基于离散余弦变换的时间序列相似性检索①

刘端阳, 张瑞强

(浙江工业大学 计算机科学与技术学院, 杭州 310023)

摘 要: 在时间序列相似性研究领域已经发展了多种方法用于时间序列的表示, 以达到降低序列维度的目的. 作为一种经典的时域-频域转换方法, 离散余弦变换目前已经在图形图像处理等领域得到了广泛的应用. 将此方法应用于时间序列的表示上, 在变换后的数据上进行相似性查询等操作. 实验表明, 相对以前的方法, 这种方法具有明显的性能提升.

关键词: 时间序列; 距离度量; 相似性; 维度约简; 离散余弦变换

Time Series Similarity Based on Discrete Cosine Transform

LIU Duan-Yang, ZHANG Rui-Qiang

(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: In the research domain of time series similarity, there have been many approach that are developed to represent time series in the order of reducing the dimensionality. As one of the classic time domain-frequency domain transformation method, the discrete cosine transform has been widely used in the field of graphics and image processing. In this paper, we used this method to represent time series, and implenmet querying operation etc. on the transformed data. Our experiments indicate that this method obviously improve the performance compared with previous method..

Key words: time series; distance measure; similarity; dimension reduction; discrete cosine transform

1 前言

在生产和生活中, 大量的时间序列数据源源不断地产生, 吸引了越来越多的研究者的兴趣. 对时间序列数据相似性问题的研究成果已广泛应用在语音处理、医学、金融、传感器网络等领域, 产生了巨大的经济和社会价值.

时间序列 S 是指按时间顺序排列的, 具有相等时间间隔的实数数据列表, 记为 $S=\{S_1,S_2,\cdots,S_n\}$, 其中时间序列长度为组成 S 的实数值个数, 记为 $|S|=n$ [1]. 给定一个查询序列 $Q=\{q_1,q_2,\cdots,q_m\}$, 一个数据序列 $S=\{s_1,s_2,\cdots,s_n\}$, 如果序列 Q 和序列 S 满足 $dist(Q,S)\le\varepsilon$, 则认为时间序列 Q 和 S 是相似的. 其中, ε 是时序相似阈值, $dist(Q,S)$ 是一个距离度量[1].

实际应用中最常使用的两种距离度量是欧几里得(Euclidean)距离与动态时间弯曲(DTW, Dynamic Time Warping). 欧几里得距离通常也被称为 L_2 范数, 计算欧几里得距离要求两个序列等长, 且两个序列中的值必须是一一对应, 每一对差值的权重相同. 欧氏距离以其简单实用被广泛采用. 动态时间弯曲源于语音识别, 其采用动态规划的思想递归定义, DTW 函数去掉了数据时序 S (长度n)和查询时序 Q (长度m)保持等长的要求(总长 L 满足条件: m, n≤L≤(m+n)), 容许序列点自我复制后再进行等长匹配, 从而具有较高精度. 但直接实现 DTW 的算法的时间复杂度为 $O(mn)$, 相对计算欧几里得距离的 $O(n)$ 的时间复杂度, 其计算效率较低.

时间序列的相似性度量是衡量两个时间序列的相似程度的方法, 它是时间序列分类、聚类、异常发现等问题的基础, 也是时间序列数据挖掘的核心问题之

① 收稿时间:2011-12-25;收到修改稿时间:2012-02-18

一. 由于时间序列所具有的高维度、高噪音、高数据量等特点, 传统的数据挖掘方法不能直接应用在时间序列的数据挖掘过程中, 因此, 快速、有效的时间序列相似性度量方法成为学术界和工业界的热点研究课题之一.

本文在已有工作的基础上, 提出了一种新型的时间序列数据变换方法, 相比传统的时间序列数据转换方法, 本文所提出的方法有效提高了相似性搜索的效率.

2 相关工作

由于时间序列是典型的高维数据, 面对海量数据, 直接去操作一个高维的数据空间是很困难的. 若直接用 SAM(Spatial Access Method)多维索引结构(如 R*树)来索引这种高维数据, 则容易导致维度灾难. 因此, 需要研究合适的数据表示形式, 进行维度约简, 在高效、方便的表示形式上进行有效的挖掘. 衡量维度约简效果的重要标准之一是要满足“无漏报”原则”, 即要求数据表示满足以下条件(下边界引理)[2]:

$$D_F(q',s') \le D(q,s)$$

即约简后的距离应不大于原先的距离. 其中: q 是查询序列, s 是数据集中的任意序列; D_F 是约简空间中的两序列距离, D 是真实的两序列距离.

时间序列的变换是时间序列研究的关键课题之一, 已经有很多学者在这方面做了很多重要的工作, 为进一步对时间序列进行数据挖掘提供了基础.

Agrawal 等人在文献[3]中使用离散傅里叶变换(DFT, Discrete Fourier Transform)将时间序列从时域空间变换到频域空间. 根据 Parseval 定理, 在频域空间中, DFT 保持原序列的 Euclidian 距离, 即满足下边界引理. 由于原始序列的能量主要集中的较低的几个频率上, 因此可取频域的前 k 个系数形成一个 k 维点来表示原序列, 相似度量就是 k 维点的 Euclidian 距离, 并建立索引. DFT 作为经典的信号变换方法, 全局性能良好, 但丢失了时间局部化的重要特征.

离散小波变换(DWT, Discrete Wavelet Transform)[4]作为一种较新的线性变换技术使用变换后生成的少数小波参数近似模拟原始信号. 其小波参数具有时间/频率特性, 可以保持比 DFT 更多的信息, 满足多分辨率的表示需求. DWT 的不足是只能处理长度为的序列, 限制了其应用范围.

奇异值分解(SVD, Singular Value Decomposition)[5]方法将一组 n 维点投影到 k 维子空间(k<n), 通过变换使得在投影维上的方差最大化. SVD 类似于主成分分析(PCA, Prime Component Analysis), 其主要缺点是计算开销大, 随着数据的增量加入, 索引性能将产生退化, 需要周期性地重组索引结构.

分段常数近似(PAA, Piecewise Aggregate Approximation)[6]方法将序列分成等长的 n 个段, 各段的平均值就构成该序列的 k 维特征向量. 这种方法的优点是易于理解和实现, 转换速度快、无漏报、线性建立索引开销、存在更灵活的距离度量等.

以上简要介绍了对时间序列进行变换时所遵循的一些基本原则及以往的研究人员提出的一些经典的变换方法. 在第 3、第 4 部分中, 我们将着重介绍利用离散余弦变换的方法对时间序列进行变换, 在变换后的序列上进行时间序列的相似性研究.

3 离散余弦变换方法

Ahmed 最早于 1974 年提出了离散余弦变换(DCT, Discrete Cosine Transform)的概念[7], 作为一种实数域变换, DCT 克服了离散傅里叶变换中复数域运算的缺点, 因此在数字信号处理、频谱分析、图形图像处理等领域得到了广泛的应用. 关于 DCT 的详细情况, 在一般的关于信号处理的参考书中都可以找到, 这里仅作一个简要的介绍.

DCT 有多种表示形式, 这里给出比较常用的一种定义如下: 设原始时间序列 x, 经变换后的序列为 X, 则有:

$$X_0 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n-1} x_i$$

$$X_k = \sqrt{\frac{2}{n}}\sum_{i=0}^{n-1} x_i \cos[\frac{(2m+1)k\pi}{2n}], k = 1,2,\cdots,n-1$$

利用公式直接计算 DCT 的时间复杂度为 $O(n^2)$, 但是类似于快速傅里叶变换, 存在时间复杂度为 $O(n\log(n))$ 的快速算法, 提高了 DCT 计算的效率.

同 DFT 一样, DCT 是一种正交变换, 因此 DCT 满足 Parseval 定理, 时域和频域的能量保持不变, 即:

$$\sum_{t=o}^{n-1}|x_t|^2 = \sum_{f=0}^{n-1}|X_f|^2$$

从而有:

$$|x-y|^2=|X-Y|^2$$

上式表明, 两个时间序列 x 和 y 的 Euclidean 距离与它们分别经过 DCT 变换后的序列 X 和 Y 的 Euclidean 距离是相等的. 对于大部分信号(时间序列)来说, 其能量都集中在 DCT 变换后的低频部分, 因此, 我们可以使用前面的较少的几个序列值来近似表示变换后的序列, 从而可以得到:

$$|x-y|^2\geq|X'-Y'|^2$$

其中, X'、Y' 为变换后序列的近似表示. 上式保证了使用 DCT 变换后的序列的近似表示能够满足下边界引理, 从而保证了没有漏报(false dismissals)的产生, 对于查询处理中产生的多报(false alarm), 可以通过一个后处理过程裁减掉.

4 实验结果与分析

为了验证我们所提出方法的有效性, 我们将我们的方法与当前流行的一些方法进行了一系列对比实验. 在前面第三部分提到的一系列变换方法中, DFT 与 DCT 同为时域-频域转换方法, 而 PAA 被普遍认为是在大多数情况下表现最好的一种变换方法, 因此, 在我们的实验中, 主要将 DCT 与 DFT、PAA 的进行了对照.

实验所采用的数据集为随机漫步(Random Walk)数据, 对于每一个序列 S, 有: $s_i=s_{i-1}+0.06z_t$, 其中 s_1 为(2, 6)之间的平均分布, z_t 为标准正态分布.

4.1 下边界紧密度(TLB, tightness of lower bounds)比较

为了保证没有漏报的产生, 变换后的序列必须下边界于原始时间序列, 我们定义下边界紧密度为: TLB=下边界距离/真实距离. 下边界紧密度越高, 变换后的序列与原始序列之间越紧密, 后处理过程所需的计算越少, 从而可以大大提高搜索的效率.

生成 400 组长度为 400 的序列, 在变换后的序列长度分别为 k=2、4、8 的情况下, 我们通过对变换后的序列的能量与原始序列进行对比, 取其平均值, 得出实验结果如表 1 所示.

表 1 DCT 与 PAA、DFT 的 TLB 比较

序列长度	*k*=2	*k*=4	*k*=8
PAA	0.37	0.51	0.72
DFT	0.52	0.65	0.80
DCT	0.64	0.77	0.83

通过对比可以看出, 使用 DCT 所得到的序列的下边界紧密度要明显优于通过 PAA 与 DFT 所得到的. 同时我们也可以看到, 在使用较少的系数的情况下, 通过 DCT 所得到的序列即可比较理想的逼近原始的时间序列.

(2) 查询效率比较

以同样的方式分别生成 200 组序列长度为 200、400 组序列长度为 400、800 组序列长度为 800 的测试数据, 采用交叉验证的方式进行查询操作. 变换后序列长度取 k=8, 相似阈值分别取 sqrt(200*4)、sqrt(400*4)、sqrt(800*4)(参照[3]). 同时, 以线性扫描(LM, Linear Scan)查询所需的时间作为时间基准. 实验结果如表 2 所示.

表 2 LM、PAA 与 DCT 查询所需时间的相对比较

	n=200	*n*=400	*n*=800
LM	1	1	1
PAA	0.36	0.38	0.47
DFT	0.44	0.52	0.61
DCT	0.34	0.41	0.55

由表中的比较可以看出, 相对于直接线性扫描的方式, PAA 与 DCT、DFT 均有较高的搜索效率. 在不同的数据集大小及序列长度下, DCT 均具有比 DFT 更高的搜索效率; 在数据集较小, 序列长度较短时. DCT 相比 PAA 有一定优势, 随着数据集增大及序列长度的增加, PAA 与 DCT 均有一定程度的退化, 由于 DCT 的计算量要比 PAA 更高, DCT 效率的退化相对更明显一些, 在今后的工作中还需要做进一步的优化以提高算法的健壮性.

5 总结

时间序列的表示是目前时间序列相似性问题研究中的一个重要的研究课题, 合理而有效的表示形式对时间序列相似性问题的研究具有重要意义. 本文采用了一种新的变换方法对时间序列数据进行变换从而达到维度约减的目的, 从而可以有效的对时间序列进行储存、索引等操作. 实验表明, 本文所提出的方法可以有效的对时间序列进行表示, 对于时间序列相似性的研究具有一定的意义.

(下转第 186 页)

4 结论

本文提出了一种基于变差函数和方向小波的多尺度边缘检测新方法. 该方法通过计算各个子区域内的平均变差函数来判断图像各个子区域内边缘的方向性, 然后根据该区域边缘的方向性, 选择合适的方向小波旋转参数, 应用方向小波对各个子区域进行不同尺度的小波变换, 进而达到在确保边缘准确定位并尽可能去除由于噪声以及图像灰度不均匀产生的伪边缘点. 仿真实验表明, 对于受高斯白噪声污染的图像, 本文所提出的边缘检测方法无论在边缘定位的准确性还是在去除伪边缘点方面, 均优于传统的小波边缘检测方法.

参考文献

1 李惠光,孙昌平.基于变差函数的噪声图像的多尺度边缘检测.光电工程,2007,34(9):108−114.

2 周杰,彭嘉雄,丁明跃.方向小波变换及其在运动弱目标检测中的应用.信息与控制,1996,25(1):21−27.

3 杨正远,郑建宏.方向小波在图像边缘提取中的应用.重庆邮电学院学报,1997,9(3):16−20.

4 付丽华,陈涛,李落清.基于方向小波变换的边缘检测.湖北大学学报,2003,25(2):95−99.

5 Journal A, Huijbregts C. Mining Geostatistics. London: Academic Press, 1978. 21−26.

6 李钟山.地质统计学中的区域化变量理论.世界地质,1997, 16(2):85−93.

7 吴刚,杨敬安,王洪燕.一种基于变差函数的纹理图像分割方法.电子学报,2001,29(1):44−46.

8 Jupp DLB, Strahler AH, Woodcock CE. Auto-correlation and regularization in digital images I Basic theory. IEEE Trans. on Geoscience and Remote Sensing, 1988,26(4): 463−473.

9 Jupp DLB, Strahler AH, Woodcock CE. Auto- correlation and regularization in digital images.II. Simple image models. IEEE Trans. on Geoscience and Remote Sensing, 1989,27(3): 247−258.

10 Lei Z, Paul B. Edge detection by scale multiplication in wavelet domain. Pattern Recognition Letters, 2002, 23(14): 1771−1784.

11 Shih M, Tseng D. A wavelet-based multiresolution edge detection and tracking. Image and Vision Computing, 2005, 23(4):441−451.

(上接第 197 页)

参考文献

1 冯玉才,蒋涛,李国徽,朱虹.高效时序相似搜索技术.计算机学报,2009,3:2107−2122.

2 潘定,沈钧毅.时态数据挖掘的相似性发现技术.软件学报,2007,18(2):246−258.

3 Agrawal R, Faloutsos C, Swami A. Efficient Similarity Search In Sequence Database. Proc. of the 4th International Conference on Foundations of Data Organization and Algorithms. 1993. 69−84.

4 Chan KP, Fu AWC. Efficient time series matching by wavelets. Proc. of 15th IEEE Int. Conf. on Data Enfineering. 1999. 126−133.

5 Korn F, Jagadish H, Falous C. Efficiently supporting and hoc queries in large datasets of time sequences. Proc. of the ACM SIGMOD International Conferences. 1997. 289 −300.

6 Keogh EJ, Chakrabarti K, Pazzani M, Mehrotra S. Dimensionality reduction for fast similarity search in large time series databses. Knowledge and Information Systems, 2001, 263−286.

7 Ahmed N, Natarajan T, Rao KR. Discrete cosine transform. IEEE Trans. on Computers, 1974. 90−93.