文章编号:1673-9469(2015)03-0070-05 doi:10.3969/j.issn.1673-9469.2015.03.017

基于模式形态距离的桥梁监测测点相似性聚类研究

屈 兵,肖汝诚,郝佳佳
(同济大学 土木工程学院,上海 200092)

摘要:为挖掘桥梁检测各测点之间的相似关系,提出基于模式形态距离的时间序列相似性度量方法。该方法首先根据监测时间序列的形态特征将序列划分成若干模式,然后以各模式形态的动态变化趋势差异为依据进行相似性的判别,并定义了各类判别结果的距离函数,最后得出各测点间的模式形态距离。在此基础上,对玉峰大桥监测点的相似性进行层次聚类分析,结果反映出的信息与桥梁的真实结构状况一致。监测点的相似性分析为桥梁结构提供了更深层次信息挖掘的可能,为传感器的坏点排查以及结构的异常数据判别提供了科学的依据。

关键词:相似性;桥梁工程;健康监测;模式形态距离;聚类分析;玉峰大桥
中图分类号:TU997 **文献标识码**:A

Bridge monitoring points similarity clustering analysis based on pattern-shape distance

QU Bing, XIAO Ru-cheng, HAO Jia-jia
(College of Civil Engineering, Tongji University, Shanghai 200092, China)

Abstract: The bridge monitoring system consists of a large amount of sensors, and there are likely to be some similarities among the data collected by different sensors. To figure out the similarity relations, a pattern-shape distance based similarity measurement approach was proposed. Firstly, the monitoring time series was divided into several patterns according to the morphological characteristics. Secondly, similarity discrimination based on the difference of dynamic change trend of each pattern shape was discussed, and the distance function of each discrimination result was defined. Thus, the pattern shape distance of different monitoring points could be calculated. On this basis, similarity analysis of monitoring points on Yufeng Bridge was discussed by means of hierarchy clustering. The clustering result shows good consistence with bridge real states. The similarity analysis of monitoring points provides the potential of deeper information mining of bridge structures and scientific basis in sensor troubleshooting and discrimination of abnormal data.

Key words: similarity; bridge engineering; health monitoring; pattern-shape distance; cluster analysis; Yufeng Bridge

在外荷载作用下,桥梁结构各构件之间或同一构件不同位置处的受力状态可能具有一定的相似性,并通过不同测点的监测数据反映出来。对这些数据的挖掘有利于从整体上把握各测点的监测状态,并为监测数据的异常判别以及传感器坏点排查提供重要的依据。桥梁监测序列为典型的时间序列,时间序列的相似性度量是相似性搜索问题中的关键所在。经研究,欧氏距离、动态时间弯曲距离、模式距离、编辑距离等在不同应用背景中均可作为度量相似性的有效方法[1-4]。其中,模式距离侧重于对数据变化趋势形态的一致性度量,因而更加适应监测序列相似性分析的需求。“模式距离”的概念由王达等[5]首次提出,他将模式定义为三元集,但其粒度较粗糙,得出的结果不

收稿日期:2015-04-15
基金项目:国家重点基础研究发展计划973项目(2013CB036300);国家自然科学基金资助项目(51378387)
作者简介:屈兵(1988-),男,湖南邵阳人,博士,主要从事桥梁监测与评估研究。

够明确;为此,丁勇伟等将模式的划分完全连续化,发展了基于弧度距离的相似性度量[6],但其条件要求过于严苛,不适应监测序列的相似性特点。本文在此基础上,结合桥梁监测数据的特征,提出基于模式形态距离的时间序列相似性度量方法。并对玉峰大桥监测系统的 44 个测点进行了相似性聚类分析,以期为桥梁结构更深层次的信息挖掘提供理论基础。

1 监测时间序列预处理

1.1 数据标准化

为消除各测点监测数据由于量纲与数量级不同带来的误差,确保其地位平等,需要在保持序列形态不变的前提下对数据进行标准化处理。

时间序列 $Y=y_1,y_2,\cdots,y_n$,序列的平均值和标准差分别为:

$$\mu_Y=\frac{1}{n}\sum_{i=1}^{n}y_i,\sigma_Y=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\mu_Y)^2}$$

设标准化变换后的序列为 Y^*,有

$$y_i^*=\frac{y_i-\mu_Y}{\sigma_Y} \tag{1}$$

变换后序列的均值为 0,标准差为 1。

1.2 缺失数据处理

进行相似性比较的各监测序列需满足齐序列要求,从而能对各对等时刻的数据进行比较。通过合理设定传感器的采集频率可达到时间的同步,数据非齐性只表现在某些时刻数据的缺失上。

对于缺失的数据,通过序列的截取,忽略缺失时间段内的数据影响,使序列齐化,如图 1 所示。

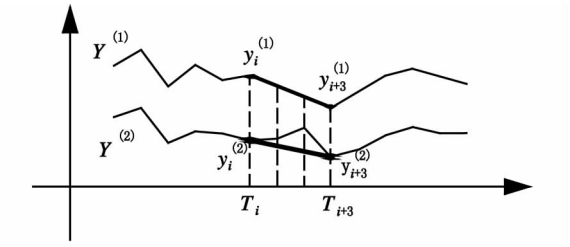

图1 缺失数据处理

Fig.1 Processing of missing data

图中,序列 $Y^{(1)}$ 在 T_i 到 T_{i+3} 内的数据是缺失的,在与 $Y^{(2)}$ 进行相似性比较时,剔除 $Y^{(2)}$ 在 T_i 到 T_{i+3} 内的数据,将 T_i 作为分段的起始时间,T_{i+3} 为终止时间,即把 $T_i^{(1)}$ 与 $Y_{i+3}^{(1)}$、$y_i^{(2)}$ 与 $y_{i+3}^{(2)}$ 当作相邻数据,然后进行比较。

2 基于模式形态距离的相似性度量

相似性度量是衡量两个序列相似性程度的依据,一般采用距离函数进行相似性的度量。给定两个时间序列 X 和 Y,如果满足 $d(x,y)\leqslant\varepsilon$,则说明两序列是相似的。其中 $d(\,)$ 是一个相似性距离度量函数,ε 是给定的相似性阈值[7]。

2.1 测点序列的选取

由于监测原始序列的数据量庞大,在进行测点相似性分析时,应提炼出能代表测点原始监测序列特征形态的特征序列进行分析。考虑到桥梁结构外荷载的周期性与随机波动性,并考虑年温差等长期效应的影响,选择各测点在每日 24 小时内监测数据的平均值组成的序列(以下简称日周期 M 值序列)作为特征序列,时间期取监测系统初始 1 年运营期。

2.2 模式形态距离

监测序列在相似性判定时,并不要求数据具有完全的一致性,只要数据的变化趋势基本一致即可认为相似。据此思想,令标准化处理后的序列为 $Y^*=\{(y_1^*,T_1),\cdots,(y_i^*,T_i),\cdots,(y_n^*,T_n)\}$,$y_i^*$ 为第 i 个周期 T_i 内的 M 值标准化后的值,则各时间单元内 y^* 值的变化趋势可用该单元直线段的斜率表示,见式(2)。认为相邻两单元的时间间隔为 1,即 $T_{i+1}-T_i=1$。

$$k_i=\frac{y_{i+1}^*-y_i^*}{T_{i+1}-T_i}=y_{i+1}^*-y_i^*\ (i=1,2,\cdots,n) \tag{2}$$

定义 $\hat{y}_i=(k_i,T_i)$ 为序列的第 i 个模式形态,用于描述序列 Y^* 在时间 T_i-T_{i+1} 内的形态特征。

对于两个待比较序列 $Y^{(1)}$ 和 $Y^{(2)}$,将序列各模式形态的相似性判别结果表示为四元集合:

$U=\{$完全相似,部分相似,不相关,相反$\}$

U 内各元素对应的量化距离集合为

$D=\{0,0.25,0.5,1\}$

其中,$ds_i\in D$,定义为 $\hat{y}_i^{(1)}$ 与 $\hat{y}_i^{(2)}$ 的模式形态距离。$\hat{y}_i^{(1)}$ 与 $\hat{y}_i^{(2)}$ 的模式形态距离可根据以下两重判别步骤进行。

2.2.1 第一重判别

根据各时间单元的斜率大小 $k_i^{(1)}$、$k_i^{(2)}$,将模式形态划分为两层次模型,如图 2 所示。

上升 ($k_i>0$)		水平 ($k_i=0$)	下降 ($k_i<0$)	
绝对上升 ($k_i>\varepsilon$)	水平上升 ($k_i\leqslant\varepsilon$)		水平下降 ($k_i\geqslant-\varepsilon$)	绝对下降 ($k_i<-\varepsilon$)

图2 模式形态的两层次模型

Fig. 2 Two-level model of pattern shape

其中,ε 为模式形态的区分阈值,可根据对序列相似性区分的容忍程度确定。一般可取各周期模式斜率绝对值均值的10%左右,见式(3)。

$$\varepsilon=\frac{1}{n-1}\sum_{i=1}^{n-1}|k_i|\times 10\% \tag{3}$$

其中,n 为总周期数;对于桥梁监测日周期 M 值序列,由于序列波动不大,ε 取值范围为0~0.1。

根据两层次模型的形态划分,可形成图3所示五类组合判别类型(A、B、C、D、E)。

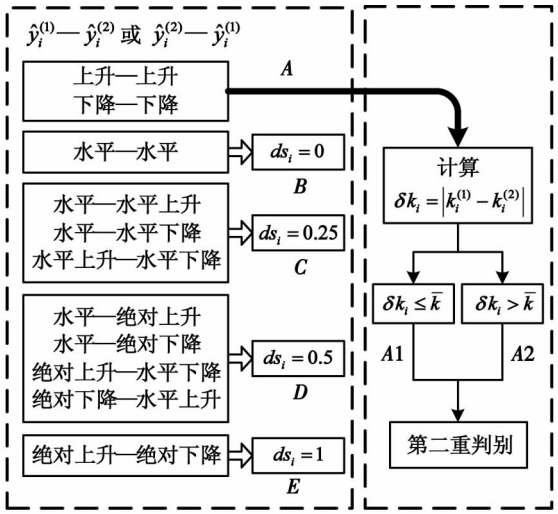

图3 模式形态相似性第一重判别

Fig. 3 The first discrimination of pattern-shape similarity

其中,δk_i 为 $\hat{y}_i^{(1)}$ 与 $\hat{y}_i^{(2)}$ 的斜率差绝对值;$\bar{k}$ 为各时间单元斜率绝对值的均值,有

$$\bar{k}=\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(|k_i^{(1)}|+|k^{(2)}|) \tag{4}$$

第一重判别中,当模式形态出现同上升或同下降的A情况时,可能形成图4所示的凹、凸曲线(y_1,y_2),两曲线不能认为完全相似。为体现此差异,进一步进行第二重判别。

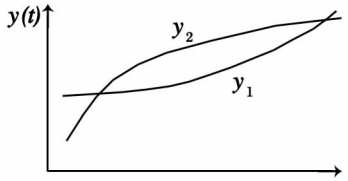

图4 两层次划分不能区分的两条曲线

Fig. 4 Two curves that two-lever model cannot distinguish

2.2.2 第二重判别

进行模式形态第二重判别前,首先计算当前时间单元与前一单元的斜率差:$\Delta k_i^{(1)}=k_i^{(1)}-k_{i-1}^{(1)}$,$\Delta k_i^{(2)}=k_i^{(2)}-k_{i-1}^{(2)}$。根据 Δk_i 的范围,将模式形态继续细分四个档次,各档次中当 k_i 值不同时,Δk_i 取值不同,模式分档见图5、表1。

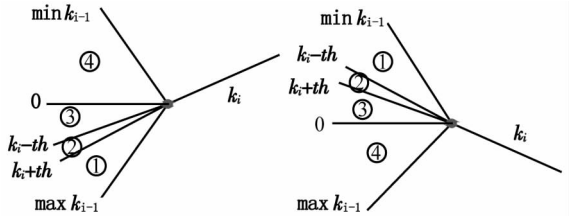

图5 斜率差模式分档示意图

Fig. 5 Classification of slope difference

表1 斜率差模式分档

Tab. 1 Classification of slope difference

编号	范围		描述
	$k_i>\varepsilon$	$0<k_i<\varepsilon$	
①	$\Delta k_i<-\varepsilon$		减速上升
②	$-\varepsilon\leqslant\Delta k_i\leqslant\varepsilon$	$-\varepsilon\leqslant\Delta k_i\leqslant k_i$	匀速上升
③	$\varepsilon<\Delta k_i\leqslant k_i$	—	加速上升
④	$\Delta k_i>k_i$		反转上升
	$k_i<\varepsilon$	$-\varepsilon<k_i<0$	
①	$\Delta k_i>\varepsilon$		减速下降
②	$-\varepsilon<\Delta k_i\leqslant\varepsilon$	$k_i<\Delta k_i\leqslant\varepsilon$	匀速下降
③	$k_i<\Delta k_i\leqslant-\varepsilon$	—	加速下降
④	$\Delta k_i<k_i$		反转下降

根据不同的 $\Delta k_i^{(1)}$、$\Delta k_i^{(2)}$,对图3中A1、A2情况进行模式形态相似性的第二重判别,见图6。

通过以上两重判别可以确定出 $\hat{y}_i^{(1)}$ 与 $\hat{y}_i^{(2)}$ 的模式形态距离 ds_i。从而得到序列 $Y^{(1)}$ 与 $Y^{(2)}$ 的模式形态距离为:

$$ds(Y^{(1)},Y^{(2)})=\frac{1}{n-1}\sum_{i=1}^{n-1}ds_i \tag{5}$$

式中,n 为总周期数,且满足 $ds(Y^{(1)},Y^{(1)})=0$,$ds(Y^{(1)},Y^{(2)})=ds(Y^{(2)},Y^{(1)})$,$ds(Y^{(1)},Y^{(2)})\in[0,1]$。$ds(Y^{(1)},Y^{(2)})$ 越接近0,两序列间的形态越接近;反之,形态差距越大。一般来说,当 $ds(Y^{(1)},Y^{(2)})<0.15$ 时,可认为两序列具有足够相似性。

2.3 反对称相似

当由式(5)算得两时间序列模式形态距离接近1时,不能简单地判定二者为非相似序列,还应考虑序列的反对称相似。取两种情况下的较小者

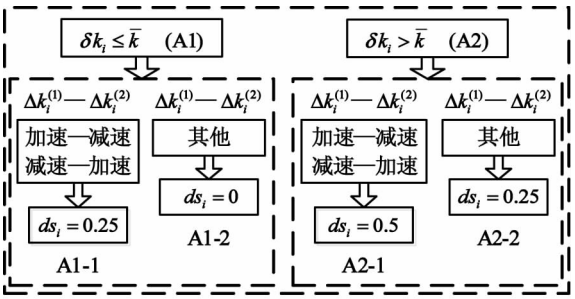

图6 模式形态相似性第二重判别

Fig. 6 The second discrimination of pattern-shape similarity

作为最终的模式形态距离,即:

$$\min[\, ds(Y^{(1)}, Y^{(2)}), ds(-Y^{(1)}, Y^{(2)}) \,] \qquad (6)$$

3 测点相似性聚类分析

聚类分析是一个将对象集划分为若干类的过程,同一个类间对象具有较高的相似性,不同类对象的相似性较小[8-9]。通过对桥梁监测测点进行聚类,将具有近似变化趋势的测点归为一类,为异常数据的判别提供指导。

3.1 测点层次聚类算法

(1)每个测点各成一类,计算每对测点序列间的模式形态距离,得到 n 个测点的距离矩阵 D_0。

(2)找出 D_0 中的最小距离元素,将其对应的两类合并成一个新类 C_K。增加对应于 C_K 的新行新列,并删除合并前两类的所在行和列。

(3)计算 C_K 与任意其他类 C_L 的距离 D_{KL},见式(7),将其作为新行新列上的元素,得到新的距离矩阵 D_1。

(4)对 D_1 重复步骤 2 和步骤 3,得到距离矩阵 D_2;如此重复下去,直到全部测点都合并为一类或满足给定聚类个数时终止。

类与类之间的距离采用平均距离法计算,设 $x_i(i=1,2,\cdots)$, $x_j(j=1,2,\cdots)$ 分别为类 C_K 与 C_L 类中的元素,d_{ij} 为 x_i 与 x_j 的模式形态距离,则类 C_K 与类 C_L 间的平均距离为:

$$D_{KL} = \frac{1}{n_k n_L} \sum_{x_i \in C_K, x_j \in C_L} d_{ij} \qquad (7)$$

3.2 测点聚类过程

玉峰大桥是一座无推力的斜靠式拱桥,主跨 110 m。其监测参数包括西南侧主拱肋(编号 1 -1 -4 -6,共 15 个测点)、斜拱肋(编号 5 -1 -7 -3,共 9 个测点)及主梁(编号 8 -1 -11 -4 共 16 个测点)的应力监测数据;主拱下三个可动支座(编号 w1、w2、w3)的位移监测数据以及环境温度监测数据(编号 tem),各测点的具体位置见图 7、图 8。

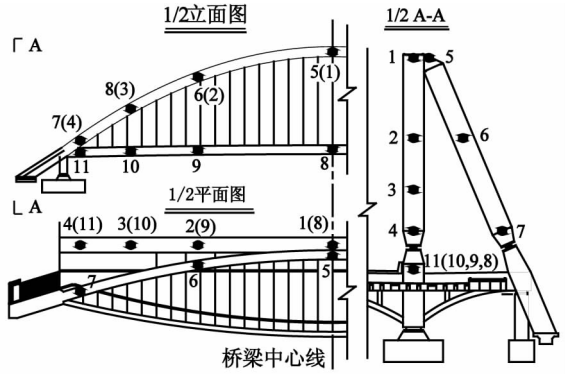

图7 测点总体布置图

Fig. 7 General layout of measuring points

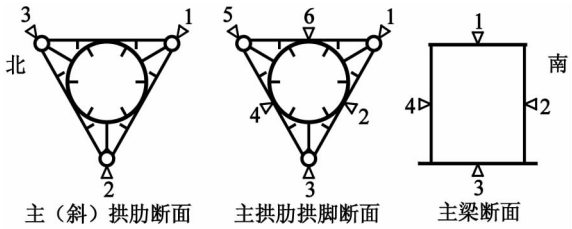

图8 测点断面布置图

Fig. 8 Sections layout of measuring points

拱肋断面中,除主拱拱脚(4 号)断面布置了 6 个测点外,其余均布置 3 个测点。

选用各测点一年期的日周期 M 值序列进行层次聚类分析,得到系统聚类树形图,如图 9 所示。

将测点分为 8 类,得到聚类结果为:

v_8 = (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3,8,6,3,8,3,3,5,3,3,4,5,7,3,2,5,7,3,2,5,7,3,2,5,7,8,8,8,8)^T

其中,第 i 个元素表示聚类序号为 i 的测点所属类别,聚类序号按照 1 -1,1 -2,…,11 -4,w1,w2,w3,tem 的顺序排列。各类别包含的测点统计见表 2。

3.3 聚类结果分析

(1)当聚类数为 8 时,类 1 包括主拱的全部 15 个测点;类 2 为主梁三个截面的南侧腹板测点;类 3 包含了斜拱和主梁的所有顶面测点;类 5 主要为主梁底板测点;类 7 为主梁北侧腹板测点;类 8 包括了斜拱底面测点、三个支座位移测点和温度测点。各类中的测点在结构中均属于同构件同类型的测点,受力行为相似,与实际情况相符。

(2)类 4 测点 8 -2 与类 6 测点 5 -3 各自单

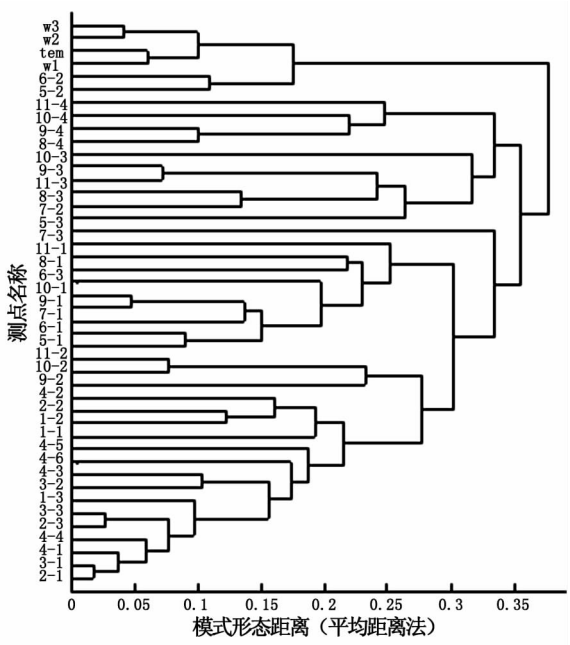

图9 测点相似性聚类树形图

Fig. 9 Clustering tree of similarity of monitoring points

表2 测点类别统计

Tab. 2 Statistics of points' categories

类别	测点编号	类别	测点编号
1	1-1,1-2,1-3,2-1,2-2,2-3,3-1,3-2,3-3,4-1,4-2,4-3,4-4,4-5,4-6	5	7-2,8-3,9-3,10-3,11-3
2	9-2,10-2,11-2	6	5-3
3	5-1,6-1,6-3,7-1,7-3,8-1,9-1,10-1,11-1	7	8-4,9-4,10-4,11-4
4	8-2	8	5-2,6-2,w1,w2,w3,tem

独成一类,其监测数据是相对孤立的。但在正常情况下,测点8-2应归入类2,测点5-3应归入类3。若排除了测点位置结构存在损伤或不同的构造形式,认为这两个测点传感器已损坏,导致监测数据有误。

(3)给出测点聚类数为15时的结果:

v_{15} = (12,12,12,12,12,12,12,12,12,12,12,12,12,12,3,15,14,3,15,3,3,11,10,4,13,8,2,3,5,7,2,3,6,7,1,4,6,8,9,15,15,15,15)^T。在v_{15}中,所有主拱测点仍归在一类(类12),而斜拱与主梁测点则较为分散。说明主拱各测点相似性程度较高。

(4)无论是v_8或v_{15},应变测点5-2、6-2,位移测点w1、w2、w3以及温度测点tem一直在一类,表明温度是影响这几个测点变化的主要因素。

使用层次聚类时,选择不同的特征序列、不同的时间段,均可能得到不一样的结果;即使在同一条件下,也可以根据需要设定不同的分类数量。实际操作时,应试验多种不同的情况,进行综合对比分析,从中筛选提炼出符合真实情况的聚类结果,从而掌握测点相似性的分布情况与特点。

4 结语

基于模式形态距离的时间序列相似性度量方法能够定量地描述时间序列的形态变化趋势差异,有效地度量桥梁监测点间的相似性程度,直观简洁。以此度量为基础,对玉峰大桥监测点进行的相似性聚类分析,结果与桥梁的真实结构状况基本一致,表明该方法具有较高的聚类精度。此外,该方法也适用于其他时间序列的相似性挖掘。

参考文献

[1] KANTARDZIC M. Data mining: concepts, models, methods, and algorithms[M]. John Wiley & Sons, 2011.

[2] ADWAN S, AROF H. On improving dynamic time warping for pattern matching[J]. Measurement, 2012, 45(6): 1609-1620.

[3] WANG X, MUEEN A, DING H, et al. Experimental comparison of representation methods and distance measures for time series data[J]. Data Mining and Knowledge Discovery, 2013, 26(2): 275-309.

[4] 李海林,郭崇慧. 时间序列数据挖掘中特征表示与相似性度量研究综述[J]. 计算机应用研究, 2013, 30(5): 1285-1291.

[5] 王 达,荣 冈. 时间序列的模式距离[J]. 浙江大学学报:工学版, 2004, 38(7): 795-798.

[6] 丁永伟,杨小虎,陈根才,等. 基于弧度距离的时间序列相似度量[J]. 电子与信息学报, 2011, 33(1): 122-128.

[7] 刘世元,江 浩. 关于时间序列相似性概念体系的探讨与研究[J]. 华中科技大学学报:自然科学版, 2004, 32(7): 75-76.

[8] WANG X, SMITH K, HYNDMAN R. Characteristic-based clustering for time series data[J]. Data Mining and Knowledge Discovery, 2006, 13(3): 335-364.

[9] DíAZ S P, VILAR J A. Comparing several parametric and nonparametric approaches to time series clustering: a simulation study[J]. Journal of classification, 2010, 27(3): 333-362.

(责任编辑 王利君)