DOI:10.19651/j.cnki.emt.2107473

基于改进动态时间规整的相似性度量及轨迹聚类

程 前 李建良
(南京理工大学 理学院 南京 210094)

摘 要：针对传统轨迹相似性计算方法度量效果不佳，且当时间序列数据过度扭曲时相似性度量难以取得好的效果。鉴此基于诸多实际应用之精度和实时性需求，基于动态时间规整算法，结合轨迹平移的思路及全局变量约束的思想，通过算法优化和参数分析给出了一种改进动态时间规整算法。数值实验结果表明改进算法在轨迹相似性度量上的识别率为 90%，与经典算法相比提高了 41.25%，度量精度明显提升。进而作为轨迹相似性度量函数结合谱聚类算法应用于轨迹数据聚类分析中，仿真轨迹数据实验结果表明基于改进算法的聚类分析能够清晰区分轨迹簇、聚类效果较为理想。

关键词：动态时间规整；数据挖掘；相似性度量；谱聚类
中图分类号：TP391 **文献标识码**：A **国家标准学科分类代码**：510.4030

Improved dynamic time warping for similar metrics and trajectory clustering

Cheng Qian Li Jianliang
(School of Science, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: For the traditional trajectory similarity calculation method, the measurement effect is not good, and the similarity measurement is difficult to achieve good results when the time series data is excessively distorted. Based on the accuracy and real-time requirements of many practical applications, this article is based on the dynamic time warping, combined with the idea of trajectory translation and the idea of global variable constraints, and gives an improved dynamic time warping algorithm through algorithm optimization and parameter analysis. Numerical experiment results show that the improved algorithm has a recognition rate of 90% in the measurement of trajectory similarity, which is an increase of 41.25% compared with the classic algorithm, and the measurement accuracy is significantly improved. Furthermore, as a trajectory similarity measurement function combined with spectrum clustering algorithm, it is applied to trajectory data clustering analysis. Experimental results of simulated trajectory data shows that clustering analysis based on the improved algorithm can clearly distinguish trajectory clusters and the clustering effect is ideal.

Keywords: dynamic time warping; data mining; similarity measure; spectrum clustering

0 引 言

近年来数据挖掘技术如分类、聚类、模式识别[1]和异常检测[2]等在金融、医疗、气象等领域[3]进行了广泛的尝试与应用。随着各类传感器、移动互联网、卫星定位等技术的高速发展，针对某些特定时域、空域的各类型移动目标跟踪分析[4]、时空轨迹数据序列采集、预测[5]及异常检测[6]。其中轨迹的相似性度量及聚类分析对时间序列数据进行实时性地跟踪、预测、分析和挖掘有着愈发重要的意义和明确的应用前景[7]，进而要求轨迹的相似性度量函数或聚类分析不仅仅确保必要的精度且需要较好的实时性。

最长公共子序列和动态时间规整算法(dynamic time warping，DTW)是相似性度量中使用较为广泛和有效的方法[8]。经典 DTW 算法的时间复杂度为 $O(nm)$，Sakoe-Chiba 约束和 Itakura-Parallelogram 约束通过全局约束方式减少了 DTW 算法复杂度、限制了轨迹的扭曲程度[9]。为提高算法度量的效果，Li 等[10]提出将 DTW 算法改进在粗粒度空间进行规划，将算法的复杂度优化到了 $O(n)$；Wang 等[11]提出将轨迹转化为连续的网格单元对 DTW 算法进行优化，但都存在相似性度量精度上的问题，为此

收稿日期：2021-08-03

文献[12]提出改进的 DTW 算法提升了算法对于异常点和噪声点的鲁棒性。针对聚类分析问题，基于划分方法、层次方法、密度方法[13]和网格方法是常见的轨迹聚类方法[14]，但针对轨迹数据聚类效果不佳。谢英红等[15]提出在流行空间上分析聚类数据点间的相似性问题，王培等[16]基于 Hausdorff 距离和谱聚类对汽车轨迹进行聚类却在精度或实时性两难全。

鉴此，基于诸多精度和实时性需求，本文运用轨迹平移的思想结合全局变量约束的思想，通过优化算法和参数分析对经典 DTW 算法进行改进，使得算法能够较好地对二维轨迹相似度进行识别，并结合谱聚类算法应用于轨迹聚类中，提升了轨迹相似性度量及聚类分析的适用性。

1　经典动态时间规整算法与平移技术

1.1　经典动态时间规整算法

动态时间规整算法被广泛应用于模式识别和时间序列的数据挖掘中，基本思想是通过扭曲时间序列来进行重复采样，对轨迹进行拓展或压缩，然后通过迭代的方式从所有可能的变换路径中找到使得总距离最小的匹配路径。

设 A,B 为维数为 n,m 的特定移动目标的轨迹数据，分别表示为 $A=(a_1,\cdots,a_n)$、$B=(b_1,\cdots,b_m)$，设 $Head(A)=(a_1,\cdots,a_{n-1})$，$Head(B)=(b_1,\cdots,b_{m-1})$，则轨迹 A,B 的 DTW 算法距离定义如式(1)所示。

$$DTW(A,B)=\begin{cases}\sum_{k=1}^{m}d(a_1,b_k), & n=1\\ \sum_{k=1}^{n}d(a_k,b_1), & m=1\\ d(a_n,b_m)+\min\begin{Bmatrix}DTW(Head(A),B)\\ DTW(A,Head(B))\\ DTW(Head(A),Head(B))\end{Bmatrix}, & \text{其他}\end{cases} \tag{1}$$

其中，$d(a_i,b_j)$ 是轨迹数据 A,B 中，点 a_i,b_j 之间距离的度量函数。

基于式(1)，可以得到 DTW 算法定义的两个轨迹之间的距离，由于 DTW 距离会受不同轨迹长度影响，故定义轨迹 A,B 之间的相似度距离 D_1 如式(2)所示。

$$D_1(A,B)=\frac{DTW(A,B)}{\max(n,m)} \tag{2}$$

1.2　基于平移的算法改进思路

为了克服 DTW 算法对一定距离范围内轨迹间的平行相似性度量效果不佳的问题，从平移轨迹中得到启发，在 DTW 算法中引入轨迹平移的思路定义一个更适合的轨迹相似度度量函数。首先考虑二维轨迹间的平移，设 F 为一族位移函数族，$c_1<c_2,d_1<d_2$ 且均为实数，$A=((x_1,y_1),\cdots,(x_n,y_n))$ 一条二维轨迹数据，对 $\forall c,d\in[c_1,c_2]\times[d_1,d_2]$，则 $\exists f_{c,d}\in F$ 使得式(3)成立。

$$f_{c,d}(A)=((x_1+c,y_1+d),\cdots,(x_n+c,y_n+d)) \tag{3}$$

则轨迹 A,B 之间的相似度距离 D_2 如式(4)所示。

$$D_2(A,B)=\min_{c,d\in[c_1,c_2]\times[d_1,d_2]}D_1(A,f_{c,d}(B))=\min_{c,d\in[c_1,c_2]\times[d_1,d_2]}\frac{DTW(A,f_{c,d}(B))}{\max(n,m)} \tag{4}$$

进一步轨迹 A,B 之间的距离 D_2 可以推广到 n 维空间的轨迹间相关性度量，如图 1 所示，通过对轨迹的平移来改进 DTW 算法，我们可以更好地识别在多维空间种有平行关系的运动轨迹之间的相似性。

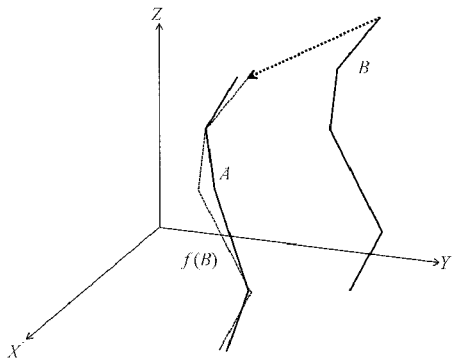

图 1　轨迹的平移

由于 DTW 算法时间复杂度较高，故中想要得到准确解 $c,d\in[c_1,c_2]\times[d_1,d_2]$，使得式(4)成立是困难的。故首先考虑在 $[c_1,c_2]\times[d_1,d_2]$ 上进行网格搜索，寻找轨迹 A,B 之间的距离 D_2 的近似解；若取搜索步长为 δ，则在区域 $[c_1,c_2]\times[d_1,d_2]$ 上进行网格寻找的解 D_3 如式(5)所示。

$$D_3=\min_{k_i\in\left[0,\cdots,\left\lceil\frac{c_2-c_1}{\delta}\right\rceil\right]}\ \min_{k_j\in\left[0,\cdots,\left\lceil\frac{d_2-d_1}{\delta}\right\rceil\right]}D_1(A,f_{c_1+k_i\delta,d_1+k_j\delta}(B)) \tag{5}$$

根据数值实验，基于平移的 DTW 算法相似度度量效果有较为明显的提升；但由式(5)知求解相似度距离 D_3 需要计算 $\left\lceil\frac{(c_2-c_1)}{\delta}\right\rceil\times\left\lceil\frac{(d_2-d_1)}{\delta}\right\rceil$ 次经典 DTW 距离，计算复杂度较高，仍需对其进行优化。

2　改进的动态时间规整算法

2.1　算法优化

设 $G=g_1,g_2,\cdots,g_k,\cdots,g_K$ 为轨迹 $A=(a_1,\cdots,a_n)$，和轨迹 $B=(b_1,\cdots,b_m)$ 通过局部缩放获得的一组对应点[17]。其中 $g_1=(1,1),g_k=(i(k),j(k)),g_K=(n,m)$。则由 DTW 算法定义的两个轨迹 A,B 之间的 DTW 距离公式可以改写为式(6)：

$$DTW(A,B)=\min_{G}\left[\sum_{k=1}^{K}d(a_{g_k(1)},b_{g_k(2)})\right] \tag{6}$$

通过动态规划算法方式求得相似度距离 D_1 后，可以得到最佳路径 G^* 如式(7)所示。

$$G^*=g_1^*,g_2^*,\cdots,g_K^* \tag{7}$$

设 $G^* = g_1^*, g_2^*, \cdots, g_K^*$，则轨迹 A, B 之间的相似性度量距离 D_2 的改进算法的度量距离 D_4 如式(8)所示。

$$D_4(A,B) = \begin{cases} \dfrac{\sum_{k=1}^{K} \left| d(a_{g_K^*(1)}, b_{g_K^*(2)}) - M \right|}{\max(n,m)}, & M < \sigma \\ D_1(A,B), & \text{其他} \end{cases} \tag{8}$$

其中，$M = Mean(d(a_{g_1^*(1)}, b_{g_1^*(2)}), \cdots, d(a_{g_K^*(1)}, b_{g_K^*(2)}))$ 是最佳路径中的所有对应点之间距离的平均值；σ 是与平移范围区域 $[c_1, c_2] \times [d_1, d_2]$ 相关的距离阈值参数。

2.2　改进算法的参数选取

下面针对改进算法的距离范围阈值参数 σ，通过数值实验选取较为合理的参数取值。实验环境为 Python 3.7 软件，Windows 10 操作系统，ThinkPad E480(处理器 Inter(R)Core(TM)i7-8550UCPU@1.80 GHz 2.00 GHz)。数据集选取欧洲 Dublin 地区一组火车轨迹数据集共 1 000 条真实轨迹；σ 分别取值为 1、2、3、4、5，Sakoe-Chiba 约束的弯曲限制 $r = \max(\omega \times \max(n,m), \| n-m \|)$ 中 ω 分别取值为 1、3/4、1/2、1/4，使用改进算法进行计算得到的算法的识别率如表 1 所示，其中识别率如式(9)所示。

$$R = \frac{\sum_{k=1}^{a} Num(N_k \cap M_k)}{\sum_{k=1}^{a} Num(M_k)} \tag{9}$$

其中，a 为测试集元素个数取为 5，M 为算法得到的相似轨迹集合元素个数取为 20，N 为仿真得到的相似轨迹集合元素个数取为 20，Num 为集合中轨迹元素的个数。

如表 1 数据所示，在改进算法中，距离范围阈值参数 σ 取值和 Sakoe-Chiba 约束的弯曲限制取值对算法的准确率有较为明显的影响，当 σ 取值为 4 时，改进算法的识别率 90%高于其他取值情况。故在后续实验中，针对相同数据集，改进算法的距离阈值参数 σ 取值为 4。

表 1　改进算法在不同参数取值下的识别率　%

参数取值	改进算法($\sigma=1$)	改进算法($\sigma=2$)	改进算法($\sigma=3$)	改进算法($\sigma=4$)	改进算法($\sigma=5$)
$\omega=1$	36.25	43.75	68.75	78.75	77.50
$\omega=3/4$	36.25	43.75	70.00	78.75	77.50
$\omega=1/2$	40.00	45.00	76.25	80.00	78.75
$\omega=1/4$	48.75	52.50	77.50	90.00	87.50

3　数值实验与分析

3.1　相似性度量的有效性和时效性

首先，需要设计实验以验证改进算法的有效性。数据集选取二维数据集基于欧洲 Dublin 地区火车轨迹仿真生成的 100 组相似轨迹对，分别使用 DTW 算法，基于网格寻找的平移算法和本文给出的改进 DTW 算法分别计算轨迹数据的相似性度量。通过多次实验调试，选择的较好的实验参数如下，通过基于网格平移 DTW 算法中 $\delta = 0.02$，$c_1 = -0.1, c_2 = 0.1, d_1 = -0.1, d_2 = 0.1$，得到的 3 种算法的度量距离如图 2 所示、计算时间如图 3 所示。

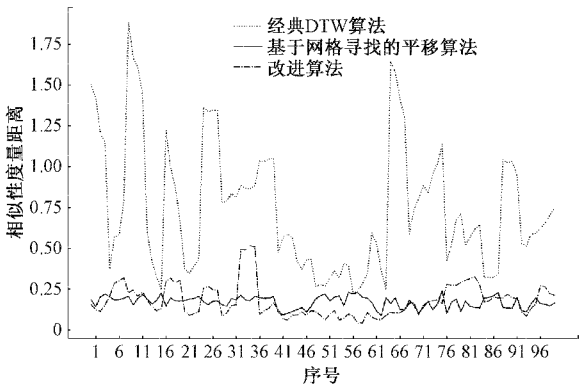

图 2　3 种算法的相似性度量

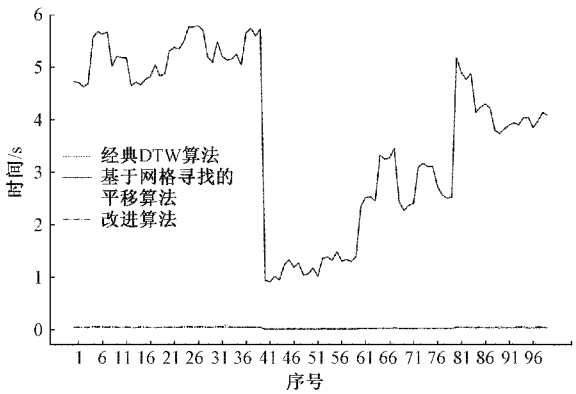

图 3　3 种算法的计算时间

实验中选取点 $a = (Lat1, Lon1), b = (Lat2, Lon2)$ 之间的距离度量函数 $d(a,b)$ 如式(10)所示。

$$d(a,b) = 2M\arcsin\sqrt{N} \tag{10}$$

其中，N 如式(11)所示。

$$N = \sin^2\left(\frac{Lat1 - Lat2}{2}\right) + \cos(Lat1) \cdot \cos(Lat2) \cdot \sin^2\left(\frac{Lon1 - Lon2}{2}\right) \tag{11}$$

其中，M=6 378.137 km 为地球半径。

图 2 数据指出，在 100 组相似轨迹对的相似性度量中，基于网格平移的 DTW 算法的平均度量为 0.167 8，结果显示轨迹间相似性较强，即该算法能够较好地识别轨迹间的相似性关系。同时改进算法的平均度量为 0.176 4，针对相似轨迹的度量效果优于经典 DTW 算法。图 3 中数据指出，经典 DTW 算法平均耗时为 0.027 5 s；基于网格寻找的平移算法通过重复进行多次经典 DTW 距离计算，平均耗时为 3.560 4 s，是经典 DTW 算法平均耗时的上百倍，而本文提出的改进算法平均耗时为 0.028 8 s 与经典 DTW 算法

运算效率相差较小，比基于网格寻找的平移 DTW 算法平均耗时低得多。通过上述实验数据说明，本文提出的改进 DTW 算法，对轨迹间一定范围内的类平移关系有更好的识别能力；同时计算效率较好，适用于实际的问题计算中。

3.2　改进算法在轨迹识别中的应用

为验证改进算法在轨迹识别中的有效性，在与 4.1 实验同样的实验环境下，数据集选取欧洲 Dublin 地区一组火车轨迹数据集共 1 000 条真实轨迹；通过选取不同的 Sakoe-Chiba 约束的弯曲限制 $r=\max(\omega\times\max(n,m),\|n-m\|)$。基于经典 DTW 算法和改进算法对测试集与实验数据集进行计算；其中 ω 分别取值为 1、3/4、1/2、1/4，取值渐进达到了弯曲限制的最小值，根据 2.2 节参数选取 σ 取值为 4。DTW 算法和本文给出的改进算法的算法得到的识别率和平均用时如表 2 所示，识别率的计算公式与 2.2 节一致。

表 2　两种算法识别率和平均用时

算法类型	DTW 算法				改进算法			
	$\omega=1$	$\omega=3/4$	$\omega=1/2$	$\omega=1/4$	$\omega=1$	$\omega=3/4$	$\omega=1/2$	$\omega=1/4$
识别率/%	36.25	36.25	40	50	78.75	78.75	80	90
平均用时/s	0.110 2	0.113 3	0.101 1	0.091 1	0.113 1	0.115 9	0.103 8	0.093 0

表 2 数据指出，在轨迹弯曲限制 ω 的取值上，取值为 1/4 时渐进达到了 Sakoe-Chiba 约束弯曲限制的最小值，度量效果优于其他情况，且算法平均耗时低于其他情况。上述数据反映了在仿真轨迹数据集上，本文给出的改进算法在对有类平移关系的轨迹识别中比经典 DTW 算法的识别率高，同时改进算法与经典 DTW 算法在相同数据集上取相同弯曲限制时平均用时相差不大。综合实验考虑，本文给出的改进算法更适合轨迹间相似性的度量，同时时间复杂度可以满足实际问题的要求。

3.3　基于谱聚类的聚类分析

聚类是应用于高维数据降维和聚类的经典方法之一，谱聚类来源于谱图划分理论的思想，将轨迹数据集中的每条轨迹数据视为图的顶点 V，根据样本之间的相似性度量为顶点间的边 E 赋值权重 W，此时得到了基于轨迹数据相似性度量的有向无环图 $G=(V,E)$。基于图论的最优划分准则使得两个子图内部相似性最高，子图间的相似性最低。

首先计算基于改进 DTW 算法计算轨迹间的相似性矩阵 $\boldsymbol{A}$ 如式(12)所示。

$$\boldsymbol{A}_{ij}=\exp\left(-\frac{D_4(s_i,s_j)}{2\Delta^2}\right) \tag{12}$$

其中，s_i 表示第 i 条轨迹数据，Δ 为尺度参数。

其次通过相似性矩阵 $\boldsymbol{A}$，构建由全部度值为对角元素的度矩阵 $\boldsymbol{D}$ 和正规拉普拉斯矩阵 $\boldsymbol{L}$ 如式(13)、(14)所示。

$$\boldsymbol{D}_{ii}=\sum_j \boldsymbol{A}_{ij} \tag{13}$$

$$\boldsymbol{L}=\boldsymbol{D}^{1/2}\boldsymbol{A}\boldsymbol{D}^{1/2} \tag{14}$$

最后通过计算正规拉普拉斯矩阵 $\boldsymbol{L}$ 的特征值，取前 k 个特征值，并构建与之相对应的特征向量矩阵 $\boldsymbol{U}=(u_1,u_2,\cdots u_k)$，使用 K-means 算法对矩阵 $\boldsymbol{U}$ 的列向量进行聚类。

为验证改进 DTW 算法在谱聚类上的应用效果，数据集选取欧洲 Dublin 地区火车轨迹数据集仿真得到的 5 组 100 条轨迹数据，使用经典 DTW 算法和改进 DTW 算法作为相似性度量函数进行谱聚类，根据 3.2 节结果度量算法参数选取 $\omega=1/4$、$\sigma=4$，谱聚类参数选取 $\Delta=5$。基于经典 DTW 算法的谱聚类可视化效果如图 4 所示，基于改进 DTW 算法的谱聚类可视化效果如图 5 所示。

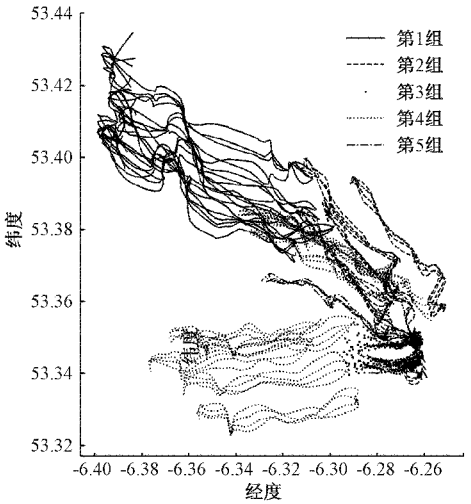

图 4　基于经典 DTW 算法的谱聚类

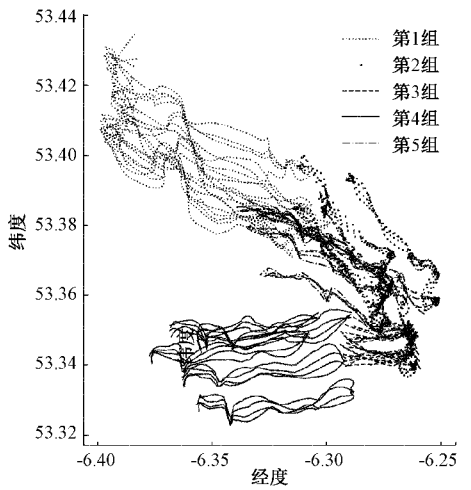

图 5　基于改进 DTW 算法的谱聚类

根据图 4、5 数据所示,在轨迹仿真数据集上基于经典 DTW 度量算法的谱聚类效果较差;而基于改进 DTW 相似性度量算法的谱聚类能够较好的反应该地区轨迹数据空间分布,通过更好地度量轨迹间的相似性关系,并对轨迹数据集进行谱聚类算法,聚类结果能够清晰区分轨迹簇、聚类效果较为理想。

4　结　　论

本文中提出了一种基于 DTW 算法和轨迹平移思想的改进算法来度量时间序列数据间的相似性,提高了对一定范围内轨迹间相关性的识别效果;同时利用谱聚类算法对仿真轨迹数据集进行聚类分析。通过多次实验结果表明,本文提出的改进算法能够很好地对具有平移相似性轨迹进行识别、结合谱聚类算法能够清晰区分轨迹簇,且改进算法在时间复杂度上与经典 DTW 算法相差不大,时效性较好。虽然经改进后的算法在有平行关系的轨迹相似度度量中有较好的表现,但对其他相似性关系能不能保证识别的准确性,还需要进一步优化与研究。本文的下一步工作是对改进算法及聚类分析进行计算效率优化[18]、轨迹预测[19]以及结合聚类分析等操作实现对热点轨迹的寻找等。

参考文献

[1] 赵驰盟,兀伟,杨银芳. 磁探测引信日标识别算法研究[J]. 国外电子测量技术,2020,39(9):47-52.

[2] JEONG H, YOO Y, YI K M, et al. Two-stage online inference model for traffic pattern analysis and anomaly detection[J]. Machine Vision & Applications, 2014, 25(6):1501-1517.

[3] CHEN X Y, BEN-CHANG W U, HAN H T. Research of rainfall weather model based on multi-dimensional time series data mining[J]. Computer Engineering and Design, 2010, 31(4):898-902.

[4] 包本刚. 融合多特征的目标检测与跟踪方法[J]. 电子测量与仪器学报,2019,33(9):93-99.

[5] LE Q I, ZHENG Z. Trajectory prediction of vessels based on data mining and machine learning[J]. Journal of Digital Information Management, 2016, 14(1):33-40.

[6] 董静怡,庞景月,彭宁,等. 集成 LSTM 的航天器遥测数据异常检测方法[J]. 仪器仪表学报,2019,40(7):22-29.

[7] AGHABOZORGI S, SHIRKHORSHIDI A S, WAH T Y. Time-series clustering-A decade review[J]. Information Systems, 2015, 53(C):16-38.

[8] NEHAL M, TAMER A, KHALED E L B. A comparative study of similarity evaluation methods among trajectories of moving objects[J]. Egyptian Informatics Journal,2018,19(3):165-177.

[9] 李正欣,张凤鸣,李克武,等. 一种支持 DTW 距离的多元时间序列索引结构[J]. 软件学报,2014,25(3):560-575.

[10] LI H, GUO C, QIU W. Similarity measure based on piecewise linear approximation and derivative dynamic time warping for time series mining[J]. Expert Systems with Applications, 2011, 38(12):14732-14743.

[11] WANG Y, LEI P, ZHOU H, et al. Using DTW to measure trajectory distance in grid space[C]. 2014 4th IEEE International Conference on Information Science and Technology, IEEE, 2014: 152-155.

[12] 郭岩,罗珞珈,汪洋,等. 一种基于 DTW 改进的轨迹相似度算法[J]. 国外电子测量技术,2016,35(9):66-71.

[13] LI M, BI X, WANG L, et al. A method of two-stage clustering learning based on improved DBSCAN and density peak algorithm[J]. Computer Communications, 2021, 167:75-84.

[14] 高强,张凤荔,王瑞锦,等. 大数据:数据处理关键技术研究综述[J]. 软件学报,2017,28(4):959-992.

[15] 谢英红,何宇清,王楠. 基于 Grassmann 流形的谱聚类分析算法[J]. 电子测量与仪器学报,2017,31(3):338-342.

[16] 王培,江南,万幼,等. 应用 Hausdorff 距离的时空轨迹相似性度量方法[J]. 计算机辅助设计与图形学学报,2019,31(4):647-658.

[17] 叶科淮,陈志,王仁杰,等. 动态时间规整算法优化[J]. 软件导刊, 2021,20(1):132-135.

[18] RAKTHANMANON T, CAMPANA B, MUEEN A, et al. Searching and mining trillions of time series subsequences under dynamic time warping[C]. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012: 262-270.

[19] 张志远,倪国新,徐艳国. 轨迹预测技术的现状及发展综述[J]. 电子测量技术,2020,43(13):111-116.

作者简介

程前,硕士研究生,主要研究方向为信息处理与数据挖掘。

E-mail:812967569@qq.com

李建良,教授,主要研究方向为工程数值计算与优化、数值模拟和仿真。

E-mail:ljl6006@njust.edu.cn