一种基于 DTW 的符号化时间序列聚类算法 *

李 迎
(辽宁师范大学 计算机与信息技术学院，辽宁 大连 116081)

摘 要：提出了一种基于 DTW 的符号化时间序列聚类算法，对降维后得到的不等长符号时间序列进行聚类。该算法首先对时间序列进行降维处理，提取时间序列的关键点，并对其进行符号化；其次利用 DTW 方法进行相似度计算；最后利用 Normal 矩阵和 FCM 方法进行聚类分析。实验结果表明，将 DTW 方法应用在关键点提取之后的符号化时间序列上，聚类结果的准确率有较好大提高。

关键词：时间序列；DTW；SAX；Normal 矩阵；FCM
中图分类号：TP311.131 文献标识码：A 文章编号：1674-7720(2011)18-0003-03

Symbolization time series clustering based on DTW

Li Ying
(Department of Computer and Science Technology, Liaoning Normal University, Dalian 116081, China)

Abstract: A method of clustering symbolization time series based on DTW is proposed to cluster the unequal dimensional time series obtained by reduction. The key points of the time series are firstly extracted and symbolized. Then the similarity between the two time series is calculated by DTW method. Lastly, the normal matrix and FCM algorithm are employed to cluster the time series. The experimental results show that the accuracy of cluster result obtained by the proposed method is good.

Key words: time series; DTW; SAX; normal matrix; FCM

时间序列(Time Series)挖掘是数据挖掘中的一个重要研究分支，有着广泛的应用价值。近年来，时间序列挖掘在宏观的经济预测、市场营销、客流量分析、太阳黑子数、月降水量、河流流量、股票价格变动等众多领域得到了广泛应用[1]。

时间序列的相似性是衡量两个时间序列相似程度的一个重要指标，它是时间序列聚类、分类、异常发现等诸多数据挖掘的基础，也是研究时间序列挖掘的核心问题之一[2]。欧氏距离(Euclidean)和动态时间弯曲距离(Dynamic Time Warping)是计算时间序列相似性时经常被采用的两种度量方式。欧氏距离对时间轴上的轻微变化非常敏感，一些轻微的变化可能使欧氏距离的变化很大，而动态时间弯曲距离可以有效地消除欧氏距离这个缺陷，动态时间弯曲可以广泛应用在自然科学、医学、企业和经济等方面[3]。SAX(Symbolic Aggregate Approximation)[4]是一种运用符号化方法对时间序列进行表示、维度约简及相似性度量的方法。但 SAX 方法采用 PAA 算法将时间序列平均划分，不能很好地计算序列之间的相似度。而利用均分点和关键点对序列进行分段，既考虑了序列本身概率分布的变化，又兼顾到序列形态的变化。

本文提出一种基于 DTW 的符号化时间序列聚类算法，在提取关键点之后，再进行符号化时间序列，以达到降维的目的。降维之后得到的符号序列为不等长序列，采用动态时间弯曲距离(DTW)方法进行计算，鲁棒性好。然后通过 DTW 得到的距离矩阵构建复杂网络，并寻找其社团结构，实现了符号时间序列聚类。本文用 DTW 方法进行相似性度量比 KPDIST[4]在聚类结果的准确率上有较好大提高。

1 相关知识

1.1 时间序列关键点的选取

基于参考文献[5]可知，时间序列中的极值点 EP 成为关键点 KP 的条件为：

条件 1. x_i 保持极值的时间段与该序列长度的比值必须大于某个阈值 C；

条件 2. 若条件 1 不满足，则包含 x_i 的最小序列模

* 基金项目：国家自然科学基金(10771092)；辽宁省博士启动基金(20081079)

式<x_{i-1},x_i,x_{i+1}>中，三点连线形成的夹角小于筛选角度 α_0。

1.2 DTW 算法

动态时间弯曲方法公式如下[3]：

设有两个时间序列 Q 和 C，长度分别为 m、n，$Q=q_1,q_2,\cdots,q_i,\cdots,q_m'$，$C=c_1,c_2,\cdots,c_i,\cdots,c_n'$，$DTW(Q,C)=\min\{\sqrt{\sum_{k=1}^{K}\omega_k/K}\}$，其中，$\omega_k$ 是两个序列对应点 q_i、c_j 的距离 $d(q_i,c_j)$、K 为较长序列的长度，通常采用递归的方法计算。从两个序列的起始点开始递归计算两个序列 i、j 两点的 DTW 距离，$r(i,j)=d(i,j)+\{r(i-1,j-1),r(i-1,j),r(i,j-1)\}$。

1.3 基于 Normal 矩阵的谱平分法[6]

将一个时间序列作为一个节点，如果两个时间序列间的相似度大于给定的阈值，则认为这两个节点有边相连，否则它们之间就没有边。这样就构造了时间序列间的一个复杂网络 G。对于网络 G，有其邻接矩阵 $\boldsymbol{A}$。利用基于 Normal 矩阵的谱平分方法可以实现复杂网络的社团划分。

2 本文算法实现

2.1 关键点提取

输入：时间序列 $X=<(t_1,x_1),\cdots(t_i,x_i),\cdots,(t_n,x_n)>(0<i<n)$筛选夹角 α_0，预设数据压缩率 p；

输出：关键点集合 $KPS=<KP_1,\cdots,KP_i,\cdots,KP_n>$

(1)根据推论 1，由 p 计算系数 x；

(2)初始化，$KP_1=x_1,x_N,x_{2N},\cdots,x_{\omega N},\omega$ 是均分段数，N 是每个平均分段内数据的个数；

(3)从 $KP_1=x_1$ 开始判断时间序列的单调性，获得包含 3 个极值点 x_{i-p},x_i,x_{i+q} 的局部时间序列 $X=<x_{i-p},\cdots,x_{i-1},x_i,x_{i+1},\cdots,x_{i+q}>$，待考察的极值点为 x_i，包含该点的最短时间序列为<x_{i-1},x_i,x_{i+1}>；

(4)计算 $\max(|x_i-x_{i-1}|,|x_i-x_{i+1}|)$，假设返回$|x_i-x_{i+1}|$；

(5)若 $\frac{2|x_i-x_{i+1}|}{|x_i-x_{i+1}|^2-1}>\mathrm{tg}\alpha_0$，则 x_i 一定不是关键点，返回(1)，对下一个极值点进行判断；

(6)若 $\frac{1}{|x_i-x_{i-1}|}>\frac{|x_{i+1}-x_i|\mathrm{tg}(\alpha_0)-1}{|x_{i+1}-x_i|\mathrm{tg}(\alpha_0)}$ 则 x_i 一定不是关键点。返回(1)，对下一个极值点进行判断；

(7)将点 x_i 并入集合 KPS，更新区间$[u-x\sigma,u+x\sigma]$；返回(1)，对下一极值点进行判断。

2.2 基于 DTW 的符号化聚类算法

输入：时间序列集。

输出：聚类结果。

(1)对每个序列，运用上面的算法得到最终的关键点序列；

(2)计算序列 C 在各区间$[KP_{ci},KP_{cj})$内的均值，并表示为符号序列；

(3)对序列 C 和序列 Q 的符号序列进行相似性距离计算(DTW 计算和 KPDIST 计算)；

(4)根据相似度，构建复杂网络 G；此处要给相似度赋予一个阈值，相似性小于阈值的点则认为无边连接。

(5)用 Normal 矩阵方法 FCM 算法对复杂网络 G 进行社团划分，得到聚类结果。

3 实验结果与分析

本文实验采用 Keogh 博士的 Synthetic Control 和 ECG 数据集。实验环境为 2.66 GHz CPU Pentium® 4 PC 机，1 GB 内存，操作系统为 Windows XP Professional。算法实现软环境为 matlab 7.0 和 VC++6.0。Synthetic Control 数据集的实验数据为 300 条，每条时间序列长度为 60。ECG 数据集有 100 个样本序列，每条时间序列长度为 96(http://www.cs.ucr.edu/~eamonn/time_series_data/)。原时间序列维度为 60 和 96，经过关键点提取、符号化之后，维度大大降低，这为后期处理带来了很大的方便。在本实验中，关键点提取时筛选角度为 45°，预设的压缩率为 80%，划分了 4 个区间段，用符号表示时为 a,b,c,d 四种字母。由于实验数据的样本个数很多，这里只显示 synthetic control 的部分实验结果。表 1 为降维后的前 4 个符号序列实验结果。

表 1 Synthetic Control 序列 1-4 KP_SAX 字符串结果

1-10	11-20	21-30	31-40	41-50	51-60
bc	ac	bb	d	cc	acc
bc	bb	c	c	dd	d
bb	bc	ba	bb	c	dac
cc	d	dc	ab	bb	ac

表 2 为 Normal 矩阵得到的非平凡特征值对应的非平凡特征向量，根据谱平分算法思想，同一社团内的节点相应的元素 x_i 非常接近。从特征向量的分析中可以看出，将 DTW 与复杂网络知识应用在符号化时间序列上是一种较好的创新。

由 DTW 距离矩阵得到的网络中，第一非平凡特征值取值为：0.252 9，而通过 KPDIST 距离矩阵得到的复杂网络中，第一非平凡特征值取值为：0.125 7，从特征值中就可以初步判断，DTW 得到的特征值更为准确，这两个特征值对应的特征向量的区间表如表 2 所示。

表 3 为两种算法对同样数据集进行聚类得到的结果。数据集 Synthetic control 采用本文方法正确率为

表 2 synthetic control 的特征向量分布表

基于 DTW 得到的特征向量		基于 KPDIST 得到的特征向量	
向量区间	正确个数	向量区间	正确个数
(-0.05~-0.022 3]	42	(-0.643 4~-0.430 7]	37
(-0.022 3~-0.019]	39	(-0.430 7~-0.029 7]	40
(-0.019~-0.017]	32	(-0.029 7~0.072 5)	30
(-0.01~-0.016)	40	(0.08~0.095]	35
(0.04~0.052 5]	37	(0.095~0.15]	33
(0.052 5~0.054]	39	(0.15~0.348 1]	32

表 3 聚类结果

数据集	方法					
	本文算法			KPDIST		
	正确样本个数	错误样本个数	正确率/%	正确样本个数	错误样本个数	正确率/%
Synthetic control	229	71	76.3	207	93	69
ECG	72	28	72	65	35	65

76.3%。而利用 KPDIST 算法正确率为 69%；数据集 ECG，本文的正确率为 72%，KPDIST 的正确率为 65%。

SAX 是一种符号化的时间序列相似性度量方法，该方法在对时间序列划分时，采用了 PAA 算法的均值划分，得出的结果不能精确地表示出原时间序列，故将关键点提取方法与 PAA 方法相结合，在对原序列降维的同时又能更准确地表示原时间序列。本文将复杂网络知识和时间序列降维方法相结合，给出了一种时间序列的聚类方法。该算法用 DTW 算法计算时间序列间的相似度，而后从时间序列的相似度得到一个复杂网络，此复杂网络表示了时间序列相互间的关系。最后采用 Normal 矩阵的方法进行网络划分，得到一个网络的社团结构。从这个社团结构中已能看出样本时间序列的归属类别，但为了结果更加清晰，用具体数字来体现，所以采用了 FCM 聚类算法进行最后的聚类。实验结果表明，用 DTW 方法计算序列之间的相似度结合在降维后的符号化时间序列上比原文 KPDIST 方法在准确率上有较好大提高。

参考文献

[1] 毛国君，段立娟，王实，等.数据挖掘原理与算法(第二版)[M].北京：清华大学出版社，2007.

[2] 刘懿，鲍德沛，杨泽红.新型时间序列相似性度量方法研究[J].计算机应用研究，2007，24(5)：112-114.

[3] KEOGH E, RATANAMAHATANA C A. Exact indexing of dynamic time warping[J]. Springer-Verlag London Ltd, 2005, 10.1007/s10115-004-0154-9:358-386.

[4] 闫秋艳，孟凡荣.一种基于关键点的 SAX 改进算法[J].计算机研究与发展，2009，46(z2)：483-490.

[5] 杜奕.时间序列挖掘相关算法研究及应用[D].合肥：中国科学技术大学，2007.

[6] 汪小帆，李翔，陈关荣.复杂网络理论及其应用[M].北京：清华大学出版社，2006：169-171.

(收稿日期：2011-05-17)

作者简介：

李迎，女，1986 年生，硕士，主要研究方向：人工智能，数据挖掘。