一种基于趋势的时间序列相似性模型*

敖环环 陈帅飞 杨少松
(河海大学计算机与信息学院 南京 211100)

摘 要：针对传统基于欧氏距离的时间序列相似性模型，搜索相似时间序列子段效率低、形态不完整的问题，本文提出了一种基于趋势的时间序列相似性模型。该模型以时间序列分段线性表示和符号化为基础，通过计算符号串的编辑距离得出时间序列在形态上的差异性。同时该模型把时间序列长度和时间序列变化幅度单独考虑，扩大了该模型的使用范围。实验表明，本文提出的基于趋势的时间序列相似性模型，在相似性匹配方面，搜索效率高于传统基于欧氏距离时间序列相似性模型；同时，该模型侧重于形态方面的相似性，所以对时间轴方向的偏移和白噪声不敏感，适用于不等长时间序列的相似性匹配。

关键词：时间序列；数据挖掘；相似性；趋势；线性表示
中图分类号：TP311 **文献标识码**：A **国家标准学科分类代码**：520.20

Similarity model based on trend of time series

Ao Huanhuan Chen Shuaifei Yang Shaosong
(College of Computer and Information, Hohai University, Nanjing 211100, China)

Abstract: In view of the problem that the time series similarity model based on Euclidean distance is less efficient and incomplete in morphology, the paper presents a similar model based on trend for time series. The model is based on piecewise linear representation and symbolic representation of time series, and the difference of time series is obtained by calculating the edit distance of the symbol string. At the same time, the model considers the time series length and the time series variation separately, expanding the use of the model. The experimental results show that the similarity model of time series based on trend is suitable for the similarity matching, and the efficiency is higher than traditional model which is based on Euclidean distance. Since the model emphasized on the similarity of morphology, it's not sensitive to the deviation of time axis and white noise, and it's suitable for the similarity matching of time series with different length.

Keywords: time series; data mining; similarity; trend; linear representation

1 引 言

时间序列的相似性[1-2]至今为止还没有一个统一的定义，如何度量时间相似性是一个带有很强主观性的问题，相似性的判定都是建立一定的模型的基础上，因为时间序列的相似性依赖的因素是多方面的，不仅依赖于用户需求，而且还依赖于目前任务的目的、数据集本身等。例如，在金融领域中，相似性通常是指序列的变化模式相同，而不是根据每个点的数值判定；但是在医学领域中，即使是两个病人的心脏跳动情况相似，但振幅差异很大，医生也认为它们不相似；还有在语音数据识别中，如果两个人说了同一个词，虽然时间长度有偏差、振幅强弱也不同，但我们依然认为它们是相似的。目前时间序列相似性的度量主要有基于欧氏距离度量[3]（euclidean），基于动态时间弯曲距离度量[4-5]（dynamic time warping），以及最大公共子序列[6]（longest common subsequence）等。但是采用这些距离度量时间序列存在着很多难以克服的问题。

各种距离度量的优缺点：欧氏距离以计算简单、通用性好著称，也可以应用到聚类[7]和分类[8]等研究领域，但最大不足之处在于只能处理等长度的时间序列，对振幅变化敏感；动态时间弯曲距离较好地克服了欧氏距离的不足，支持时间序列的动态弯曲，但该方法计算复杂、时间复杂度高，限制了其应用范围；最长公共子序列要求太高，对某一局部的变化比较敏感，会导致相似性被错误的估计。基于各种

收稿日期：2016-03
*项目基金：国家科技支撑计划(2013BAB06B04)、江苏水利科技(2013025)资助项目

距离度量的不足,本文选择编辑距离作为时间序列的相似性距离度量标准。

本文提出了一种基于趋势的时间序列相似性模型,通过基于重要点的分段线性表示方法对其拟合,进而依据每个时间子序列的变化趋势来判断相似性,从而可以避免对每个数据点的过分依赖,然后结合每个子段的斜率和时间跨度来对拟合时间离散化映射到一个二维的向量空间,再按照二维向量的两个分量的值对其近似符号化(SAX),最后计算对应时间序列符号串的编辑距离,依据相似性计算公式输出相似度。根据大量的实验表明,本文提出的基于趋势的时间序列匹配模型计算简单、实现方便,搜索效率高于传统基于欧氏距离时间序列相似性模型,且搜索出来的相似时间序列子段在形态上都是完整的。另外,该模型侧重于形态方面的相似性,所以对于时间轴方向的偏移不敏感、对于白噪声不敏感,同时该模型适用于不等长时间序列的相似性计算。

本文的其余部分组织如下:第二部介绍了基于重要点的时间序列线性表示方法;第三部分,提出了一种基于时间序列二维向量的符号化表示方法;第四部分,提出了基于趋势间序列相似性模型;第五部分,对模型进行实验论证;第六部分,总结全文。

2 时间序列的线性表示

时间序列特征的提取方法有很多种,例如文献[9]通过斜率来近似表示原时间序列。本文选取了基于重要点的分段线性表示[10]策略来对原始时间序列进行预处理,通过选取重要点的方式对原始序列进行分段线性表示,可以提取出时间序列的主要特征,去除细节干扰,只保留原始时间序列中的主要形态特征,有利于提高数据挖掘的效率和准确性[11]。

下面对重要点及分段线性表示概念进行描述。

定义 1(重要点):假设有时间序列 X,$< a_i, a_{i+1}, \cdots, a_j >$ 是时间序列 X 的子序列,当点 a_m 满足以下条件之一时,则 a_m 为重要点,其中 $i < m < j$

1) a_m 是极小值,同时 $a_i/a_m \geqslant R, a_j/a_m \geqslant R$

2) a_m 是极大值,同时 $a_i/a_m \leqslant R, a_j/a_m \leqslant R$

R 为选取重要点时的指定阈值,其取值横大于 1。R 值越大,选取的分段点越少,则提取的整体趋势特征越显著;反之,局部趋势特征越明显。

定义 2(时间序列分段线性表示):设有时间序 $X = < x_1, x_2, \cdots, x_n >$,分段点的集合是 $X'_i = < x'_1, x'_2, \cdots x'_m >$,其中 $x'_1 = x_1, x'_m = x_n, m < n$,

则 X 的分段线性表示为:

$$X_L = < f_1(x'_1, x'_2), f_2(x'_2, x'3), \cdots, f_{m-1}(x'_{m-1}, x'_m) > \quad (1)$$

其中 $f_{m-1}(x'_{m-1}, x'_m)$ 表示在区间 $[x'_{m-1}, x'_m]$ 内的线性拟合函数。

表 1 符号化规则

时间跨度 / 变化趋势	大于 Q	小于等于 Q
剧烈上升	A	B
缓慢上升	C	D
剧烈下降	a	B
缓慢下降	C	d

3 时间序列符号化

通过提取重要点对原始时间序列进行线性拟合,原始时间序列的主要特征被保留下来,呈现出一系列的上升下降趋势,为了对时间序列之间的相似性进行定量评估,需要对拟合后的时间序列进行离散化和符号化处理。根据时间序列上升和下降的趋势不同,可以采用不同精细程度的划分。假定升降变化趋势划分为四个等级,跨度划分为两个等级,则对应的符号化规则如表 1 所示。

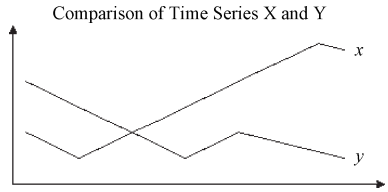


图 1 忽略时间跨度

4 基于趋势的时间序列相似性模型

4.1 相关定义

定义 3(编辑距离):编辑距离[12]是指两个字符串 S_1 和 S_2,由 S_1 转成 S_2 所需最少编辑操作次数 $D_{edit}(S_1, S_2)$。允许的三种编辑操作包括插入一个字符、删除一个字符、将一个字符替换成另一个字符。

定义 4(时间序列相似性度):假定有时间序列 $X = < x_1, x_2, \cdots, x_n >$ 和 $Y = < y_1, y_2, \cdots, y_n >$,两个时间序列的相似度计算公式如下:

$$sim(X,Y) = \left(1 - \frac{D_{edit}}{\max(L(}S_X), L(S_Y))\right) * \left(\frac{\min(L(X), L(Y)}{\max(L(X), L(Y)}\right) * \left(\frac{\min(\Delta_{Amp}(X), \Delta_{Amp}(Y)}{\max(\Delta_{Amp}(X), \Delta_{Amp}(Y)}\right) \quad (2)$$

其中 $L(X), L(Y)$ 表示原始时间序列的长度,$\Delta_{Amp}(X)$,$\Delta_{Amp}(Y)$ 表示原始序列振幅跨度,$L(S_X)$,$L(S_Y)$ 表示映射后的符号串的长度。根据应用场景不同,公式中的振幅跨度是可选项。

4.2 基于趋势的时间序列相似性模型算法

算法步骤如下:

输入:查询时间序列 X,被查询时间序列 Y

输出:时间序列 X 与 Y 的相似度 $sim(X,Y)$

第一步：对时间序列 X 和 Y 进行规范化处理，通过规范化处理可以将原始时间序列的值映射到[0,1]之间；

第二步：用基于重要的的分段线性表示方法对时间序列进行线性表示，得到原始时间序列的分段线性表示：

$X_L = < f_1(x'_1, x'_2), f_2(x'_2, x'3), \cdots, f_{m-1}(x'_{m-1}, x'_m) >$；

第三步：对时间序列线性拟合函数进行符号化，首先把线性拟合函数的每一个子段映射到一个二维向量，然后将时间序列的变化趋势划分为 M 个等级，将时间跨度划分为 N 个等级，这样整个时间序列对应的二维向量将映射到 $M \times N$ 种字符表示的字符串，对两个时间序列分别按对应规则进行符号化，从而得到对应趋势串 S_X 和 S_Y；

第四步：计算趋势串的编辑距离，得到 $D_{edit}(S_1, S_2)$，根据公式 2 计算两个时间序列的相似度 $sim(X,Y)$，并输出结果。

5　实验验证

5.1　实验目的及评价指标

本实验的主要目的在于验证本文提出的基于趋势的时间序列相似性模型在搜索相似模式时间序列子段的优势。实验主要的评价指标为：1)搜索相似时间序列模式子段所耗费的时间；2)搜索出来的时间序列子段形态的完整性。

5.2　实验数据

本文中的时间序列数据集取自于 www.cs.ucr.edu/~eamonn/tutorials.html 公布的用于数据挖掘的通用时间序列数据集(本文简称为 KData)，如表 2 所示。

表 2　KData 数据集

序列名称	序列长度	序列名称	序列长度
Burst	9382	Memory	6875
Chaotic	1800	Ocean	4096
Fluid_dynamics	10000	Earthquake	2501
Earthquake	4096	Speech	1020

5.3　实验方法

为了将基于欧氏距离的相似性模型和本文提出的基于趋势的时间序列相似性模型进行对比，本实验分别对每个时间序列随机抽取一个双峰子段作为模板序列，并分别设定子段相似性的阈值，搜索与其相似的时间序列子段。实验中设定本文相似性模型分段表示的阈值 R 为 1.05，相似度阈值设为 70%，得到的结果如表 3 所示。基于欧氏距离的相似性模型相似度阈值设为 0.04(每个点幅度的平均偏移量)，得到的相似时间序列子段统计结果如表 4 所示，图 2 展示了两种模型搜索过程的时间消耗。

表 3　本文模型相似时间序列子段搜索结果

序列名称	相似子段个数	序列名称	相似子段个数
Burst	11	Memory	0
Chaotic	8	Ocean	0
Fluid_dynamics	2	Powerplant	0
Earthquake	9	Speech	5

表 4　欧氏距离模型相似时间序列子段搜索结果

序列名称	相似子段个数	序列名称	相似子段个数
Burst	38	Memory	0
Chaotic	2	Ocean	0
Fluid_dynamics	17	Powerplant	15
Earthquake	48	Speech	0

5.4　实验结果及分析

根据图 2 的实验结果分析可知，在 10 个不同领域的数据集上，本文提出的时间序列相似性模型搜索相似时间序列子段所消耗的时间比传统欧氏距离模型的要小。之所以如此，是因为本文的时间序列相似性模型采用了基于趋势的符号化策略，将时间序列数据集进行了压缩，大大提高了搜索效率。

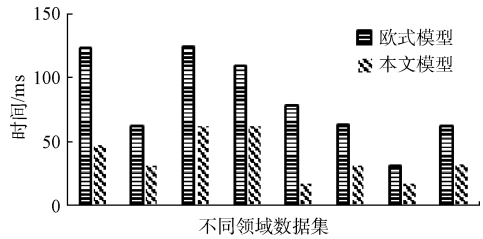
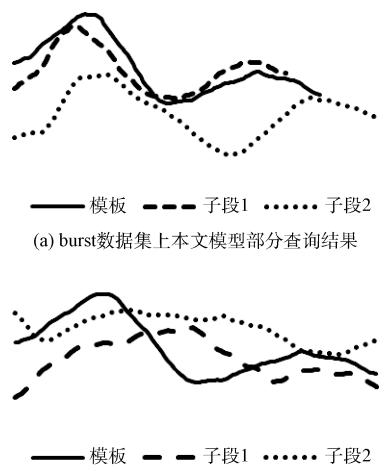


图 2　搜索相似时间序列子段消耗时间对比

为了比较两种模型搜索出来结果的差异，从实验的结果中选取 Burst 的相似时间序列子段的部分结果进行展示对比，如图 3 所示。



(a) burst数据集上本文模型部分查询结果

(b) burst数据集上欧氏距离模型部分查询结果

图 3　两种模型查询结果对比

根据图 3 的实验结果分析可知，基于趋势的时间序列

相似性模型可以搜索出完整形态的相似时间序列子段,而基于传统欧氏距离的相似性模型搜索出来的相似性时间序列子段长度是严格相等的,在形态上是不完整的。

另外,从图3的结果还可以得出如下结论:1)基于欧氏距离的时间序列相似性模型侧重于时间序列在空间上的接近程度,而本文提出的基于形态的时间序列相似性模型侧重于时间序列在形态方面的相似;2)另外本文提出的相似度模型判断两个时间序列相似性,对噪声数据不敏感,对时间轴的线性漂移不敏感,同时该模型适用于不等长时间序列相似性比较。

6 结 论

本文提出的基于趋势的时间序列相似性模型,以时间序列分段线性表示和符号化为基础,可以快速提取出时间序列的主要形态同时又将时间序列大幅度压缩,从而大大提高了搜索相似模式子段的效率,而且搜索出来的相似模式子段在形态上是完整的。另外,由于本文提出的时间序列相似性模型是以时间序列趋势的来比较相似性,所以对时间序列的局部变化不敏感,对时间轴上的偏移也不敏感。同时该模型也适用于不等长时间序列的相似性比较。下一阶段的研究重点是考虑如何更加合理的对时间序列进行趋势符号化表示,提高基于趋势的时间序列相似性匹配的精度。

参考文献

[1] SERRÀ J, ARCOS J L. Particle swarm optimization for time series motif discovery[J]. Knowledge-Based Systems, 2016(92): 127-137.

[2] LIU B, LI J, CHEN C, et al. Efficient motif discovery for large-scale time series in healthcare[J]. IEEE Transactions on Industrial Informatics, 2015, 11(3): 583-590.

[3] 谢福鼎, 李迎, 孙岩,等. 一种基于关键点的时间序列聚类算法[J]. 计算机科学, 2012, 39(3):157-159.

[4] IZAKIAN H, PEDRYCZ W, JAMAL I. Fuzzy clustering of time series data using dynamic time warping distance[J]. Engineering Applications of Artificial Intelligence, 2015, 39: 235-244.

[5] PETITJEAN F, FORESTIER G, WEBB G I, et al. Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm[J]. Knowledge and Information Systems, 2015: 1-26.

[6] ARSLAN A N. Fast algorithms for local similarity queries in two sequences[J]. International Journal of Foundations of Computer Science, 2015, 26(5): 625-642.

[7] IORIO C, FRASSO G, D'AMBROSIO A, et al. Parsimonious time series clustering using p-splines [J]. Expert Systems with Applications, 2016.

[8] LINES J, BAGNALL A. Time series classification with ensembles of elastic distance measures[J]. Data Mining and Knowledge Discovery, 2015, 29(3): 565-592.

[9] 洋洋, 陈小惠, 王保强,等. 脉搏信号中有效信号识别与特征提取方法研究[J]. 电子测量与仪器学报, 2016, 30(1): 126-132.

[10] BANKÓ Z, ABONYI J. Mixed dissimilarity measure for piecewise linear approximation based time series applications[J]. Expert Systems with Applications, 2015, 42(21): 7664-7675.

[11] JIN L I, XING F, SUN T, et al. Space high-accuracy intelligence payload system with integrated attitude and position determination[J]. Instrumentation, 2015 (1).

[12] CHATTERJEE K, HENZINGER T A, IBSEN-JENSEN R, et al. Edit distance for pushdown automata [M]. Automata, Languages, and Programming. Springer Berlin Heidelberg, 2015: 121-133.

作者简介

敖环环,硕士研究生,主要研究方向为数据挖掘。
E-mail:aohuanhuanhhu@163.com

陈帅飞,硕士研究生,主要研究方向为时间序列的表示方法、时间序列的相似性搜索。
E-mail:chenshuaifei163@163.com

杨少松,硕士研究生,主要研究方向为数据挖掘。
E-mail:489271346@qq.com