DOI:10.7544/issn1000-1239.2014.20121055
51(8):1802-1810,2014

不确定时间序列的相似性匹配问题

吴红花[1] 刘国华[1,2] 王 伟[1]
[1](东华大学计算机科学与技术学院 上海 201620)
[2](计算机软件新技术国家重点实验室(南京大学) 南京 210093)
(whhua0707@163.com)

Similarity Matching for Uncertain Time Series

Wu Honghua[1], Liu Guohua[1,2], and Wang Wei[1]
[1](*College of Computer Science and Technology, Donghua University, Shanghai* 201620)
[2](*State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing* 210093)

Abstract Similarity matching techniques for certain time series do not consider the uncertainty of data, but in the real world the time series data collected by the sensors is often not certain, To solve this problem, we perform pre-processing over uncertain time series. It is divided into horizontal and vertical dimensions, that is, time dimension and probability dimension. First, an uncertain time series is compressed by the Haar wavelet transform. On this basis, we process the obtained uncertain time series longitudinally, and put forward a kind of method of electing representatives, which adopts maximum probability method and the mean method to select a certain time sequence. After pretreatment, we carry on the dimensionality reduction and indexing with generated certain time series. According to the query sequence and each time series in the database in the combination of uncertainty, we put forward the similarity matching algorithm corresponding to a combination of them respectively.

Key words time series; uncertainty; matching; dimensionality reduction; Haar wavelet transform

摘 要 确定性时间序列的相似性匹配方法都没有考虑数据的不确定性,而现实世界中传感器采集到的数据往往是不确定的,现有的时间序列的相似性匹配方法不适用于这些领域.针对此问题,将不确定性时间序列做预处理,把它分为横向时间维和纵向概率维,首先把给定的不确定时间序列用 Haar 小波变换进行压缩变换,在此基础上,对得到的不确定性时间序列概率维作纵向处理,提出一种选代表方法,即采用概率最大法、均值法等选出一条确定的时间序列.通过这 2 种预处理后,对得到的确定性时间序列进行降维和索引,根据查询序列和数据库中的时间序列中的各自的不确定性进行组合,分别提出对应组合的相似性匹配算法.

关键词 时间序列;不确定性;匹配;降维;Haar 小波变换

中图法分类号 TP311.13

时间序列(time series)是按时间顺序排列的实数序列,它反映了实体属性在时间顺序上的特征.时间序列匹配是对现实世界中大量与时间相关数据进行挖掘的基础,其中时间序列的相似性匹配问题是一个备受关注的问题,目前研究成果主要集中在确定性时间序列上.但现实世界中时间序列不一定是

收稿日期:2012-11-20;**修回日期**:2013-12-16
基金项目:国家自然科学基金项目(61070032);计算机软件新技术国家重点实验室(南京大学)开放课题(KFKT2010B06)

确定的,如在位置定位服务(location based service, LBS)系统、环境监测、钢厂轧钢、物联网等领域中,时间序列数据均存在不确定性.近年来不确定性数据管理逐渐成为数据库领域的研究热点[1-8],但关于不确定性时间序列的研究才刚刚起步,有关不确定性时间序列的相似性匹配问题还未见有效的解决方法.

降维是解决时间序列相似性匹配问题的一个重要手段[9-14].文献[9-10]中提出了基于离散傅里叶变换(DFT)的匹配算法,它为时间序列相似性匹配问题的研究奠定了基础.Chan 等人[11-12]首先提出了离散小波分析方法,它同时在时域和频域上反映了信号的特性,而且具有良好的局部化性质,且得到了比离散傅里叶变换更好的结果.文献[10-13]主要是采用 Haar 小波变换对确定性时间序列进行相似性搜索.文献[15]提出了一维和二维 Haar 小波方法对关系表进行压缩变换,最后进行近似查询处理.这些文献并未考虑时间序列中数据的不确定性,也没有讨论不确定时间序列的相似性匹配问题.

本文主要研究了不确定性时间序列的相似性匹配问题.先将不确定性时间序列的每一时间点的取值进行建模构成概率模型,把不确定性时间序列分为横向和纵向 2 个维度,然后对不确定性时间序列进行预处理操作,最后,将得到压缩后的不确定性时间序列规约为一个确定性时间序列,再用确定性时间序列相似性匹配方法进行匹配.

1 相关工作

1.1 确定性时间序列的相似性匹配

时间序列相似性匹配问题最早是由 Agrawal 等人[9]在 1993 年提出的,这一问题被描述为给定一个查询的时间序列,从大规模时间序列数据库中找出与之相似的序列,"相似"是基于距离函数来衡量的,2 个时间序列之间的距离是确定的.时间序列相似性匹配分为完全匹配和子序列匹配.该文献研究了完全匹配的查询,先用离散傅里叶变换(DFT)将原始序列变换到特征空间,实现对时间序列的降维;并提出一种 F 索引,即对特征空间中的点进行索引.文献[7-9,16-17]主要研究关于概率数据库的相似性查询.文献[14]研究了子序列匹配的查询,也是首先采用 DFT 对时间序列进行降维,然后用 R*-tree[18]和最小边界矩形(minimum bounding rectangle, MBR)对特征空间中的点进行索引,最后对从索引文件中查出的候选结果,再用相应的原始数据进行后处理,得到最终查询结果.Keogh 等人[19]通过对确定性时间序列逐段线性化,进行相似性搜索.

通过对确定时间序列相似性的研究,人们认识到降维处理是时间序列相似性匹配的关键步骤[14].长度为 n 的时间序列通常被映射为 n 维空间的一个点,利用 R-tree 等建立索引,提高查询效率.但是若 n 比较大,容易导致维度灾难(curse of dimensionality)问题,索引效率急剧下降.通过提取时间序列的主要特征,将时间序列从高维空间变换到低维特征空间,降低数据维数.在此基础上,将这些降维之后的 DFT 系数或小波系数组织为多维索引结构 R*-tree 等索引.

以上的相关工作主要是确定性时间序列相似性匹配的情况,随着不确定时间序列的出现,若采用欧式距离来度量,则 2 个不确定时间序列之间的距离也是不确定的,因此现有的时间序列的相似性匹配方法不能直接运用于不确定性时间序列中.

1.2 不确定性时间序列的相似性匹配

2008 年文献[20]首次提出了不确定性时间序列的相似性匹配问题及一种不确定性时间序列隐身高维度的通用模型——隐性时间序列(cloaked time series).他们假设在每个时间点,每个随机变量的描述信息仅知其概率分布的均值和方差,在此基础上,提出一个新的基于隐式时间序列数据库的模式匹配查询,即隐式范围查询(cloaked range query, CRQ),并利用U-tree[21]索引进一步加快匹配的效率.

文献[22]提出了不确定性时间序列的相似性匹配方法,他们把不确定时间序列视为 n 维空间中的不确定性向量,认为 2 个不确定性时间序列之间的距离也是不确定的,据此得出不确定时间序列之间的欧氏距离和动态时间弯曲距离表示法,且提出了不确定性时间序列的 2 种查询方法,即概率性的有界范围查询(probabilistic bounded range query, PBRQ)和概率性的排序范围查询(probabilistic ranked range query, PRRQ).前者主要是返回那些距离小于给定阈值 ε 且查询结果出现的概率至少为 τ 的所有时间序列集合;后者返回满足相同条件的前 m 个的有序的时间序列集合.

Yeh 等人[16]提出了 PROUD 算法——不确定性时间流(uncertain time streams)的概率相似性查询处理,将不确定性流数据看成是随机变量的有序序列,且已知它的期望和方差,并认为 2 个时间序列之间的距离(欧氏距离和 DTW 距离)也是一个随机变量.他们采用数学统计方法逐步地剪枝那些候选的序列,以减少计算代价,并将 PROUD 算法运用到流环境中可用流数据的处理当中.该算法能有效地

计算 PBRQ 的结果,但不能直接计算 2 个时间序列之间的距离;另外,PROUD 算法增加一个限制,即同一条不确定时间序列的所有时间点的不确定方差是相同的. 由于流数据与时间序列数据是 2 个不同的概念,因此还要研究不确定性时间序列的查询处理.

Sarangi 等人[17]关于不确定性时间序列的相似性的广义概念提出一套理论框架,并提出 DUST 算法,计算不确定性时间序列之间的距离,它主要用于同一个不确定时间序列的不同点服从不同的概率密度函数的情况.

文献[23]主要探索出不确定性时间序列的小波分解构造方法,由于不确定数据集合的表示法占用很大的存储空间,所以需要对不确定性数据集合进行压缩表示,其思想主要运用 Haar 小波分解方法,将时间序列数据以分层方式进行正交分解,提出 2 种不同模式来优化结果中的不确定性,并通过实验得出小波表示法的压缩率. 但他们并没有提出具体的相似性匹配方法.

2 不确定时间序列建模

通常时间序列是确定的 n 维点的序列,不确定时间序列是不确定的 n 维点的序列,即在 n 维向量空间中每个点的取值是不确定的,并由这些点构成的序列. 这种不确定性表示为每个时间点的样本观测值的集合. 每一个时间点的取值用一个随机变量来表示,把不确定时间序列都认为是具有时间特性的随机变量的有序序列. 关于随机变量的描述信息有均值、概率、期望和方差等.

定义 1. 不确定时间序列[22]. 长度为 n 的不确定时间序列是由一条包含 n 个元素的序列组成的,不确定时间序列记为: $TS_U=\{(t_1,X_1),(t_2,X_2),\cdots,(t_n,X_n)\}$,其中 t_i 代表第 i 个时间点,每条元组中的属性用变量 X_t 表示,代表时刻 t 观察值的集合,记为 $X_t=\{x_{t,1},x_{t,2},\cdots,x_{t,s}\}$,$s$ 为集合 X_t 的基数即样本观察值的个数. 不确定性时间序列的原始数据集如表 1 所示:

Table 1 Uncertain Time Series Data Sets Table

表 1 不确定时间序列数据集表

Time Points	X_t
1	{0.8,0.9,1,1.5,1.8}
2	{1.0,1.08,1.1,1.8,2.0,2.1}
⋮	⋮

根据不确定性时间序列相邻时刻的值有连续性特性,如前后两时刻的值可能是相等的,这对于本文提出的选代表方法起到了很大的作用. 下面讨论相关性的定义.

定义 2. 相关性. 给定一个不确定时间序列 T_U,设前一时刻 t_i 的取值为集合 X_i,它的域为 $Dom(X_i)=\{x_{i,1},x_{i,2},\cdots,x_{i,s}\}$,若不确定时间序列前一时刻值出现 $x_{i,2}$,则后一时刻 X_{i+1} 的域值集合中也恰好包含有 $x_{i,2}$ 这个值,称之为时刻 i 的观察值与时刻 $i+1$ 的观察值具有相关性.

不确定性时间序列在数据库中的存储方式是:在时间序列数据库中,每个元组仅含 2 个属性,一个表示时间,另一个表示实数值,不确定性发生在表示实数值的属性上. 因此,可通过把普通的不确定性数据库建模方法进行精简得到适合于不确定性时间序列的建模方法.

不确定时间序列建模过程形式化描述如下:给定一个不确定时间序列 T,假定它的长度为 N,其第 i 个元素 X_i 的概率密度函数用 $f_i(\cdot)$来表示,通常这个函数是不容易得知的,本文通过迭代统计法建立概率分布,以建立不确定时间序列的概率模型. 假定每个函数 $f_i(\cdot)$表示时刻 i 有 w 个观察值的概率分布函数,$f_i(\cdot)$对应于观察值的频率记为 $p(i,1),\cdots,p(i,w)$,这里 $p(i,j)$的值表示时间序列第 i 个元素中第 j 个观察值的概率,故时刻 i 所有的概率 $p(i,j)$的累加和等于 1,有:

$$\sum_{j=1}^{w} p(i,j)=1. \tag{1}$$

本文根据所提出的概率模型,得到不确定时间序列在数据库中的存储形式如表 2 所示:

Table 2 An Uncertain Time Series

表 2 一条不确定时间序列 $T(N=3)$

Time Points		Sample Observation	Probability
t	s		
1	1	0.028	0.10
	2	0.023	0.13
	3	0.006	0.27
	4	0.038	0.42
	5	0.046	0.08
2	1	0.017	0.10
	2	0.036	0.50
	3	0.027	0.40
3	1	0.045	0.4
	2	0.053	0.6

3 不确定时间序列的相似性匹配

3.1 问题定义

第1节已经介绍了确定性时间序列的相似性匹配方法,只适应于确定的数据,对于不确定性时间序列其查询处理技术不能直接运用以前的方法,特别是当不确定时间序列的维数很高时.鉴于此,首先将问题形式化描述,并提出一种有效的方法,以降低数据的维度,并插入一个索引(R*-tree),执行相似性匹配.

由于查询时间序列和数据库存储的时间序列都可能是确定的或不确定的,根据查询中和数据库中时间序列的确定性与否,将相似性匹配问题分为以下3种类型:

1) 确定性查询序列和不确定性时间序列;

2) 不确定性查询序列和确定性时间序列;

3) 不确定性查询序列和不确定性时间序列.

对于以上3种情况,均要执行不确定时间序列的相似性匹配操作,要做的主要工作是对于数据库中已经存在的数据进行查询处理.查询处理要考虑的首要因素就是查询效率,若不确定时间序列的纵向不确定性的数据量很大时,这时不确定时间序列的可能世界更多,数据量更大,查询效率降低.但可以采取先对不确定时间序列进行Haar小波分解,这样使得数据的存储空间大大减小,使得查询效率提高.另外,要想提高查询查询结果的准确性,需要从不确定性时间序列中抽取或者选出一条(或者几条)确定的时间序列,这样可以利用现有的确定性的时间序列相似性匹配技术来处理不确定性时间序列的相似性匹配操作.

3.2 小波分解

本文提到的小波分解思想主要来自文献[23],该文采用Haar小波分解方法把不确定性时间序列数据按横向和纵向2个方向进行压缩变换,且将不确定时间序列分为2个构件:横向上是时间构件(time component),纵向上是概率构件(probabilistic component).

对于不确定性的小波表示法,每个系数表示时间范围段和概率范围段的数据行为变化.小波分解过程中特定次序的划分要么沿着时间构件,要么沿着概率构件.构造小波树的过程中,若给定小波树的某一层,则这一层的所有切分沿着时间构件或沿着概率构件是确定的,最终形成的全局切分序列与小波树的高度 $h=\mathrm{lb}\,N$ 有相同的顺序.

用一个长度为 h 的二进制串表示切分次序(cut-order),给定一个切分顺序 $q_1\cdots q_r$,该串中当某一位是0时对应着概率维切分,当某一位是1时沿着时间维切分.若不确定时间序列的长度是 N,每个元素的桶数是 w,桶为每个时隙的观察值的概率是 $p(i,j)$,小波树的超根结点是所有 $N\times w$ 个桶的均值,所有的小波系数的个数就等于 $N\times w$,构造完全的小波树的高度为 $\mathrm{lb}\,N+\mathrm{lb}\,k$.小波树的叶子层的每个叶子结点对应于在 $N\times w$ 个柱状图空间中一对相邻的桶,小波系数的值就等于这2个桶概率值的差值的一半.任何一个桶概率值的准确值都可以由沿着整个树高的小波系数的累加和重构出来,其中左分枝对应加法,右分枝对应减法.

总的来说,给定一个小波树,压缩方法就是去掉那些不能反映出原始值特征的小的系数,以最小误差法挑选出最优的小波系数.设原始的概率值为 $p(i,j)$,即第 i 个数据点的第 j 个桶,从小波分解中得到的系数为概率 $p(i,j)'$,定义误差为均方误差 E,根据误差最小,系数最大,找出需要的小波系数.E 如式(2)表示:

$$E=\sum_{i=1}^{N}\sum_{j=1}^{w}(p(i,j)-p'(i,j))^2. \tag{2}$$

例1. 假定不确定时间序列的长度为 $N=4$,每个元素的桶数 $w=4$,每个桶对应概率值如表3所示:

Table 3 The Probability Value of Each Bucket

表3 每个桶的概率值

w	Length of Time Series, N			
	1	2	3	4
1	0.1	0.2	0.3	0.4
2	0.2	0.2	0.2	0.1
3	0.6	0.3	0.1	0.2
4	0.1	0.3	0.4	0.3

由表3所知,要建立的小波树的高度为lb4+lb4=4,超根结点为所有桶的概率平均值0.25.全局切分次序的二进制序列为 $q_1q_2q_3q_4$.判断每一位取0还是取1的方法是:先取 $q_i=0$ 计算给定层的小波系数 $f_1,\cdots,f_l$;然后取 $q_i=1$ 计算给定层的小波系数 $g_1,\cdots,g_l$,再分别计算这2种情况小波系数的平方和 $\sum_{i=1}^{l}f_i^2$ 和 $\sum_{i=1}^{l}g_i^2$,q_i 取两者中的较大数.

若 $q_4=0$,按概率维切分,将上述16个数纵向划分成8个有序对,即((0.1,0.2),(0.6,0.1),

(0.2,0.2),(0.3,0.3),(0.3,0.2),(0.1,0.4),(0.4,0.1),(0.2,0.3)),每个有序对求和的一半,即对应均值为(0.15,0.35,0.2,0.3,0.25,0.25,0.25,0.25);每个有序对求差的一半,即对应小波系数为(−0.05,0.25,0,0,0.05,−0.15,0.15,−0.05),此时每个系数为 f_i,则 $\sum_{i=1}^{l} f_i^2 = 0.115$.

若 $q_4=1$,按时间维切分,将上述 16 个数横向划分成 8 个有序对,即((0.1,0.2),(0.3,0.4),(0.2,0.2),(0.2,0.1),(0.6,0.3),(0.1,0.2),(0.1,0.3),(0.4,0.3)),每个有序对求和的一半,即对应均值为(0.15,0.35,0.2,0.15,0.45,0.15,0.2,0.35);每个有序对求差的一半,即对应小波系数为(−0.05,−0.05,0,0.05,0.15,−0.05,−0.1,0.05),此时每个系数为 g_i,则 $\sum_{i=1}^{l} g_i^2 = 0.045$.

显然 $\sum_{i=1}^{l} f_i^2 > \sum_{i=1}^{l} g_i^2$,取 $q_4=0$,该层按概率维切分.

其他各层依照上述计算过程,得出 $q_3=0$,这一层按概率维切分.

综上所述,其他的每一层树的构造方法都类似计算,得出 $q_2=0$(或 1),$q_1=0$(或 1).最后得到不确定时间序列数据集的小波树,其中小波树的每个结点都是小波系数,用小波系数序列代替原有时间序列.根据式(2),选择那些使误差 E 最小的系数,这样可以得出一个不确定时间序列的压缩表示法.

小波分解表示法使原始的不确定性数据量大大减少,存储空间也相应降低,但得到的结果仍然是不确定的时间序列.在此基础上,本文提出选代表方法进行特征提取,选出一条确定的时间序列,这样在相似性查询过程中,匹配算法的时间复杂度会有所降低.下面介绍 3 种选代表方法.

3.3 选代表方法

在进行不确定时间序列相似性匹配时,需要把已创建的不确定时间序列模型的可能世界(可能世界:假设不确定时间序列的长度为 N,每一个时间点的观察值是 m 个,那么从每一个时间点任意选出一个,再将选出的 N 个时间点组合起来形成的确定时间序列有 m^N 条.)都展开成确定的时间序列,但是得到确定的时间序列的数目可能为指数级,那么在查询过程中要进行蛮力搜索,查询效率太低,查询算法的时间复杂度为指数级.这条路目前是行不通的,因此要选用其他的方法,从不确定时间序列数据库抽取出一条确定的时间序列,便于查询的时间序列与之匹配.其核心思想是通过选代表方法选出一条确定的时间序列,作为不确定时间序列的代表,选出的确定时间序列作为真实值的近似表示.

本文提出了 3 种选代表方法:

1) 概率最大法.根据不确定时间序列模型,得到每个时隙的每一个检测值的概率,选出概率最大的那个检测值,作为确定时间序列的候选值,最终得到一条确定的时间序列 $T_1=\{(t_1,v_1),(t_2,v_2),\cdots,(t_n,v_n)\}$.

2) 考虑相关性.由于不确定时间序列的每一个点是一个集合,由相关性得出确定的时间序列实际是集合与集合之间的笛卡儿积.由于不确定时间序列的连续性,后一时刻的值可能与前一时刻出现相关的情况.例如,对于同一条不确定时间序列,前一时刻的值从观察值集合中选择了值 a,而后一时刻就选出值 a,根据此特性,易选出一条确定的时间序列 T_2.

3) 均值法.均值法也是最常用的方法,在统计理论和实际应用中,它使用一个随机变量的平均值,以代表其真实值.根据不确定时间序列数据模型表,容易算出每一时隙的检测值的均值,这样得到一条确定的时间序列 $T_3=\{(t_1,\mu_1),(t_2,\mu_2),\cdots,(t_n,\mu_n)\}$.

对于选出的确定时间序列,已经排除了查询过程中时间序列的不确定性,使得查询效率有很大程度地提高,但是确定的时间序列的维度仍然与不确定时间序列相同.因此在时间序列相似性查询中,还要对确定的时间序列做降维处理,然后建立索引,再进行时间序列的查询处理.这里采用典型的时间序列降维方法,主要有离散傅立叶变换(DFT)、离散小波变换(DWT)等.

在不确定性时间序列的相似性匹配时,本文采取直接对不确定时间序列进行 Haar 小波分解,达到压缩的目的,从横向和纵向 2 个方向来降低维度,经过小波分解后得到的不确定时间序列,再对其选代表,选出一条确定的时间序列,然后执行查询操作.不确定时间序列数据集经过 2 次预处理后,它也能有效地进行查询处理,下面介绍 3 种类型的查询处理.

4 匹配算法

4.1 不确定时间序列的相似性匹配过程

这里采用概率范围(probabilistic range query, PRQ)查询[22],其定义如下:给定一个时间序列数据库 D,一个查询时间序列 Q,它们至少有一个是不确定的,查询半径 $\varepsilon\geqslant 0$,概率阈值 $\tau\in(0,1]$,

$PRQ_{\varepsilon,\tau}(Q,D)=\{T\in D\mid P(dist(Q,T)\leqslant\varepsilon)\geqslant\tau\}$,这里 $dist(\cdot,\cdot)$是欧氏距离函数,$P(\cdot)$是一个事件发生的概率.

1) 确定性查询序列和不确定性时间序列数据库

对于这种情况,首先使用小波分解方法将原始的不确定时间序列压缩变换,这样减少了可能世界,也减少了存储空间.然后用选代表方法中的概率最大法(或相关性、均值法)从压缩后的不确定性时间序列数据库中选出一条确定的时间序列,并将最大概率(或均值)作为另一个属性.再用常用的降维方法(如 APCA,Haar 小波变换)得到代表序列的特征序列,将这些序列插入到一个特定的索引(如 R*-tree).最后,抽取查询时间序列的特征序列,用确定的时间序列匹配方法进行匹配.

2) 不确定性查询序列和确定性时间序列数据库

首先将数据库中确定的时间序列用已有的降维方法降低维度,将它们插入到一个特定的索引中.然后对不确定性查询序列进行预处理,即先用小波分解方法压缩变换,接着再进行选代表,最终得到一条确定的查询序列,并降低它的维度.最后用确定的时间序列匹配方法进行匹配.

3) 不确定性查询序列和不确定性时间序列数据库

这种类型是以上 2 种情况的组合,这里不再赘述.

4.2 相似性匹配算法

算法 1. 不确定性时间序列相似性匹配.

输入:给定一个查询时间序列 Q、一个时间序列数据库 D、查询半径 ε、概率阈值 τ、小波分解子程序 *Wavelet_Decom*、选代表子程序 *Elect_Representatives*、抽取子程序 *Extract* 和相似性查询子程序 *Similarity_Search*;

输出:*PRQ* 的结果子集 *ResultSet*.

① if Q is certain and D is uncertain
② $D_1\leftarrow Wavelet\_Decom(D)$;
③ $D_2\leftarrow Elect\_Representatives(D_1)$;
④ $Q'\leftarrow$dimensionality reduction of Q;
⑤ $D_2\leftarrow$Extract all time series in D_2;
⑥ $ResultSet\leftarrow Similarity\_Search(Q',D_2,\varepsilon)$;
⑦ else if Q is uncertain and D is certain
⑧ $D\leftarrow$dimensionality reduction of D;
⑨ $Q_1\leftarrow Wavelet\_Decom(Q)$;
⑩ $Q_2\leftarrow Elect\_Representatives(Q_1)$;
⑪ $Q'\leftarrow Extract(Q_2)$;
⑫ $ResultSet\leftarrow Similarity\_Search(Q',D,\varepsilon)$;
⑬ else if Q is uncertain and D is uncertain
⑭ $Q'\leftarrow Extract(Q)$;
⑮ $D\leftarrow$Extract all time series in D;
⑯ $ResultSet\leftarrow Similarity\_Search(Q',D,\varepsilon)$;
⑰ end if

算法 1 在最坏情况下的时间复杂度是 $O(n^2)$,其中 $n=|D|$.

5 实验数据

实验数据是来自钢厂轧钢过程中凸度参数的变化情况,将原始数据文件转换成我们研究所用的实验数据.在实际钢厂轧钢过程中,将每一卷钢板作为一个周期,每一卷的检测数据是按时间顺序变化的,统计所有卷的检测数据,得到每一时间点的取值是一个可能的集合,该集合体现出不确定性,形成一个不确定的时间序列.

本文构建不确定性时间序列数据库的过程如下:给定 m 组检测数据 $G_1,G_2,\cdots,G_m$,它们的长度均为 N,每组数据均是一个二元组$\langle t_i,v_i\rangle$,包含时间 t 和值 v_2 个属性.将每一组检测数据看作一个常规时间序列,每组数据的每一时间点都是一个确定的值.读取这 m 组检测数据,用柱状图的形式统计这 m 组数据,通过迭代统计法建立概率分布,以建立不确定时间序列的概率模型.统计每组数据在每一时刻所出现的不同值的次数,并将每一时刻每个出现值的次数作为该值的概率,且每一时刻所出现的不同值反映了该时间序列的不确定性.最终能够建立一个不确定时间序列数据库 D,D 如表 4 所示:

Table 4 Probabilistic Databases *D* of Uncertain Time Series

表 4 不确定时间序列概率数据库 *D*

Time Points		Sample Observation	Probability
t	s		
1	1	0.028	0.10
	2	0.023	0.13
	3	0.006	0.27
	4	0.038	0.42
	5	0.046	0.08
2	1	0.017	0.10
	2	0.036	0.50
	3	0.027	0.40
3	1	0.045	0.4
	2	0.053	0.6
⋮	⋮	⋮	⋮
N	…	…	…

本文实验已建立的不确定时间序列数据库中有1 000条数据,每条时间序列的时间点都是130.任意给定一个查询序列Q,其序列长度也是130,查询半径$\varepsilon=0.002$.

直接对不确定时间序列执行选代表操作,然后再执行本文提出的相似性匹配过程,得到的相似性匹配结果如图1所示.根据本文提出的相似性匹配思想,先将不确定时间序列数据库进行小波压缩变换,然后再对得到的不确定性时间序列执行选代表操作,相似性匹配结果如图2所示.

预处理前后相似性匹配算法,后者的效率更高,执行效率如图3所示.以查全率和查准率作为标准,和已有的不确定时间序列相似性匹配算法PROUD以及DUST在不同标准偏差下的查全率进行比较如图4所示,查准率比较如图5所示,该算法的效果要优于已有的2种算法.在查询执行时间方面,如图6所示,所提出算法的查询执行时间也优于上述2种已有算法.

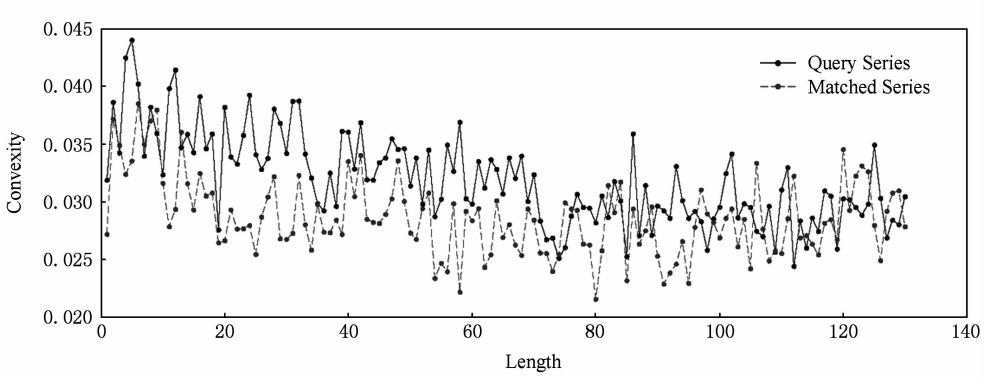


Fig. 1 Similar sequence after performing elected representatives.

图1 直接执行选代表查询出的相似性序列

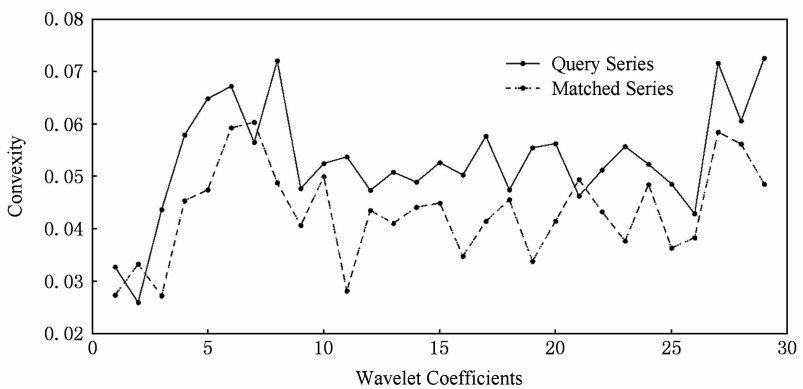


Fig. 2 Similar sequence after performing the wavelet transform and elected representatives.

图2 小波变换后执行选代表查询出的相似性序列

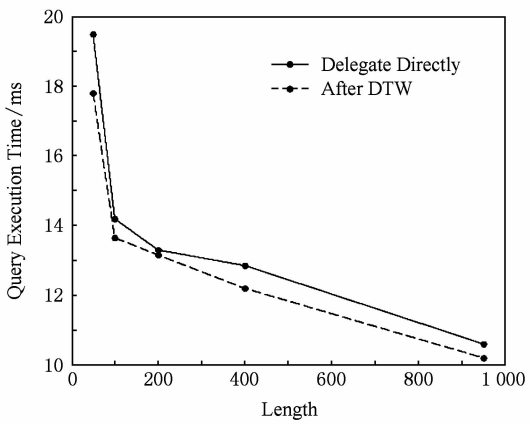


Fig. 3 Query execution time.

图3 查询执行时间

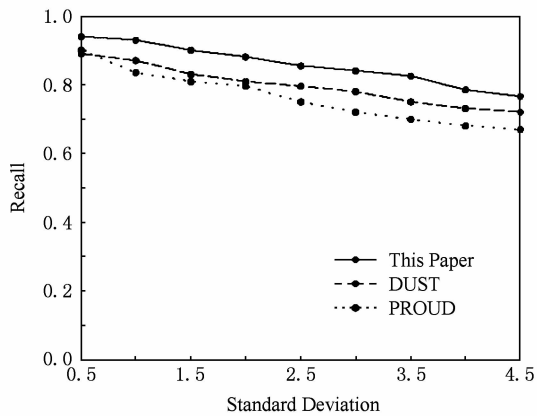


Fig. 4 Recall.

图4 查全率

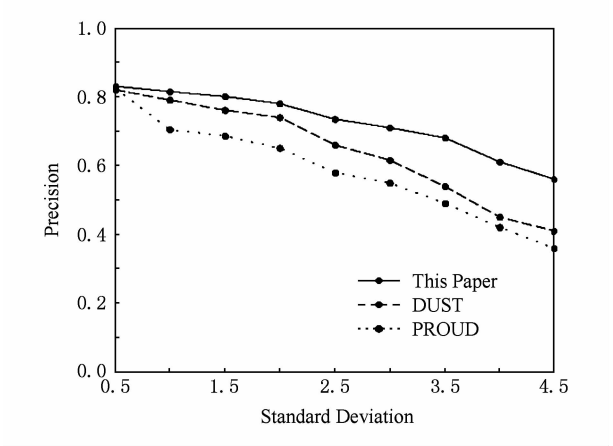


Fig. 5 Precision.

图 5 查准率

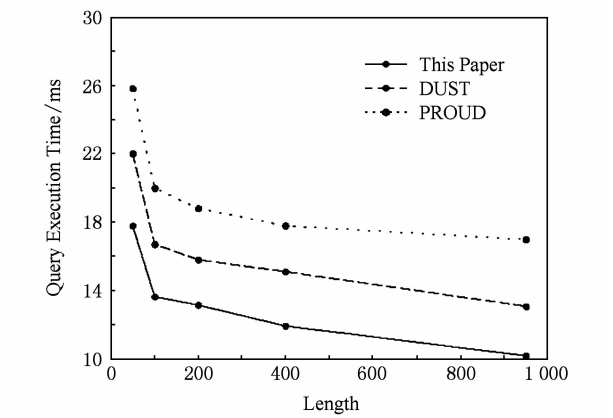


Fig. 6 Comparison on query execution time.

图 6 查询执行时间比较

实验结果表明:从数据库中 1 000 条不确定时间序列中,以标准偏差为 0.5 为例,与给定查询序列 Q 相似的序列有 27 条,实际查询出的时间序列有 30 条,其中与给定查询序列 Q 相似的序列有 25 条,其查全率为 92.5%;查准率为 83.3%;随着标准误差从 0.5 增加到 4.5,本文算法查全率从 92.9%下降到 76.2%,查准率从 83.3%下降到 50.8%;DUST 算法的查全率从 89.6%下降到 72.5%,查准率从 82.6%下降到 40.5%;PROUD 算法的查全率从 90.7%下降到 66.4%,查准率从 83.7%下降到 34.4%.从执行时间角度来看,在同等长度的不确定时间序列情况下,本文算法要比 DUST 算法大约减少 3 ms,比 PROUD 算法节省大约 7 ms.

6 总结与展望

为了减少不确定性时间序列数据库占用的空间,也为了解决时间序列数据的不确定性,同时能有效地匹配,本文根据小波分解的压缩变换方法减少了存储空间,给出了由不确定性时间序列规约为确定性时间序列的方法,并提出了不确定时间序列的相似性匹配方法.未来将致力于不确定性时间序列的其他数据挖掘任务.

参 考 文 献

[1] Renz M, Cheng R, Kriegel H P, et al. Similarity search and mining in uncertain databases [J]. Proceedings of the VLDB Endowment, 2010, 3(2): 1653-1654

[2] Cheng R, Kalashnikov D, Prabhakar S. Evaluating probabilistic queries over imprecise data [C] //Proc of the 2003 ACM SIGMOD Int Conf on Management of data. New York: ACM, 2003: 551-562

[3] Re C, Dalvi N, Suciu D. Efficient top-k query evaluation on probabilistic data [C] //Proc of the 23rd Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2007: 886-895

[4] Hua Ming, Pei Jian, Zhang Wenjie, et al. Ranking queries on uncertain data: A probabilistic threshold approach [C] //Proc of 2008 ACM SIGMOD Int Conf on Management of data. New York: ACM, 2008: 673-686

[5] Sarma A D, Benjelloun O, Halevy A, et al. Representing uncertain data: Models, properties, and algorithms [J]. The VLDB Journal, 18(5): 989-1019

[6] Ngai W K, Kao B, Chui C K, et al. Efficient clustering of uncertain data [C] //Proc of the 6th IEEE Int Conf on Data Mining (ICDM 2006). Piscataway, NJ: IEEE, 2006: 436-445

[7] Sathe S, Jeung H, Aberer K. Creating probabilistic databases from imprecise time-series data [C] //Proc of the 27th Int Conf on Data Engineering (ICDE 2011). Piscataway, NJ: IEEE, 2011: 327-338

[8] Bernecker T, Emrich T, Kriegel H, et al. A novel probabilistic pruning approach to speed up similarity queries in uncertain databases [C] //Proc of the 27th Int Conf on Data Engineering (ICDE 2011). Piscataway, NJ: IEEE, 2011: 339-350

[9] Agrawal R, Faloutsos C, Swami A. Efficient similarity search in sequence databases [C] //Proc of the 4th Int Conf on Foundations of Data Organization and Algorithms (FODO 1993). Heidelberg: Springer, 1993: 69-84

[10] Rafiei D, Mendelzon A. Efficient retrieval of similar time sequences using DFT [C] //Proc of the 5th Int Conf on Foundations of Data Organizations and Algorithms (FODO 1998). Heidelberg: Springer, 1998: 249-257

[11] Chan K P, Fu A W. Efficient time series matching by wavelets [C] //Proc of the 15th Int Conf on Data Engineering (ICDE 1999). Piscataway, NJ: IEEE, 1999: 126-133

[12] Chan K P, Fu A W, Yu C T. Haar wavelets for efficient similarity search of time-series: With and without time warping [J]. IEEE Trans on Knowledge and Data Engineering, 2003, 15(1): 686-705

[13] Popivanov I, Miller R J. Similarity search over time-series data using wavelets [C] //Proc of the 18th Int Conf on Data Engineering (ICDE 2002). Piscataway, NJ: IEEE, 2002: 212-221

[14] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast fubsequence matching in time series databases [C] //Proc of the ACM SIGMOD Int Conf on Management of Data. New York: ACM, 1994: 419-429

[15] Chakrabarti K, Garofalakis M N, Rastogi R, et al. Approximate query processing using wavelets [J]. The VLDB Journal, 2001, 10(2/3): 199-223

[16] Yeh M Y, Wu K L, Yu P S, et al. PROUD: A probabilistic approach to processing similarity queries over uncertain data streams [C] //Proc of the 12th Int Conf on Extending Database Technology (EDBT 2009). New York: ACM, 2009: 684-695

[17] Sarangi S R, Murthy K. DUST: a generalized notion of similarity between uncertain time series [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD 2010). New York: ACM, 2010: 383-392

[18] Bec*k*mann N, Kriegel H P, Schneider R, et al. The R*-tree: An efficient and robust access method for points and rectangles [C] //Proc of ACM SIGMOD Int Conf on Management of Data. New York: ACM, 1990: 322-331

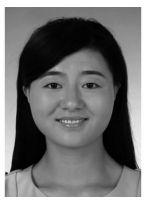
[19] Keogh E, Chakrabarti K, Mehrotra S, et al. Locally adaptive dimensionality reduction for indexing large time series databases [C] //Proc of 2001 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2001: 151-162

[20] Lian Xiang, Chen Lei, Yu X J. Pattern matching over cloaked time series [C] //Proc of the 27th Int Conf on Data Engineering (ICDE 2008). Piscataway, NJ: IEEE, 2008: 1462-1464

[21] Tao Yufei, Cheng R, Xiao Xiaokui et al. Indexing multi-dimensional uncertain data with arbitrary probability density functions [C] //Proc of the 31st Int Conf on Very Large Data Bases. New York: ACM, 2005: 922-933

[22] Aβfalg J, Kriegel H, Kröger P, et al. Probabilistic similarity search for uncertain time series [C] //Proc of the 21st Int Conf on Scientific and Statistical Database Management (SSDBM 2009). Berlin: Springer, 2009: 435-443

[23] Zhao Yuchen, Aggarwal C C, Yu P S. On wavelet decomposition of uncertain time series data sets [C] //Proc of the 19th ACM Conf on Information and Knowledge Management (CIKM 2010). New York: ACM, 2010: 129-138



Wu Honghua, born in 1986. Master in the College of Computer, Donghua University. Her main research interest is the database for uncertain time series.

Liu Guohua, born in 1966. Professor and PhD supervisor in the College of Computer, Donghua University. His main research interests include theory of database, uncertain database, web data management, and business process management (BPM).

Wang Wei, born in 1982. PhD candidate in the College of Computer, Donghua University. His main research interests include uncertain database.