. 669~684.

[10] Li Wei, Kumar N., Lolla V, et al. A Practical Tool for Visualizing and Data Mining Medical Time Series. In: Computer-Based Medical Systems, 2005.

[11] Jessica Lin, Keogh E. Londardi S, et al. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms.In: Proc. of the 8th ACM SIGMOD workshop on Research(DMKD), San Diego, CA, USA. 2003.

[12] Jessica Lin, Keogh E, Londardi S. Visualizing and Discovering Non-Trivial Patterns in Large Time Series Databases.In: Information Visualization, 2005.

[13] Kamel, I. and Faloutsos, C. Hilbert R-tree: An improved R-tree using fractals. In: Proc. Very Large Database (VLDB), 1994. 500~509.

[14] Keogh E. Exact indexing of dynamic time warping. In: Proc. of 28th International Conference on Very Large Databases Conference(VLDB), Hong Kong, China.2002. 406~417.

[15] Keogh E., Kasetty, S. On the need for time series data mining benchmarks: a survey and empirical demonstration. In: 8th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002. 102~111.

[16] Keogh E. Pazzani, M. A simple dimensionality reduction technique for fast similarity search in large time series databases. In: 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Kyoto, Japan,2000. 122~133.

[17] Keogh E. Chakrabarti K, Pazzani M. Locally adaptive dimensionality reduction for indexing large time series databases. In: Proc. of ACM SIGMOD Conference on Management of Data, 2001. 151~162.

[18] Keogh E. Similarity Search in Massive Time Series Databases: [Ph.D. Thesis]. University of California, Irvine. 2002.(Importance)

[19] Keogh E. Lonardi S. Ratanamahatana A.C. Towards Parameter-Free Data Mining. In: Proc. of KDD, Seattle, WA. USA. 2004.

[20] Keogh E. Selina Chu, David Hart, et al. An Online Algorithm for Segmenting Time Series. In: Proc. of IEEE International Conference on Data Mining(ICDM), 2001.

[21] Kevin B. Pratt and Eugene Fink. Search for Patterns in Compressed Time Series. In: International Journal of Image and Graphics. 2(1), 2002. 89~106.

[22] Korn, F., Jagadish, H. and Faloutsos, C. Efficiently supporting ad hoc queries in large datasets of time sequences. In: Proc. ACM SIGMOD International Conference on Management of Data, Tuescon, AZ. USA, 1997. 289~300.
[23] Refiei, D. On similarity based queries for time series data. In: Proc. 15th IEEE International Conference on Data Engineering (ICDE), Sydney, Australia, 1999. 410~417.
[24] Yi, B, K. Faloutsos, C. Fast time sequence indexing for arbitrary LP norms. In: Proc. Of the 26th International Conference on Very Large Databases(VLDB), Cairo, Egypt. 2000.
[25] Zhu Yunyue. High Performance Data Mining in Time Series: Techniques and Case Studies: [Ph.D. Thesis], New York University. 2004.
[26] 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.电路与系统学报.5(2), 2000.9~14.
[27] 李爱国，覃征.大规模时间序列数据库降维及相似搜索.计算机学报,2005.1147~1475.
[28] 黄超，朱扬勇．基于回归系数的时间序列维约简与相似性查找.模式识别与人工智能.2006.52~57.
[29] 王元珍，李俊奎，曹忠升．RPAA：一种基于时间特性的时间序列表示..计算机科学.(录用待发表)

Progress and Comparative Study of Time Series Representations: In the Context of Time Series Mining and Modeling

LI Junkui WANG Yuanzhen LIU Chengcheng CAO Zhongsheng
Research Institute of Database & Multimedia, Huazhong University of Science & Technology, Wuhan, Hubei, 430074, China

Abstract

The representation of time series is one of basic and fundamental questions in mining time series sequences. The paper gives a detailed survey on the existing methods for representing time series in the process of mining, and conducts a comparative study in different views. The results show that, most of the representing methods try their best to approximate time series into a lower-dimensional style, and they all domain application depended. There should undergo a transformation process of the representations in the real system construction
Keywords: Data mining Time series Representation Modeling

李俊奎：男，1981 年生，博士研究生，主要研究方向是数据挖掘、机器学习。
王元珍：女，1945 年生，教授，博士生导师，主要研究方向是现代数据库理论与实现技术、数据挖掘中间件技术。
刘城成：男，1982 年生，硕士研究生，主要研究方向是时间序列建模，多媒体数据库技术。
曹忠升：男，1965 年生，副教授，主要研究方向是空间和多媒体数据库技术。