文章编号: 1001-0920(2002)05-0527-05

利用反馈的时序模式挖掘算法研究

郑斌祥, 席裕庚, 杜秀华
(上海交通大学 自动化研究所, 上海 200030)

摘 要: 针对时序数据相似性挖掘方法进行研究, 提出一种利用反馈的时序数据相似性挖掘算法, 由用户赋予各初始范围查询得到的相似序列相应的权值, 通过反馈与给定序列叠加产生新的查询序列, 再次进行范围查询, 获得相似序列。将该算法用于某钢铁企业的电力负荷时序数据, 计算结果表明了算法的有效性。

关键词: 时间序列; 相似性挖掘; 反馈; 数据挖掘; 知识发现
中图分类号: TP 311 文献标识码: A

Research on similarity mining in time series data sets

ZHENG Bin-xiang, XI Yu-geng, DU Xiu-hua
(Institute of Automation, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: A new algorithm based on feedback is proposed for the time series data similarity mining. The algorithm lets the users set weights for the time sequences obtained through the original range query. Similarity time sequences are got by using range query based on the time sequence through feedback. The algorithm is applied for similarity mining of the time series data of the electrical loads for a steel plant. The simulation results show the effectiveness of the algorithm.

Key words: time series; similarity mining; feedback; data mining; knowledge discovery

1 引 言

数据挖掘就是从大型数据库的数据中提取人们感兴趣的、隐含的和事先未知的知识[1]。在许多现实数据库中, 数据常常是按时间顺序记录的一系列观测值, 对象的属性值可能会随时间而变化。因此, 从时序数据中挖掘潜在的有用知识具有重要的理论和实践意义。

针对时序数据模式挖掘的研究, 主要集中于时序数据相似性模式挖掘。时序数据相似性模式挖掘就是在数据库中发现与给定时序数据模式很相似的时序序列。它具有广泛的实用价值, 例如: 对观测的空间数据库记录的恒星和行星的时序数据进行相似性分析, 能帮助天文学家发现新的星星; 在电力系统中帮助技术人员发现具有相似用电模式的负荷, 从而适当调整这些负荷的用电时间, 以使电力系统达到经济运行。

针对时序数据相似性模式挖掘的研究已取得一些成果。Berndt 和 Clifford[2]提出一种动态时间弯曲技术, 允许沿时间轴进行伸缩变换, 获得与参考时序模式匹配的时序序列; Faloustson[3]提出了时间窗口的概念, 将时间序列分解为一系列子序列, 再由子序列抽取特征量进行相似性匹配; Agrawal[4]研究了时序数据存在偏离和噪声等情况时的时序数据

收稿日期: 2001-04-10; 修回日期: 2001-06-22

作者简介: 郑斌祥(1973—), 男, 江西余江人, 博士生, 从事数据仓库和数据挖掘的研究; 席裕庚(1946—), 男, 上海人, 教授, 博士生导师, 从事复杂系统控制理论和智能机器人等研究。

匹配问题。

然而,实际用户往往开始时并不很明确所要查询的时序序列,当由时序数据相似性模式挖掘算法获得的时序序列不能满足用户要求时,目前的时序数据相似性模式挖掘算法缺乏有效手段进一步获得使用户满意的结果,而且在上述算法中,用户难以有效地参与时序数据相似性模式的挖掘过程。为此,本文提出一种利用反馈思想的时序数据相似性挖掘算法,并用其对某钢铁企业电力负荷时序数据进行了相似性挖掘,计算结果表明本文算法是有效的。

2 利用反馈的时序数据相似性挖掘

时序数据相似性挖掘就是在数据库中发现与给定时序序列的模式很相似的序列。在进行序列相似性挖掘之前给定一个相似性评价函数和一个阈值 ϵ,如果函数值小于等于 ϵ,则表明序列相似。通常用 X 与 Y 之间的距离函数 $D(X,Y)$ 作为序列 X 与 Y 的相似性判别函数。距离函数 $D(X,Y)$ 常用 X 与 Y 之间的欧几里德空间距离等来代替[2~6],如果计算结果小于等于给定的阈值 ϵ,则表明 X 与 Y 相似。一般情况下,数据库中的时序序列都很长,因而计算距离需要较长的时间。如果能从序列中抽取少量主要特征则可以大大提高序列的查找速度,因此将时序数据的相似性挖掘分为以下两个步骤:

1) 依据应用要求将时序数据划分为一系列子序列,用离散傅立叶变换将子序列时序数据从时域空间变换到频域空间。根据 Parseval 的理论,时域能量函数与频域能量函数相同,且频域空间的大部分能量集中在前几个系数上,因此可以考虑只选用傅立叶变换得到的前 k 个系数。将这些系数看作从时间序列上提取的特征,于是从每个序列获得 k 个特征。进一步将它们作为 k 维空间上的一个映射,即将时序数据的子序列映射为 k 维空间上的点。这样便保留了时序数据的主要特征,而且降低了时序数据的维数,减小了计算的复杂性。

2) 针对 k 维空间上的时序数据点,本文采用多维索引方法 R 树来存储这些多维空间的点。通过范围查询检索与给定序列相似的时序序列,将检索的相似时序序列展现给用户,由用户赋予其感兴趣的序列相应的权值,并通过反馈与给定序列叠加产生新的查询序列,再次进行范围查询,获得相似序列。无反馈与利用反馈的时序数据相似性挖掘过程相比,无反馈的时序数据相似性挖掘由范围查询获得满足相似性评价函数和阈值 ϵ 的相似时序序列后,时序数据相似性挖掘便结束了。若用户对所获得的结果不满意,算法无法为用户提供有效手段进一步获得满意的结果。而利用反馈思想的时序数据相似性挖掘算法则为用户提供了有效参与挖掘过程的有效手段,用户由范围查询获得满足相似性评价函数和阈值 ϵ 的相似时序序列后,可根据经验和应用要求对所获得的感兴趣的各序列赋予相应的权值,并通过反馈与原给定序列叠加获得新的查询序列(该查询序列经用户参与的反馈修正后更真实地反映了用户的意愿),再通过范围查询,直到获得令用户满意的结果。

2.1 时序数据的离散傅立叶变换(DFT)

给定一个时间序列 $\vec{X}=\{x_t \mid t=0,1,\cdots,N-1\}$,序列 $\vec{X}$ 划分为子序列 $\vec{X}_1,\cdots,\vec{X}_m$,其中 $\vec{X}_i=\{x_{ij} \mid j=0,1,\cdots,n-1\}$,$N=nm$。

对 $\vec{X}$ 的子序列 $\vec{X}_i$ 进行离散傅立叶变换,得

$$x_{if}=\frac{1}{n}\sum_{t=0}^{n-1}x_{it}\exp(-i*2\pi f t/n)$$

$$f=0,1,\cdots,n-1,\quad i=1,2,\cdots,m \tag{1}$$

其中,$\vec{X}_i$ 代表时域信息,x_{if} 为子序列 $\vec{X}_i$ 离散傅立叶变换的傅立叶系数。

$$\vec{X}_{iF}=\{x_{if} \mid f=0,1,\cdots,n-1\} \tag{2}$$

根据 Parseval 理论,时域能量谱函数与频域能量谱函数相同。

$$\|\vec{X}_i-\vec{Y}_i\|^2=\|\vec{X}_{iF}-\vec{Y}_{iF}\|^2 \tag{3}$$

即

$$D(\vec{X}_i,\vec{Y}_i)=(E(\vec{X}_i-\vec{Y}_i))^{1/2}=(E(\vec{X}_{iF}-\vec{Y}_{iF}))^{1/2}=D(\vec{X}_{iF},\vec{Y}_{iF}) \tag{4}$$

经离散傅立叶变换,时序数据由时域空间变换到频域空间,其距离保持不变。因此可考虑只选用傅立叶变换得到的前 k 个系数,把 k 个系数看作从时间序列上提取的特征,将时序数据的子序列映射为 k 维空间上的点。

2.2 定 义

对映射到 k 维空间上的时序点,本文用范围查询获得给定序列的相似序列。

定义 1 对于有 N 个数据点的集合 Ω,任取 $p\in\Omega$,且记 $D(p,x)$ 为点 p 与 $x\in\Omega$ 的点之间的距离,对于给定大于零的 ϵ,若 $D(p,x)<\epsilon$,则称 x 为点 p 的 ϵ 范围内的相似数据点。

定义 **2**　以欧几里德距离作为距离函数, 用 $\mathrm{dist}(p_i, p_j)$ 表示 δ 维空间的点 p_i 和 p_j 之间的距离, δ 维空间的点 p 记为 $[p_1, p_2, \cdots, p_\delta]$。

$$\mathrm{dist}^{\delta}(p_i, p_j) = \left(\sum_{k=1}^{\delta}(p_{ik} - p_{jk})^2\right)^{1/2} \quad (5)$$

2.3　算　法

利用反馈思想的时序数据相似性挖掘算法的思想是, 采用多维索引方法 R 树通过范围查询检索与给定序列相似的时序序列, 将得到的相似时序序列展现给用户, 由用户赋予各序列相应的权值, 并通过反馈将各序列按相应的权重与给定序列叠加产生新的查询序列, 再次进行范围查询, 获得满足用户要求的相似序列。算法主要包括以下两部分:

1) 计算反馈时序数据点;

2) 计算经过反馈得到的相似时序数据点。

算法 **1**　计算经过反馈得到的相似时序点 computSimSeq(Root, P, r, ϵ)

输入: 待查询的时序点集合 P, 给定时序序列点 r, ϵ;

输出: 相似时序点集合 SimSet。

Step1: 设置集合 SimSet 和 SimSettmp 初始值为空;

Step2: 将 P 中的每个点 p 和 MBR 插入 R-Tree, 建立一链表 nodelist;

Step3: 当链表非空时, 若结点为叶子结点, 对于结点中的每个点 p;

Step4: 计算点 p 与点 r 的距离 Dist(p, r), 若 Dist(p, r) < ϵ, 则将点 p 插入相似时序点集合 SimSettmp;

Step5: 对相似时序点集合 SimSettmp, 调用算法 computFeedbackSeq(SimSettmp, W, r) 计算由相似时序点集合中的点经用户赋予相应权重后与 r 叠加获得的时序点 C;

Step6: 对于 P 中的每个点 p, 计算点 p 与点 C 的距离 Dist(p, r), 若 Dist(p, r) < ϵ, 则将点 p 插入相似时序点集合 SimSet;

Step7: 返回 SimSet。

算法 computSimSeq(Root, P, r, ϵ) 通过对 R 树的结点计算时序序列点 r 的相似时序点集合, 建立一链表 nodelist, 当结点是叶子结点时, 计算叶子结点中存储的点 r 与点 p 的距离, 当点 r 与点 p 的距离小于 ϵ 时, 将点 p 插入相似时序点集合 SimSettmp。对由初始范围查询检索的相似时序点, 调用算法 computFeedbackSeq(SimSettmp, W, r) 计算由相似时序点集合中的点经用户赋予相应权重后与 r 叠加获得的时序点 C。针对新的参考时序点 C, 再次进行范围查询找到点 r 的相似时序点, 返回相似时序数据。

算法 **2**　计算反馈时序点 C, computFeedbackSeq(SimSettmp, W, r)

输入: 由初始范围查询获得的相似时序点集合 SimSettmp, 用户设定的权值 W, 给定时序序列点 r;

输出: 由反馈获得的时序点 C。

Step1: 对由初始范围查询获得的相似时序点集合的点进行傅立叶反变换, 得到时序数据曲线, 由用户根据对各时序数据曲线的感兴趣程度赋予各相似时序点集合的点相应的权值;

Step2: 对于相似时序点集合 SimSettmp 中的点 A 和参考时序点 r, 对应的权重为 WA 和 WR, 计算反馈后的时序点 C 为

$$C = (A * WA + B * WR)/(WA + WR)$$

Step3: 返回时序点 C。

算法 2 由算法 computSimSeq(Root, P, r, ϵ) 调用, 计算经反馈获得的时序点 C。对应于初始范围查询得到的相似时序点和参考时序点 r 的权重由用户根据经验和应用要求设定。

3　仿真算例

运用本文算法, 对某钢铁企业的电力负荷时序数据库挖掘其中的相似时序序列。我们以负荷的功率数据(取 1999 年 4 月的数据) 为例, 其时序曲线如图 1 所示。

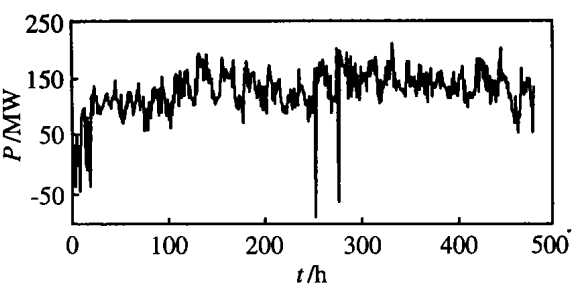
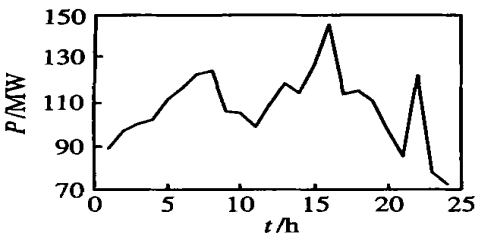
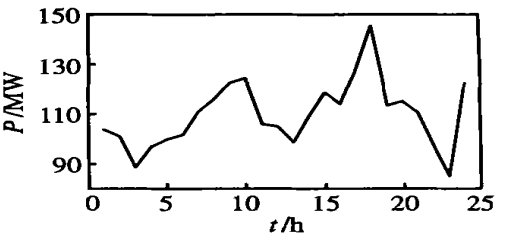


图 **1**　电力负荷功率曲线

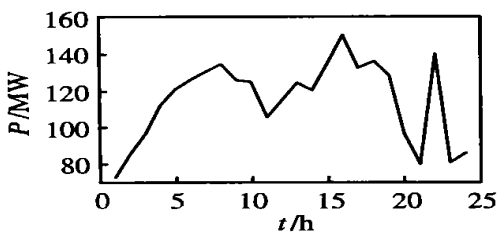
对上述时序数据, 利用反馈的时序数据相似性挖掘算法, 给定参考查询序列, 对时序子序列和参考查询序列用 DFT 变换到频域多维空间上的点, 由范围查询从时序数据库中获得与参考序列相似的序列, 并将经过傅立叶反变换的相似时序数据曲线展现给用户。限于篇幅, 本文仅选出其中 3 条最为相似的时序序列, 如图 2 所示。



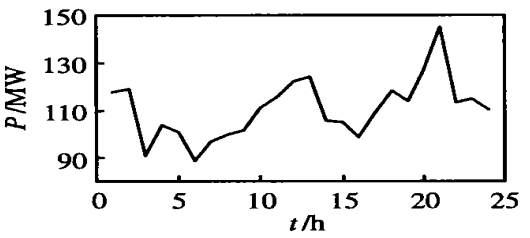
(a) 参考序列



(b) 相似序列 1

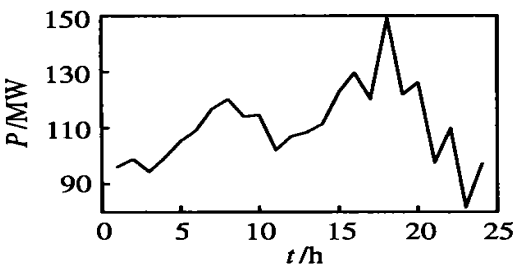


(c) 相似序列 2

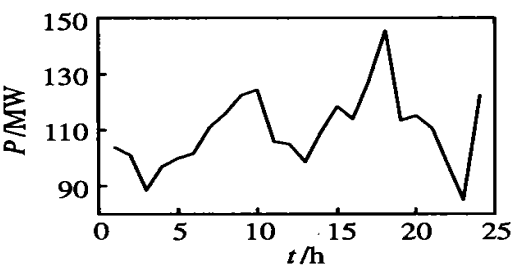


(d) 相似序列 3

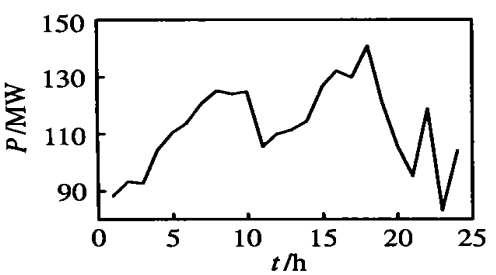
图 2 初始范围查询获得的相似时序曲线



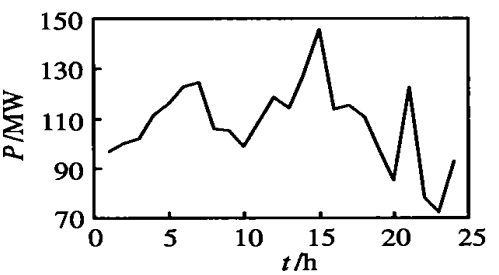
(a) 经反馈叠加后的时序序列



(b) 相似序列 1



(c) 相似序列 2



(d) 相似序列 3

图 3 利用反馈获得的相似时序曲线

图 2(a) 为参考序列数据曲线, 表示电力负荷在某时间段的用电模式, 其余为与参考序列相似的时间序列。用户发现得到的相似序列中(图 2(b)) 晚上 22∶00 左右负荷电力消耗有上升趋势。用户希望发现更多的具有这种趋势的时序序列。利用反馈的算法使得用户可以对各相似序列赋予相应的权值, 由于用户对图 2(b) 相似序列兴趣最大, 其权值也最大, 权值选为[- 2, 2] 之间, 再次进行相似性挖掘, 得到的相似序列如图 3 所示。

图 3(a) 为经过反馈叠加后的时序序列, 图 3(b) ～ 图 3(d) 为获得的满足用户需求的相似序列, 即在晚上 22∶00 左右负荷电力消耗有上升趋势的相似时序序列。通过对生产数据库进行查询, 发现在这些时间区间内均有生产任务。因此用户可相应地调整电力系统用电负荷, 使电力系统经济运行。

4 结 语

时序数据相似性挖掘是数据挖掘中的重要研究内容。本文提出一种利用反馈的时序数据相似性挖

掘算法, 为用户进一步获得满意的结果提供了有效手段, 使用户能有效地参与和干预时序数据相似性模式的挖掘过程。我们将算法用于某钢铁企业的电力负荷时序数据, 计算结果表明了算法的有效性。

参考文献(References):

[1] R Agrawal, M Mehta, J Shafer, et al. The QUEST data mining system[A]. *Proc of Int Conf on Data Mining and Knowledge Discovery* (*KDD'96*)[C]. Oregon, 1996. 244-249.

[2] D J Berndt, J Cliffod. Finding patterns in time series: A dynamic programming approach[A]. *Advances in Knowledge Discovery and Data Mining*[C]. Menlo Park: AAAI Press, 1996. 229-248.

[3] C Faloutsos, M Ranganathan, Y Manolopoulos. Fast subsequence matching in time-series databases[A]. *Proc of ACM SIGMOD Conf on Management of Data* (*SIGMOD'94*)[C]. Minneapolis: ACM Press, 1994. 419-429.

[4] R Agrawal, Lin K I, Sawhony Shim K, et al. Fast similarity search in the presence of noise, scaling and translation in time series databases[A]. *Proc of 21st Int Conf on Very Large Data Bases*[C]. Zurich, 1995. 490-501.

[5] N Roussopoulos, S Kelley, F Vincent. Nearest neighbour queries[A]. *Proc of ACM SIGMOD*[C]. San Jose, 1995. 71-79.

[6] R Agrawal, C Faloutsos, A Swarni. Efficient similarity search in sequence database[A]. 4th *Int Conf on Foundations of Data Organization and Algorithms*[C]. Evanston, 1993. 69-84.

[7] D Rafiei, A Mendelzon. Similarity-based queries for time series data[A]. *Proc of ACM SIGMOD Conf on Management of Data* (*SIGMOD'97*)[C]. Arizona: ACM Press, 1997. 13-25.

(上接第 516 页)

[22] S Lloyd. Coherent quantum feedback[J]. *Physical Review A*, 2000, 62(9): 2108-2114.

[23] A Doherty, J Doyle, H Mabuchi, et al. Robust control on the quantum domain[A]. *Proc 39th IEEE CDC*[C]. Sydney, 2000. 1: 949-954.

[24] M Yanaqisawa, H Kimura. Transfer function approach to quantum control systems[A]. *Proc 40th IEEE CDC*[C]. Orlando, 2001. 1595-1600.

[25] D D Alessandro, M Dahleh. Optimal control of two-level quantum systems[J]. *IEEE Trans on Automatic Control*, 2001, 46(6): 866-876.

[26] R W Brochett. System theory on group manifolds and coset spaces[J]. *SIAM J Control*, 1972, 10(2): 265-284.

[27] D Cheng, W P Dayawansa, C F Martin. Observability of systems on Lie groups and coset spaces[J]. *SIAM J Control*, 1990, 28(3): 570-581.

[28] N Khaneja, R Brockett. Time optimal control in spin systems[J]. *Physical Review A*, 2001, 63(3): 2308.

[29] L Liola, S Lloyd. Dynamical generation of noiseless quantum subsystems[J]. *Physical Review Letters*, 2000, 85(16): 3520-3524.

[30] Ting Yu. Decoherence and localization in quantum two-level systems[J]. *Physica A*, 1998, 248(3): 393-418.

[31] L Viola, S Lloyd. Dynamical suppression of decoherence in two-state quantum systems[J]. *Physical Review A*, 1998, 58(4): 2733-2744.

[32] G Alber. *Quantum Information: An Introduction to Basic Theoretical Concepts and Experiments*[M]. New York: Springer, 2001.

[33] M Hirvensalo. *Quantum Computing*[M]. New York: Springer, 2001.

[34] I Pastirk, E J Brown, Q Zhang, et al. Quantum control of the yield of a chemical reaction[J]. *J Chemical Physics*, 1998, 108(11): 4375-4378.