时间序列中快速模式发现算法的研究

黄 河[1] 黄 轲[2] 杭小树[1] 熊范纶[1]
[1](中国科学院合肥智能所,合肥 230031)
[2](宜春大学生物化学系,江西宜春 336000)
E-mail:huanghe@mail.iim.ac.cn

摘 要 针对长时间序列,该文提出了一种新的能快速发现序列中时序模式的检索方法。首先将时间序列分成若干等长的子序列;接着从每个子序列中提取特征序列,该特征序列能够反映子序列中数据的变化趋势;然后根据每个特征序列将相应的子序列分配到一系列盒子中,使得不同盒子中的子序列因数据变化趋势不同而不相似,而在同一盒子中的序列由于数据变化趋势相同而有可能相似;最后通过计算每个盒子中任意两个子序列间的欧几里德距离来发现所有的模式。有关实验证明该算法是行之有效的。

关键词 时间序列 时序模式 特征序列 欧几里德距离

文章编号 1002-8331-(2003)21-0192-03 **文献标识码** A **中图分类号** TP311

Algorithm for Fast Time-Series Patterns Recovery in A Long Sequence

Huang He[1] Huang Ke[2] Hang Xiaoshu[1] Xiong Fanlun[1]
[1](Institute of Intelligent Machine, Chinese Academy of Sciences, Hefei 230031)
[2](Department of Biochemics, College of Medicine, Yichun University, Yichun, Jiangxi 336000)

Abstract: In this paper a fast algorithm is presented for recovering time-series patterns in a long sequence. First, the sequence is segmented into same-length subsequences. Then a feature series is extracted from each subsequence to show its changing properties. Third, all the subsequences are distributed into a set of boxes according to their feature series, so that those in a box are possibly similar, while those in different boxes are impossibly similar. Finally the algorithm discovers all time-series patterns by computing Euclidean distance between any two subsequences in each box. The experiment results prove that it can be put into practice and work very efficiently.

Keywords: Time sequence, time-series patterns, feature series, Euclidean distance

1 引言

时间序列是一组有序的随着时间改变的序列值或事件,而时序数据库是由时间序列组成的数据库。时序数据库广泛运用于各种领域,例如:科学实验的数据分析、股票市场的波动分析等等。为了进行预测,首先必须建立一个适当的预测模型。因此,如何从时序数据中挖掘出时序模式,就成为一个重要的研究课题。

目前,国际上有关时间序列的研究已取得了一些成果。在文献[1]中,R.Agrawal 和 C.Faloutsos 利用离散傅利叶变换(DFT)将时间序列由时域映射到频域,再用 R^* 树结构对模式进行匹配,但这种方法仅适用于整体序列匹配。在文献[2]中,C.Faloutsos 和 M.Ranganathan 第一次提出了时间窗口(sliding window)的概念,该算法适用于子序列匹配。有些研究[3,4]采用时间弯曲距离(time warping distance)和编辑距离(edit distance)作为评价两序列是否相似的标准,该技术适用于比较不等长分段间的距离,但时间复杂度是时间序列长度的平方。

尽管上述算法在其各自的应用领域内取得了很大的成功,但笔者认为它们都有一个主要的缺陷,即它们均从一个给定的模式着手,在匹配过程中调整各类参数以使得时间序列与给定的模式相匹配,无法或不适合于动态地挖掘出所有的模式集合。文献[5]提出了一种时序模式发现算法,该方法通过分段线性表示法,将时序曲线拟和为线段序列,从而以相对应的线段的斜率反正切值作为模式的逻辑表示。该算法能自动地发现所有(子)模式,但所花费的时间与序列长度成平方关系。该文提出了一种新的检索方法,能大大提高算法的效率。

文章安排如下:第 2 节具体描述了时序模式发现算法;第 3 节分析了算法的时间复杂度;第 4 节给出了实验结果;第 5 节是结论。

2 时序模式发现算法

为了分析长时间序列,首先将它分成若干个等长的子序列,并从每个子序列中提取特征序列;接着将具有相同特征序列的子序列分配到同一盒子里,这样,就得到了一系列盒子,其中每个盒子中至少包含一个子序列,且使得同一盒子中的子序

基金项目:国家自然科学基金重点项目(编号:69835001);国家 863 高技术研究发展计划(编号:2001AA110464)资助
作者简介:黄河(1977-),女,硕士研究生,主要研究方向:数据挖掘。

列可能相似而不同盒子中的子序列不相似。最后，通过计算每个盒子里子序列间的欧几里德距离来发现时间序列中的所有时序模式。

2.1 序列的分段

给定一个时间序列 $L=\{(x_1,y_2),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中 x_i 表示时间，y_i 表示数值(实数)，n 表示序列的长度，根据先验知识将 L 分成 s 段，即分成 s 个长度为 m 子序列 $L_i^s=\{(x_{(i-1)\cdot m+1},y_{(i-1)\cdot m+1}),\cdots,(x_{i\cdot m},y_{i\cdot m})\}$ $i=1,2,\cdots,s(n=s*m)$。例如，L 表示 10 年间某商品的日销售量，可以根据用户的要求将序列按周或月分段。

2.2 特征序列的提取

给定一个子序列 L_i^s，它的特征序列 F_i^s 由子序列中的极值点组成。此处，极值点是指子序列的局部最小值点或局部最大值点，它们能简单、快速地反映出子序列曲线的起伏变化。

令 $F_i^s=\{(x_{i_1},y_{i_1}),\cdots,(x_{i_s},y_{i_s})\}$，其中 $y_{i_j-1}<y_{i_j}<y_{i_j+1}$ $j=1,2\cdots,s$。

提取了特征序列之后，采用文献[6]中的方法来平滑特征序列。该方法描述如下：给定特征序列 $F_i^s=\{(x_{i_1},y_{i_1}),\cdots,(x_{i_s},y_{i_s})\}$，最小时间间隔 θ_{time} 和最小数值变化率 θ_{value}，对于极值点 (x_{i_j},y_{i_j}) 和 $(x_{i_{j+1}},y_{i_{j+1}})$，若它们满足条件 $x_{i_{j+1}}-x_{i_j}<\theta_{time}$ 和 $\frac{|y_{i_{j+1}}-y_{i_j}|}{\frac{|y_{i_{j+1}}|+|y_{i_j}|}{2}}<\theta_{value}$，则从特征序列中去除这两个数值点。如图 1 所示。

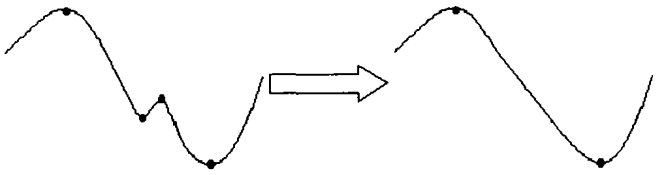

图 1 平滑特征序列

与滑动平均相比，这种方法保留了那些能描述子序列曲线起伏特点的特殊点。而滑动平均则趋向于平滑掉序列曲线的波峰和波谷，并且丢失了序列的开始点和结束点。

2.3 子序列的分配

给定子序列 L_i^s 及其特征序列 F_i^s，在 F_i^s 中每个点处用垂直线将 L_i^s 分隔，如图 2 所示：

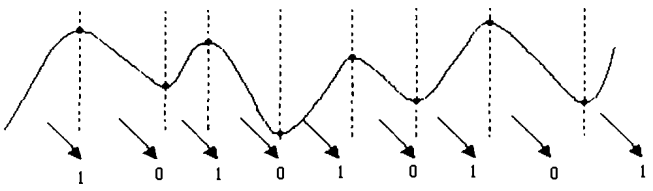

图 2 用 0、1 字符串表示子序列

对于分割后的每一条子曲线，若上升则用“1”表示，若下降则用“0”表示。因此，每一个子序列可以用一个 0、1 字符串来表示曲线的起伏状况。将具有相同字符串的子序列放入同一个盒子，且该盒子用这个字符串标识。这样就得到一个盒子集 $B=\{B_1,B_2,\cdots,B_b\}$，其中每个盒子 B_i 至少装有一个子序列，且同一盒子中的子序列有相似的变化曲线。显然，不同盒子中的曲线是不相似的。因此，仅需要比较同一盒子中的子序列是否相似，从而大大提高了算法的执行效率。

2.4 相似性的计算

既然子序列的长度相等，一个简单的方法就是将子序列看成 m-维空间中的点，再采用欧几里德距离来比较子序列间的相似性。给定两个子序列 L_i^s and L_j^s，它们之间的距离定义为：$d(L_i^s,L_j^s)=(\sum_{k=1}^{m}(y_{i_k}-y_{j_k})^2)^{1/2}$。笔者认为：两个子序列曲线尽管它们的基线或振幅不同，但若具有相似的变化趋势，这两个子序列仍是相似的。采用如下方法：首先将子序列 L_i^s 的所有值标准化，用 $\lambda(L_i^s)$ 表示 L_i^s 的标准化子序列；然后再计算标准化子序列间的欧几里德距离，即给定 $\lambda(L_i^s)$ 和 $\lambda(L_j^s)$，$d(\lambda(L_i^s),\lambda(L_j^s))=(\sum_{k=1}^{m}(\lambda(y_{i_k})-\lambda(y_{j_k}))^2)^{\frac{1}{2}}$。所采用的标准化方法是 $\lambda(y_{i_k})=(y_{i_k}-Ey_i)/Dy_i$，其中 Ey_i 是 L_i^s 的均值，Dy_i 是 L_i^s 的标准方差。这样可以使得标准子序列 $\lambda(L_i^s)$ 的曲线落在-1 与 1 之间，且序列的均值和方差分别为 0 和 1。显然，计算 $d(\lambda(L_i^s),\lambda(L_j^s))$ 的时间复杂度为 $O(m)$。

2.5 时序模式的发现

时序模式的发现算法描述如下：

输入：一个长时间序列 $L=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，子序列长度 m，最小时间间隔 θ_{time}，最小数值变化率 θ_{value}，距离允许误差 DT。

输出：时序模式集。

```
PatternRecovery(L,m,θtime,θvalue,DT){
将 L 分成 s 个长度为 m 的子序列 Li^s  i=1,2,…,s;
对于每个子序列 Li^s{
  从 Li^s 中提取特征序列 Fi^s;
  平滑 Fi^s;
}
将 s 个子序列分配到一系列盒子 B={B1,B2,…,Bb}中;
ResultSet=ϕ;//设置结果集合 ResultSet 为空;
对于每个盒子 Bi(i=1,2,…,b){
  Bi 中任意两个子序列 Lj^s 和 Lk^s{
    If d(λ(Lj^s),λ(Lk^s))<DT then ResultSet=ResultSet∪{Lj^s,Lk^s};
  }
}
Join(ResultSet);//将结果集合中有公共元素的子序列对合并
Output(ResultSet);
}
```

总的序列模式即为结果集合中子序列模式之和。

3 算法的分析

令 n 为序列的总长度，m 为子序列的长度，则有 $s=n/m$ 的子序列。时间序列的分段，特征序列提取和子序列分配需花费

的时间为 $O(n)$。时序模式发现部分的时间复杂度与子序列的分配结果有关。首先考虑最好情况：若每个盒子里仅有一个子序列，即每个子序列都是一个时序模式，则该部分的时间花费为 $O(n/m)$；接着考虑最差情况：若仅有一个盒子，即所有子序列都分配在一个盒子里，则可以组成 $s\cdot(s-1)/2$ 个子序列对，计算每个子序列对的欧几里德距离所需的时间为 $O(m)$，总的时间复杂度为 $O(n(n-m)/2m)$；最后考虑一般情况：若所有子序列平均地分配到 b 个盒子里，那么每个盒子中有 s/b 个子序列，可以组成 $s(s-b)/2b^2$ 个子序列对，总的时间为 $O(n(n-bm)/2b^2m)$。与模式发现部分相比，分段、特征序列提取和子序列分配所花费的时间可以忽略不计。

4 实验结果

实验的主要目的是：(1)通过与文献[5]的方法进行计较来说明本算法的优越性；(2) 对一个现实的数据库应用上述算法来说明该算法的有效性。

4.1 与文献[5]中方法的比较

由随机数组成的时间序列分别应用文献[5]中的方法和该文方法进行模式发现。图 3 显示了对于不同长度的时间序列进行时序模式发现时，两种方法所需的时间。可以看出，该文方法所花费的时间与序列的长度几乎成线性关系，而文献[5]中的方法所花费的时间与序列的长度成平方关系。

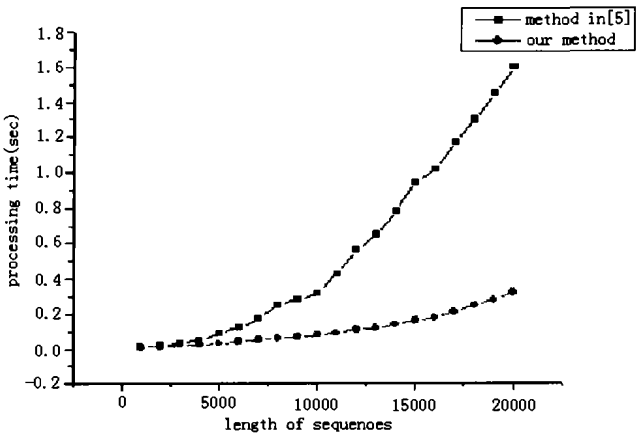

图 3 对不同长度的时间序列两种方法所需的时间

4.2 算法的有效性

文章以安徽某县的温度数据库为例，该数据库是由 1980 年到 1988 年期间，从 4 月到 10 月每天的平均温度组成。为了方便起见，这里丢弃 5、7、8 和 10 月的最后一天的平均温度，并对序列按月进行分段。令 $\theta_{time}=5$，$\theta_{value}=0.1$，$m=30$，$DT=3$，得到表 1 中的结果：同一行中显示的月份表明，在这些月中每天的平均温度变化曲线相似。例如，在第 2 行中，1980 年 7 月、1981 年 9 月、1982 年 9 月、1983 年 8 月、1984 年 5 月和 8 月的每天平均温度变化曲线相似，它们属于同一个模式。

表 1 算法的执行结果

1980	1981	1982	1983	1984	1985	1986	1987	1988
7	9	9	8	5,8				
8	5							
9	8,10	5	4					
		7,8	5					
		10	9	6				
						6,7	6,8	7
						8	9	5,6
						9	7	
							5	10

5 结论

如何从序列中快速挖掘时序模式，建立适当的检索结构是极其重要的。笔者认为两段时序数据尽管它们的基线或振幅不同，但若具有相似的变化趋势，这两段数据仍是相似的。因此，提出了一种新的检索方法能快速发现序列中的时序模式。有关实验证明该算法是行之有效的。(收稿日期：2002 年 11 月)

参考文献

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

2.C Faloutsos,M Ranganathan,Y Mandoponlos.Fast subsequence matching in time-series databases[C].In:Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data,Minneapolis,Minnesota,1994:419~429

3.Tolga Bozkaya,Nasser Yazdani,Meral zsoyoglu.Matching and Indexing Sequences of Different Lengths[C].In:Proceedings of the Sixth International Conference on Information and Knowledge Management,Las Vegas,Nevada,1997:128~135

4.B-K Yi,H Jagadish,C Faloutsos.Efficient retrieval of similar time sequences under time warping[C].In:Proceedings of the Fourteenth International Conference on Data Engineering,Orlando,Florida,USA,1998:201~208

5.蔡智，岳丽华，王熙法.时序模式发现算法研究[J].计算机研究与发展，2000;37(9):1107~1112

6.Chang Shing Perng,Haixun Wang,Sylvia R Zhang et al.Landmarks:a new model for similarity-based pattern querying in time series database.In Proceedings of the 16th International Conference on Data Engineering,San Diego,California,USA,2000:33~42

7.黄河，熊范纶，杭小树.时序数据库中快速相似搜索的算法研究[J].模式识别与人工智能，2002;15(4)

(上接 158 页)

1.Andrew S Tanenbaum.Computer Networks[M].Third Edition,ISBN 7-302-03035/TP.1618 Prentice Hall International,Inc,2000:116~122

2.Scott Fluhrer,Itsik Mantin,Adi Shamir.Weaknesses in the Key Scheduling Algorithm of RC4.URL:http://www.drizzle.com/~aboba/IEEE/rc4_ksaproc.pdf,2001-07-25

3.Kristin Burke.Wireless Network Security 802.11/802.1x".URL:http://www.cs.fsu.edu/~yasinsac/wns02/19b.pdf,2002-05:31

4.Michel MONLY,Marie-Bernadette PAUTET.The GSM System for Mobile Communications[M].ISBN7-5053-3634-7/TP.1499,电子工业出版社，1996:259~271

5.RFC 2196.Site Security Handbook[S].URL:http://www.ietf.org/rfc/rfc2196.txt

6.Jim Thompson,CTO,Musenki.802.11 Roaming and Shared Use Access Points,URL:http://www.eyeforwireless.com/docs/musenki.pdf,2001-11