不确定时间序列的统计降维方法

肖 瑞 刘国华 陈爱东 宋 转
(东华大学计算机科学与技术学院 上海 201620)

摘 要 由于不确定时间序列的长度很长,并且每个采样点的取值具有不确定性,导致了维度灾难和庞大的可能世界集,给不确定时间序列相似性匹配带来了巨大的困难,因此对不确定时间序列降维是实现对其方便存储、快速查询和相似性匹配的首要任务。不确定时间序列普遍采用小波变换的降维方法,但是该方法没有考虑到采样点之间的相关性。为解决该问题,提出一种基于概率统计和数据相关性的降维方法,该方法将不确定时间序列分为概率维度和时间维度,并分别对两维度进行降维。在时间维度,根据采样点之间的相关性,使用某个采样点代表后续相关度高的采样点;在概率维度,使用大概率点表示相邻的小概率点。实验效果表明:使用该方法对不确定时间序列进行降维后,降维序列可以保持原序列的变化趋势,压缩程度显著,并且可近似地恢复原序列。

关键词 时间序列,不确定性,降维,统计,相关性
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.028

Statistic Reduction for Uncertain Time Series

XIAO Rui LIU Guo-hua CHEN Ai-dong SONG Zhuan
(College of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract Due to the length of uncertain time series and the uncertainty of values in each sample point, time complexity is very high when matching two uncertain time series. So the dimension reduction is the primary task to match fast for uncertain series. Now, always taking wavelet transform reduces dimension for uncertain time series, but the method does not consider the correlation between every sample points. We put forward a new method based on statistics and data correlation. It divides uncertain time series to probability dimension and time dimension and performs dimension reduction respectively in the two dimensions. We used a sampling point to represent the subsequent sampling points with high correlation in time dimension, and used large probability point to represent the adjacent small probability points in probability dimension. Experimental results show that the compression ratio is remarkable when using the method to reduce uncertain time series. In addition, it can approximately recover the uncertain time series with reduced outcomes.

Keywords Time series, Uncertainty, Reduction, Statistics, Correlation

1 引言

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。由于时间序列的长度很长并且不确定时间序列在每个时间点的取值具有不确定性,导致了维度灾难和庞大的可能世界集。因此,对不确定时间序列降维是实现对其方便存储、快速查询和相似性匹配的首要任务。

本文主要研究了不确定时间序列基于统计和相关性的降维问题。先将不确定性时间序列分为时间维度和概率维度,然后提出了在时间维度和概率维度上分别降维的算法。在时间维度,统计降维方法有效地保持了序列趋势变化时的关键观察点;在概率维度,统计降维方法有效地保持了观察点中的大概率观察值,并且消除了大量的小概率观察值和奇异点。

本文第2节介绍了不确定时间序列降维的相关工作;第3节介绍了不确定时间序列的建模方法;第4节介绍了对不确定时间序列时间维度的降维方法;第5节介绍了对不确定时间序列概率维度的降维方法;第6节分析了统计降维方法;第7节是实验验证;最后是结论。

2 相关工作

因为难以对不确定时间序列这种连续且高维的数据进行有效的处理,所以对不确定时间序列进行降维是完成对不确定时间序列数据挖掘的必要环节[9]。不确定时间序列通常被映射为 n 维空间内的多个点,但是由于序列的长度很长,容易导致维度灾难(Curse of Dimensionality)[1],而目前的高维数据索引方式很难处理这种情况。

离散傅里叶变换(DFT)[1,2]是对时间序列降维的经典算

到稿日期:2013-06-16 返修日期:2013-07-28

肖 瑞(1987-),男,硕士生,主要研究方向为不确定性时间序列,E-mail:qingdaxiaorui@163.com;刘国华(1966-),男,博士,教授,博士生导师,CCF高级会员,主要研究方向为数据库理论、数据库安全、Web数据管理;陈爱东(1990-),男,硕士生,主要研究方向为不确定数据挖掘;宋 转(1988-),女,硕士生,主要研究方向为应用层组播。

法之一。它首先对时间序列进行 DFT 变化，然后取变换后的前 $K(K>0)$个系数对原序列进行近似的表示，并且 Parseval 定理说明 DFT 变换后的前 K 个保留了原序列的大部分能量。在 DFT 变换后，用于时间序列相似性度量的 Euclidean 距离依然可以得到保持，上述优点决定了 DFT 变换在时间序列索引和相似性匹配领域有一定的优势。但是将序列由时域变换到频域的过程，仅仅保留了信号的全局特性，而忽略了信号的局部差异。

Chan K. P 等人[3]首先提出了离散小波分析方法 DWT (Discrete wavelet analysis)，它反映了信号在时域和频域上的差异，在时域和频域上均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。它将序列分为尺度部分和细节部分，尺度部分通过时间序列卷积高通滤波器获得，而细节部分则再次进行 DWT 变换。其中尺度部分反映了序列的大致走向。

文献[3-6]主要是采用 Haar 小波变换对确定时间序列进行相似性搜索，并且文献[7,8]证明了 DWT 对序列降维后的恢复性优于 DFT，但是 DFT 的滤波性能优于 DWT 降维。

文献[9]提出了一维和二维 Haar 小波方法来对关系表进行压缩变换，最后进行近似查询处理。文献[12,13]则将 DWT 运用于不确定时间序列及不确定时间序列流数据的降维中，获得了良好的降维效果。但是 DWT 无法处理任意长度的时间序列，因此在使用中还存在重大的缺陷。

Korn 等人[10]提出的奇异值分解法 SVD(Singular value decomposition)，是一种基于统计概论分布的投影方法，该方法将原数据投影到较小的空间，实现数据压缩和序列降维。但是该方法的时间复杂度很高，无法在时间复杂度要求很高的挖掘任务中使用。

PCA（主成分分析）[14]是一种经典的对高维数据进行降维的方法，它将高维数据通过线性变换的方法投影到低维空间中，从而在原属性集中得到相对较少的、彼此不相关的属性子集。PCA 方法希望降维后的数据不能失真，即降掉的维度应该是噪声或冗余的属性。

3 不确定时间序列建模

不确定时间序列的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，把不确定时间序列认为是具有时间特性的随机变量的有序序列。

定义 1(不确定时间序列) 长度为 n 的不确定时间序列由一条包含 n 个元素的序列组成，记为：$TS_U=\{(t_1,r_1,V_1,P_1,X_1),(t_2,r_2,V_2,P_2,X_2),\cdots,(t_n,r_n,V_n,P_n,X_n)\}$，其中 t_i 代表第 i 个观察点，每条元组中的属性用变量 r_t,V_t,P_t 和 X_t 表示。r_t 代表第 t 个观察点的表示能力，即能表示的后续观察点和本身的个数和，初始为 1，即只表示自身观察点。V_t 代表第 t 时刻观察值的集合，记为 $V_t=\{v_{t,1},v_{t,2},\cdots,v_{t,s}\}$，$P_t$ 代表第 t 时刻观察值取值概率的集合，记为 $P_t=\{p_{t,1},p_{t,2},\cdots,p_{t,s}\}$，$X_t=\{x_{t,1},x_{t,2},\cdots,x_{t,s}\}$，$X_t$ 代表第 t 时刻观察值表示能力的集合，即该观察点上每个观察值所能表示的邻近观察值和本身的个数和，初始值为 1，即每个观察值只表示自身观察值。s 为集合 V_t 的基数，即样本观察值的个数。不确定时间序列的数据集如表 1 所列。

表 1 不确定时间序列数据集

t	r	V	p	X
1	1	{0.8,0.9,1}	{0.2,0.3,0.5}	{1,1,1}
2	1	{1.0,3.2}	{0.65,0.35}	{1,1}
3	.	…	…	…

4 时间维度降维算法

不确定时间序列采样点之间具有相关性，使用前采样点表示邻近相关度高的采样点，可以有效地减小时间序列的存储空间。将不确定时间序列分为时间维度和概率维度，在时间维度使用某个前采样点代表后续若干个相关度高的采样点，并且在该点增加额外的开销 r_i，用于记录所表示的后面连续若干个采样点和自身的个数和。该开销不仅提高了该采样点的表示能力，还提高了降维后时间序列的恢复能力。

定义 2(邻近相关度) 长度为 n 的不确定时间序列，根据两相邻采样点的相对位置不同，分为邻前相关度和邻后相关度，其中采样点 $i(1\leqslant i<n)$和采样点 $i+1$ 的邻前相关度为：$C_{(i,i+1)}=p_i'\times(|X'_{(i,i+1)}|/|X_{i+1}|)$，其中 $|X_{i+1}|$ 为第 $i+1$ 个采样点观察值的个数，$|X'_{(i,i+1)}|$ 为观察点 i 和观察点 $i+1$ 中近似相等的取值个数，p'_i 为两观察点近似相等的取值在观察点 i 的概率和，当两观察点的某两个取值满足 $|X_{i,j}-X_{i+1,k}|\leqslant\alpha$，两取值近似相等，其中 α 为比较阈值；采样点 $i(1\leqslant i<n)$和采样点 $i+1$ 的邻后相关度为：$C_{(i+1,i)}=p'_{i+1}\times(|X'_{(i+1,i)}|/|X_i|)$，其中 $|X_i|$ 为第 i 个采样点观察值的个数，$|X'_{(i+1,i)}|$ 为观察点 i 和观察点 $i+1$ 中近似相等的取值个数，p'_{i+1} 为两观察点近似相等的取值在观察点 $i+1$ 的概率和，当两观察点的某两个取值满足 $|X_{i,j}-X_{i+1,k}|\leqslant\alpha$，两取值近似相等，其中 α 为比较阈值。

当 $C_{(i,i+1)}>\beta$ 时，观察点 i 可近似地表示观察点 $i+1$，则删除观察点 $i+1$，并将观察点 $i+1$ 的记录 r_{i+1} 加至观察点 i 的记录 r_i。当 $C_{(i+1,i)}>\beta$ 时，观察点 $i+1$ 可近似地表示观察点 i，则删除观察点 i，并将观察点 i 的记录 r_i 加至观察点 $i+1$ 的记录 r_{i+1}。

定义 3(非邻相关度) 长度为 n 的不确定时间序列，根据两不相邻采样点的相对位置不同，分为非邻前相关度和非邻后相关度，其中采样点 $i(1\leqslant i<n)$和采样点 $j(1<j\leqslant n)$且 $(j-i>1)$的非邻前相关度为：如果对每个 $k(i\leqslant k<j)$，$C_{(k,k+1)}>\beta$，则非邻前相关度为：$C_{(i,j)}=(C_{(i,i+1)}+C_{(i+1,i+2)}\cdots+C_{(j-2,j-1)}+C_{(j-1,j)})/|j-i|$；否则非邻前相关度为：$C_{(i,j)}=0$。其中采样点 $i(1\leqslant i<n)$和采样点 $j(1<j\leqslant n)$且 $(j-i>1)$的非邻后相关度为：如果对每个 $k(i<k\leqslant j)$，$C_{(k,k-1)}>\beta$，则非邻后相关度为：$C_{(j,i)}=(C_{(j,j-1)}+C_{(j-1,j-2)}\cdots+C_{(i+2,i+1)}+C_{(i+1,i)})/|j-i|$；否则非邻后相关度为：$C_{(j,i)}=0$。

如果对每个 $k(i\leqslant k<j)$，$C_{(k,k+1)}>\beta$，则 $C_{(i,j)}>[|j-i|\times\beta]/|j-i|$，即 $C_{(i,j)}>\beta$，同理如果对每个 $k(i\leqslant k<j)$，$C_{(k+1,k)}>\beta$，则 $C_{(j,i)}>\beta$。

当 $C_{(i,j)}>\beta$ 时，观察点 i 可近似地表示观察点 $i+1$ 到观察点 j 的共 $|j-i|$ 个观察点，则删除该 $|j-i|$ 个观察点，并将该 $|j-i|$ 个观察点的记录加至 r_i。当 $C_{(j,i)}>\beta$ 时，观察点 j 可近似地表示观察点 i 到观察点 $j-1$ 的共 $|j-i|$ 个观察点，则删除该 $|j-i|$ 个观察点，并将该 $|j-i|$ 个观察点的记录加至 r_j。

通过定义邻近相关度和非邻相关度，我们可以构建不确

定时间序列的相关度矩阵如下：

$$c=\begin{pmatrix} c_{1,1} & c_{1,2}\cdots c_{1,n-1} & c_{1,n} \\ \vdots & \ddots & \vdots \\ c_{n,1} & c_{n,2}\cdots c_{n,n-1} & c_{n,n} \end{pmatrix}$$

其中，$C_{(i,j)}(1\leqslant i,j\leqslant n)$，如果 $i=j-1$，则为邻前相关度，如果 $i=j+1$，则为邻后相关度；如果 $j-i>1$，则为非邻前相关度，如果 $i-j>1$，则为非邻后相关度；如果 $i=j$，则 $C_{(i,j)}=1.0$。

通过相关度矩阵可以方便地分析不确定时间序列每个采样点之间的相关性。如果 $C_{(1,n)}>0$，则第一个采样点可以近似地表示后续 $n-1$ 个采样点，说明不确定时间序列在区间 $[1,n]$ 内变化趋势平缓，并且每个采样点的观察值集合近似相等。

对长度为 n 的不确定时间序列进行时间维度的统计降维，可采用两种方式来遍历整个时间序列的每个观察点，分别为前序遍历和后序遍历。其中前序遍历从观察点 $i=1$ 开始，计算每两个相邻观察点的邻后相关度，到观察点 $i=n$ 结束；其中后续遍历为从观察点 $i=n$ 开始，计算每两个相邻观察点的邻前相关度，到观察点 $i=1$ 结束。

对不确定时间序列时间维度降维的后序遍历算法如下，关于前序遍历的算法只需要更改迭代处理的顺序即可。由于非邻相关度是在邻近相关度的基础上得到的，使用非邻相关度进行时间维度降维需要消耗更多的时间，因此该算法采用邻近相关度作为相关性的度量方式。

算法名称：不确定性时间序列时间维度降维

输入：不确定时间序列 $T=\{(t_1,r_1,V_1,P_1,X_1),(t_2,r_2,V_2,P_2,X_2),\cdots,(t_n,r_n,V_n,P_n,X_n)\}$，比较阈值 α，相关度阈值 β

输出：T 在时间维度降维后的序列 T′

```
1. for(i=n-1;i>0;i--){
2. if(C(i,i+1)≥β){
3. ri=ri+ri+1;
4. delete ti+1;}
5. else {
6. record ti+1;}
7. }
```

5 概率维度降维算法

对概率维度进行降维的主要思想是在每个时间点上使用大概率取值近似表示临近的小概率取值，并且消除由于测量误差和传输噪声造成的奇异点。奇异点为在每个时间点上偏离该时刻取值整体分布的观察值。在降维过程中，为序列每个时间点的全部取值增加额外的开销 X_i，记录每个取值所表示的临近小概率取值和自身的个数和，该开销提高了采样值的表示能力。

对不确定时间序列进行概率维度降维时使用的变量如表 2 所列。

表 2 概率维度降维参数

参数	符号	计算方式	分析
平均距离	d_i	$[\max(V_i)-\min(V_i)]/(s-1)$	计算邻近点
平均概率	p_i	$1/s$	计算小概率点
倍数阈值	k	输入参数	计算奇异点

对不确定时间序列 T 的每个观察点 $t_i(1\leqslant i\leqslant n)$，计算平均距离 d_i 和平均概率 p_i，并按照 V_i 集合内的观察值进行排序；检查 T 的每一个观察点 $t_i(1\leqslant i\leqslant n)$ 中所有观察值 $v_{i,j}$（$v_{i,j}\in V_i$ 且 $1\leqslant j\leqslant s$）和观察值的概率 $p_{i,j}$（$p_{i,j}\in P_i$ 且 $1\leqslant j\leqslant s$），如果 $|v_{i,j}-v_{i,k}|<d_i(k=j+1$ 或 $k=j-1,1\leqslant k\leqslant s)$ 且 $p_{i,j}<p_i$，则删除 $v_{i,j}$，$p_{i,j}$，$x_{i,j}$，并将观察值 $v_{i,j}$ 的记录 $x_{i,j}$ 加至观察值 $v_{i,k}$ 的记录 $x_{i,k}$。

对 T 的每个观察点 $t_i(1\leqslant i\leqslant n)$，检查该观察点最大观察值 $v_{i,s}$ 和最小观察值 $v_{i,1}$（假设每个观察点按照观察值升序排列），如果 $|v_{i,s}-v_{i,s-1}|>k\times d_i$ 且 $p_{i,s}<p_i$，则 $v_{i,s}$ 为奇异点，删除 $v_{i,s}$，$p_{i,s}$，$x_{i,s}$，并将观察值 $v_{i,s}$ 的记录 $x_{i,s}$ 加至观察值 $v_{i,s-1}$ 的记录 $x_{i,s-1}$；如果 $|v_{i,1}-v_{i,2}|>k\times d_i$ 且 $p_{i,1}<p_i$，则 $v_{i,1}$ 为奇异点，删除 $v_{i,1}$，$p_{i,1}$，$x_{i,1}$，并将观察值 $v_{i,1}$ 的记录 $x_{i,1}$ 加至观察值 $v_{i,2}$ 的记录 $x_{i,2}$。

对长度为 n 的不确定时间序列进行概率维度的统计降维，可采用两种方式来遍历每个观察点的观察值，分别为上序遍历和下序遍历。其中上序遍历为从观察值 $j=1$ 开始，计算每个观察值的概率和该观察值与相邻观察值的距离，直到观察值 $j=s$ 结束；其中下序遍历从观察值 $j=s$ 开始，计算每个观察值的概率和该观察值与相邻观察值的距离，直到观察值 $j=1$ 结束。

定义 4（重启策略） 对长度为 n 的不确定时间序列进行概率维度的统计降维，对每个观察点进行多次遍历和消除邻近小概率点及奇异点，直到该观察点的观察值不能被消减才停止遍历，该策略称为重启策略。

采用重启策略的有益方面：

(1)如果观察点存在奇异点，则在初次统计分析时求得的平均距离较大，导致其余概率很小的观察值由于没有距离小于平均距离的邻近值而无法消减。重启策略可有效解决该问题。

(2)如果观察点中观察值概率在初次统计分析时略大于平均概率，但是由于与邻近观察值的距离远远小于平均距离而不满足统计降维条件，使得该观察值无法消减，则重启策略可通过重新统计分析将其合并至邻近观察值。

对不确定时间序列概率维度降维的算法如下，该算法采用上序遍历，关于下序遍历的算法只需要更改迭代处理的顺序即可。

算法名称：不确定性时间序列概率维度降维

输入：不确定时间序列 $T=\{(t_1,r_1,V_1,P_1,X_1),(t_2,r_2,V_2,P_2,X_2),\cdots,(t_n,r_n,V_n,P_n,X_n)\}$，距离倍数阈值 k

输出：T 在概率维度降维后的序列 T″

```
1.  for(i=1;i≤n;i++){
2.  {
3.  pi=1/s;
4.  di=[max(Vi)-min(Vi)]/(s-1)
5.  sort(Vi,Pi) by Viasc;
6.  m=s;
7. for(j=1;j≤s;j++){
8.  if(pi,j≤ pi){
9.    if(|vi,j- vi,j+1|≤di){
10.      xi,j+1=xi,j+xi,j+1;
11.      delete vi,j,pi,j,xij;
12.      m--;
13.      }
14.   }
```

```
15.     s=m;
16.   }
17.   if((p_{i,s}≤p_i)&&(|v_{i,s}− v_{i,s−1}|≥k×d_i)){
18.       x_{i,s−1}=x_{i,s}+x_{i,s−1};
19.       delete v_{i,s},p_{i,s},x_{i,s};
20.       m−−;
21.     }
22.   if((p_{i,1}≤p_i)&&(|v_{i,1}− v_{i,2}|≥k×d_i)){
23.       x_{i,2}=x_{i,1}+x_{i,2};
24.       delete v_{i,1},p_{i,1},x_{i,1};
25.       m−−;
26.     }
27.   s=m;
28. } while(X_i changed)
29. }
```

6 统计降维分析

对不确定时间序列的整体降维过程为：首先对时间维度降维，然后对概率维度降维。采取该顺序的原因是：在时间维度上的降维是通过相邻观察点取值的相关性计算，而在概率维度上的降维是通过统计分析观察值的概率和距离进行，并且尽可能地消减邻近小概率取值。如果首先进行概率维度上的降维，则会降低相邻采样点之间的相关性，使得时间维度上的降维效果变差。

对不确定时间序列 T 进行时间维度降维的算法首先需要遍历每个观察点，然后计算相邻观察点的相关度；其中计算相邻观察点相关度的最大时间复杂度为 $O(s^2)$，则时间维度降维算法的时间复杂度为 $O(ns^2)$。对 T 进行概率维度上的降维，最坏的情况是时间维度上的降维没有消除任何观察点，则对 n 个观察点都需要遍历所有观察值，假设重启的次数为 m，则概率维度上降维算法的时间复杂度为 $O(nsm)$，综上所述，对 T 进行统计降维的时间复杂度为 $\mathrm{Max}(O(ns^2),O(nsm))$。

不确定时间序列经过时间维度和概率维度降维后，序列恢复则将时间维度上的前采样点替代后续已删除并合并至该点的多个采样点。

下面通过对图 1 和表 3 所示的不确定时间序列进行统计降维，来进一步说明降维的效果；图 2 是对其进行降维后的效果图，其中输入参数 α 为 0.1，β 为 0.7，k 为 3。可以看出，对不确定时间序列进行统计降维后，可以有效地保持时间序列的整体变化趋势，并且通过消除小概率点和奇异点，使得每个观察点的观察值分布趋于均匀，提高了观察值的表示能力。图 3 是对降维后序列进行恢复后得到的序列，可以看出，恢复后的时间序列保持了原序列的趋势特征，并且近似地保持了原序列观察点中观察值的取值分布。

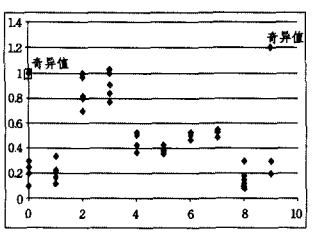


图 1 不确定时间序列示例

表 3 示例序列概率集合

t	P(观察值概率集合)	t	P
0	{0.3,0.1,0.4,0.1,0.1}	5	{0.13,0.27,0.5,0.1}
1	{0.1,0.3,0.3,0.1,0.2}	6	{0.1,0.45,0.45}
2	{0.3,0.2,0.1,0.3,0.1}	7	{0.35,0.4,0.25}
3	{0.5,0.2,0.1,0.1,0.1}	8	{0.2,0.1,0.3,0.1,0.3}
4	{0.4,0.3,0.2,0.1}	9	{0.2,0.7,0.1}

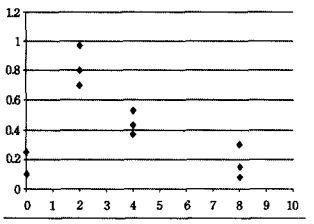


图 2 示例序列统计降维效果

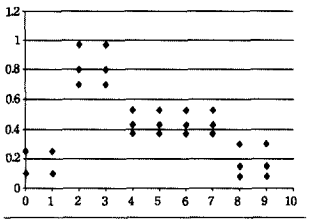


图 3 示例序列降维后恢复效果

7 实验

7.1 实验环境

本次实验的环境为：Windows7 32 位操作系统；英特尔酷睿 I3-370 处理器；NVIDIA Geforce GT330M 显卡。

7.2 实验数据

实验数据是来自钢厂轧钢过程中一卷钢板的凸度值变化情况。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，每一时隙的变化是不确定的，形成一个不确定的时间序列。

假定检测过来的原始数据与数据库中的时间范围是相同周期的，每一卷检测值是一个时间序列，每一条元组都是一个 2-tuple$\langle t_i,v_i\rangle$，循环读取这组检测数据中的每一个二元组，与数据库中的对应时刻的值进行比较，更新数据库中的值，统计出每一时刻所出现的观察值的频率。

我们主要通过统计每一组值来找到这样的经验值。具体做法是：将原始检测数据通过统计计算，得到每个时刻样本的可能出现值和可能值的概率，实际每个周期大概是 150 个时间点，这样形成了一条不确定的时间序列数据。最后得到 1000 条不确定时间序列，每条时间序列的时间点都是 150。

7.3 实验结果

本文通过两种参数：压缩度和恢复度来进行结果分析。压缩度为降维前后时间序列所占的空间差值与降维前所占空间的比值；恢复度通过以下方法来衡量。

(1)给定查询序列 Q 和序列集合 S，通过一定的相似性匹配算法，使用 Q 对 S 进行相似性查询，获得输出集合 TS。

(2)对集合 TS 内的序列进行时间维度和概率维度的降维，并根据降维后的结果将序列恢复到集合 TS' 中。

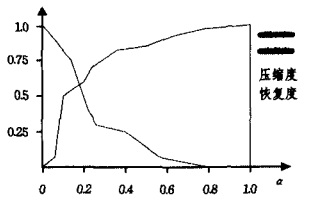
(3)通过同样的相似性匹配算法，使用 Q 对 TS' 进行相似性查询，获得输出集合 TS''。则恢复度为 TS'' 集合内序列个数与 TS 集合内序列个数的比值。

实验结果表明：对 1000 条不确定时间序列执行时间维度和概率维度进行降维后，压缩度为 0.72。通过期望匹配算法验证的恢复度为 0.8，其中输入参数 α 为 0.1，β 为 0.7，k 为 3。

当改变输入参数 α 时，压缩度和恢复度的变化如图 4 所示，压缩度随 α 单调递增，而恢复度随 α 单调递减；当改变输入参数 β 时压缩度和恢复度的变化如图 5 所示，压缩度随 β 单调递减，而恢复度随 β 单调递增。所以要选取合适的 α 和 β

值，保证压缩度和恢复度都可以满足需求。

最后通过与 haar 小波变换降维方法的对比实验，验证了统计降维方法具有很低的时间复杂度和空间复杂度。图 6 为统计降维方式和小波变换降维方式的执行效率对比，表明统计降维方式有很低的时间复杂度；图 7 为在恢复度都为 0.8 时，统计降维方式和小波变换方式占用存储空间的对比，表明统计降维方式有很低的空间复杂度。



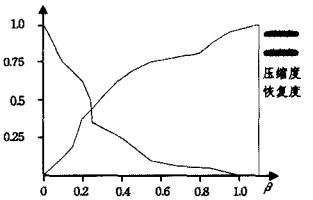
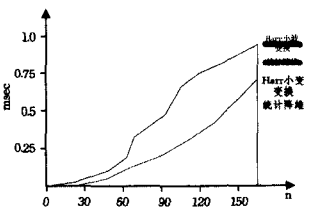


图 4　压缩度和恢复度与 α 的关系　图 5　压缩度和恢复度与 β 的关系



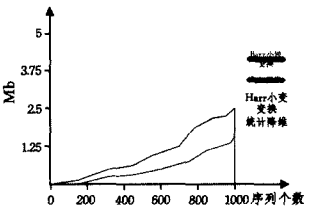


图 6　降维效率对比图　　图 7　存储空间对比图

结束语　本文提出了对不确定时间序列基于统计和相关性的降维方法，该方法可以有效地降低不确定时间序列存储空间，相较于其他针对不确定时间序列的降维方法，有很低的时间复杂度和空间复杂度，并且可以处理任意长度的时间序列。后续的主要任务是将该降维算法应用到不确定时间序列的索引和相似性匹配过程中。

参 考 文 献

[1] Agrawal R, Faloutsos C, Swami A. Efficient Simillarity Search in Sequence Databases[C]// FODO. 1993:69-84

[2] Rafiei D, Mendelzon A O. Querying Time Series Data Based on Sililarity[J]. IEEE TKDE, 2000, 12(5):675-693

[3] Chan K-P, Fu A W-C. Efficient time series matching by wavelets [C]// Proceedings of the 15th International C on ference on Data Engineering. Sydney, Australia, 1999:126-133

[4] Popivanov I, Miller R-J. Similarity Search Over Time-Series Data Using Wavelets[C]// Proceedings of the 18th International Conference on Data Engineering. San jose, CA, 2002:212-221

[5] 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 Transactions on Knowledge and Data Engineering, 2003, 15(1):686-705

[6] Zhao Yu-chen, Aggarwal C-C, Yu P-S. On wavelet decomposition of uncertain time series data sets[C]// Proceedings of the 19th ACM Conference on Information and Knowledge Management. Toronto, Canada, 2010:129-138

[7] Popivanov I. Similarity search over time-series data using wavelets. Data Engineering [C] // Proceedings. 18th International Conference. 2002:212-221

[8] Kawagoe E, Ueda S. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration[J]. Data Mining and Knowledge Discovery, 2002, 7:349-371

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

[10] Korn F, Jagadish H, Faloutsos C. Efficiently supporting ad hoc queries in large datasets of time sequences[C]// Joan P, ed. Proceedings of ACM SIGMOD International Conference on Management of Data. Tuescon, AZ USA: Morgan Kaufmann Publishere, 1997:286-300

[11] Faloutsos C, Ranganathan M, Manolopoulos Y. Fast Subsequence Matching in Time Series Databases [C]// Proceedings of the ACM SIGMOD International Conference on Management of Data. Minneapolis, 1994:419-429

[12] Cormode G, Garofalakis M. Histograms and Wavelets on Probabilistic Data[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(8):1142-1157

[13] Liao Kang-li, Chen Hua-hui, Qian Jiang-bo, et al. Wavelet Decomposition Algorithm for Uncertain Data Streams[M]. 2011: 965-970

[14] Vidal R, Ma Ya, Sastry S. Generalized principal component analysis (GPCA)[C]// Proceeding of IEEE Conference on Computer Vision and Pattern Recognition. 2003:18-20

（上接第 121 页）

[2] 薛艳，雷红轩，李永明. 基于可能性测度的计算树逻辑[J]. 计算机工程与科学，2011，33(9)：70-75

[3] 李丽君，雷红轩，李永明. 基于可能测度的 LTL 模型检测[J]. 计算机学报，2012(3)：33-39

[4] 希利尔，等. 数据，模型与决策[M]. 北京：中国财政经济出版社，2012

[5] 宋伟. 工程管理案例[M]. 北京：机械工业出版社，2012

[6] Li Li-jun, Li Yong-ming. Model-checking of linear-time properties in possibilistic Kripke structure. Quanitative Logic and soft computing[C]// World Scientific Publishing Co. Pte. Ltd. 2012: 287-294

[7] Xue Yan, Lei Hong-xuan, Li Yong-ming. Possibilistic Kripke structure decision processes. Quanitative Logic and soft computing[C]// World Scientific Publishing Co. Pte. Ltd. 2012: 295-302

[8] 徐泽水. 基于期望值的模糊多属性决策法及其应用[J]. 系统工程与实践，2004(1)：109-113

[9] Baier C, Katoen J-P. Principles of Model Checking[M]. Cambridge: The MIT Press, 2007:816-832