

◎工程与应用◎

改进DTW算法的心电信号相似性度量

涂 辉¹, 刘 丽^{1,2}, 张正金¹TU Hui¹, LIU Li^{1,2}, ZHANG Zhengjin¹

1.江南大学 物联网工程学院, 江苏 无锡 214122

2.无锡市第四人民医院 信息科, 江苏 无锡 214122

1.School of IOT Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China

2.Department of Information, Wuxi Fourth People's Hospital, Wuxi, Jiangsu 214122, China

TU Hui, LIU Li, ZHANG Zhengjin. Improved DTW algorithm to measure similarity of ECG signal. Computer Engineering and Applications, 2015, 51(16): 215-218.

Abstract: In view of when use the Dynamic Time Warping (DTW) algorithm to measure the time series similarity, but the sequence has a larger excursion it will cause abnormal path and inaccurate matching. According to the nature of the electrocardiogram (ECG) signal, the most significant feature based on ECG priority matching method is proposed. In order to reduce the time consumption, according to the location of R waves, restrictive windows for curved path adaptively divided, compared with the traditional method, the proposed algorithm increases the accuracy, and meanwhile the running time is significantly reduced.

Key words: time series; dynamic time warping; ElectroCardioGram (ECG); priority matching; adaptive

摘 要: 针对时间序列相似性度量中动态时间规整 (DTW) 算法在序列时间轴偏移较大时易产生病态路径及匹配不准确的问题, 根据心电信号自身的特性, 提出一种基于心电信号最显著特征的优先匹配法。为减小算法时间复杂度, 根据心电信号 R 波特征点位置自适应地划分弯曲路径的约束窗口。与传统算法比较, 在提高准确率的同时, 运行时间大幅降低。

关键词: 时间序列; 动态时间规整; 心电信号; 优先匹配; 自适应

文献标志码: A **中图分类号:** TP391.4 **doi:** 10.3778/j.issn.1002-8331.1308-0239

1 引言

时间序列是一类常见且与时间相关的高维数据, 也是数据挖掘领域中主要的研究对象, 广泛存在于金融、医学、气象、网络安全等领域中^[1]。时间序列相似性度量是时间序列分析中的重要任务之一, 对时间序列的分类、聚类、检索具有重要意义。

目前时间序列相似性的度量方法主要有: 欧氏距离法^[2] (Euclidean Distance), 直接计算时间轴上对应点间的距离, 算法简单、快速, 但它只能应用于等长的时间序列, 且对序列在时间轴上的偏移及序列突变敏感。动态

时间规整 (Dynamic Time Warping, DTW) 算法根据最小代价的时间弯曲路径进行匹配, 对时间序列发生弯曲后的相似性度量具有很好的鲁棒性。算法的时间复杂度高且序列在时间轴偏上偏移较大时容易产生病态路径及匹配不准确的问题。如图 1 欧几里德距离与 DTW 距离对比图。

1994 年, Berndt 和 Clifford 首先把动态时间弯曲距离引入到时间序列分类的问题中^[3]。针对 DTW 计算量大的问题, Huang 和 Kinsner 提出一种调节窗口法^[4]。Thanawin 等人提出一种在大数据量下的时间序列挖掘

基金项目: 国家自然科学基金 (No.61300150)。

作者简介: 涂辉 (1987—), 男, 硕士研究生, 研究领域为临床医学的非结构化数据挖掘; 刘丽 (1975—), 女, 博士, 副教授, 研究领域为医学图像处理; 张正金 (1986—), 男, 硕士研究生, 研究领域为模糊系统。

收稿日期: 2013-08-16 **修回日期:** 2014-01-16 **文章编号:** 1002-8331(2015)16-0215-04

CNKI 网络优先出版: 2014-01-20, <http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1308-0239.html>

算法^[5]。这些工作的前提都是比较时间序列的相似性。

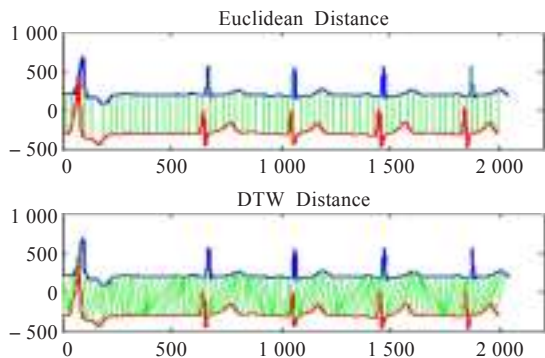


图1 欧式距离与DTW距离对比图

2 相关知识

2.1 动态时间弯曲距离

动态时间弯曲距离的主要思想为通过调整时间点之间的对应关系,找出两个任意长时间序列中数据之间的最佳匹配路径,从而度量时间序列的相似性^[6]。时间序列 $X=(x_1, x_2, \dots, x_m)$, $Y=(y_1, y_2, \dots, y_n)$ 长度分别为 m, n 。距离矩阵 D 由 x_i 与 y_j 的欧式距离的平方 $d_{i,j}=d(x_i, y_j)=(x_i - y_j)^2$ 构成。在 D 中寻找一条弯曲路径 $W=(w_1, w_2, \dots, w_k)$ 使得 X 与 Y 的匹配度最大。其中 $\max(m, n) \leq k \leq m + n - 1$, $w_l=(i, j)$ 表示 x_i 与 y_j 匹配。 w_i 的约束为:

(1)边界约束

$$w_1=(1, 1); w_k=(m, n)$$

(2)单调性、连续性约束

$$w_k=(a_k, b_k); w_{k+1}=(a_{k+1}, b_{k+1})$$

其中 $0 \leq a_{k+1} - a_k \leq 1, 0 \leq b_{k+1} - b_k \leq 1$ 。

DTW 路径是 D 中满足以上条件的最短路径,如图2:

$$DTW(X, Y) = \min_k \left(\frac{1}{k} \sum_{p=1}^k w_p \right)$$

最优路径的查找是通过动态规划来实现的^[7]。定义累积矩阵 $R=\{r(i, j)\}_{m,n}$ 来记录最短路径,即

$$r(i, j) = d(x_i, y_j) + \min \begin{cases} r(i, j-1) \\ r(i-1, j) \\ r(i-1, j-1) \end{cases}$$

实现算法的时间复杂度为 $O(mn)$ 。

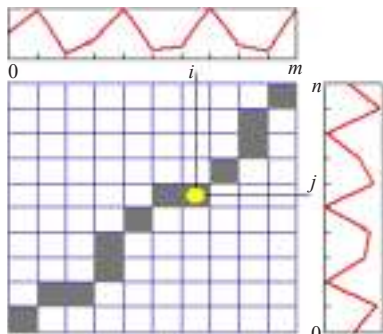


图2 DTW弯曲路径示意图

2.2 心电信号

心电信号是心肌的电活动在体表的表現,包含了許多关于心脏疾病的信息。心电图的一个周期的波形由P-QRS-T波构成^[8],其中QRS波群是单个心拍中最显著的特征。最典型心电图如图3所示。

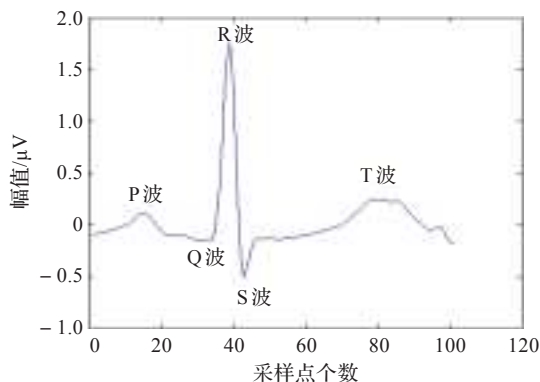


图3 典型心电信号

ECG 信号已被证实是一种非平稳信号,其可视化图形是一种非参数曲线。由于信号微弱易受噪声影响,病变种类多,个体差异大,分布不均衡,非稳定并且以某种不规则的复杂方式波动^[9]等原因使得心电信号的相似性度量更加困难。目前对心电信号的研究主要有3大类^[10]: (1)基于机器学习的方法,如文献[11]提出基于高阶累计量和希尔伯特基函数获取同一信号的两组特征集,分别训练得到两个SVM分类器,用加权投票法对分类器结果集成。(2)数据挖掘方法,如文献[12]提出一种Swift-Rule规则挖掘方法,针对不同种类的时间序列信号挖掘,自动得出复杂的分类规则,通过对ECG信号的挖掘,实现了正异常分类。(3)基于数学模型的方法,如隐马尔科夫模型。

形态学的方法更符合人直观的理解,由于起步晚,准确率目前的方法不尽如人意而使用较少^[13]。

3 改进DTW算法

本文在研究总结已有算法不足基础上提出改进型DTW算法,该算法首先遵循最大特征点优先匹配原则,然后根据信号特征自适应确定弯曲窗口的范围。实验表明,算法在提高匹配精度的同时减少了运算时间。

3.1 最大特征点优先匹配原则

传统的时间序列相似性度量算法将序列看作是静态的,严格按照顺序进行匹配。ECG信号可看作是动态的时间序列,信号中一旦出现病变的波形则诊断为某种疾病。大量实验发现对曲线相似性度量精度影响最大的是曲线中特异性较大的点。心电信号中QRS波群幅值高,斜率大是信号中最明显的特征。病变的心电信号如室性心律失常患者的QRS波群要比正常QRS波群宽大,但这种病变波并不在每个周期稳定存在^[14],如图4序列1中A波和序列2中B波为病变信号,也是特征最明

显的波群(本文序列横轴使用采样点个数表示,因为在一定采样频率下,采样点个数与时间存在对应关系)。如直接运用DTW路径计算将出现误匹配,如图5。

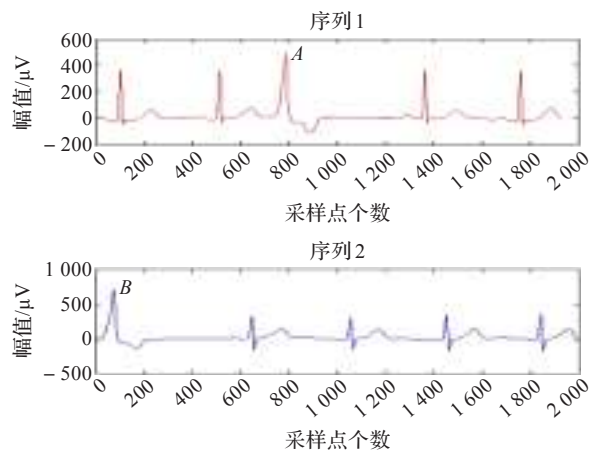


图4 两例室性心律失常心电图

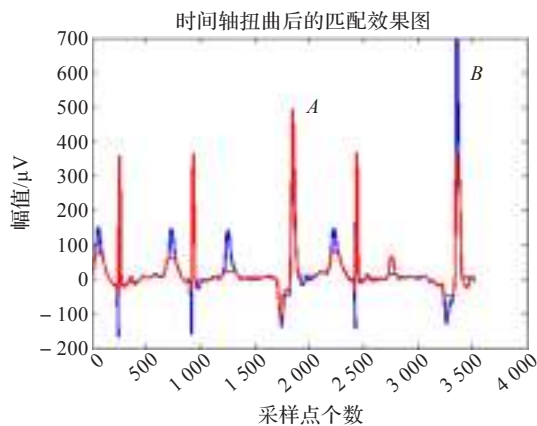


图5 错误匹配结果

最大特征点优先匹配原则是将序列1中A点平移至与B点横轴相当的位置,再计算DTW距离。

3.2 自适应弯曲窗口约束

临床上广泛使用的12导联心电图机可记录十几秒的心电数据。若按2.1小节中传统算法计算,当弯曲路径偏离对角线较大时会产生病态匹配,针对该问题的改进算法中加入弯曲窗口(Warping Window),可改善病态弯曲并轻微减少算法的运行时间^[15],但对匹配的精度会产生影响,不同情况下弯曲窗口的范围是难以确定的。最常见的弯曲窗口约束如图6。

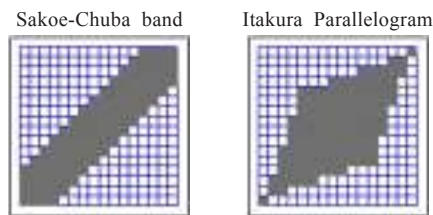


图6 两种常见的弯曲窗口约束

本文提出一种自适应弯曲窗口约束,该方法根据不同ECG信号QRS波群位置可自动调整约束窗口范围,

不仅具有普遍适应性,还能够降低计算时间。

研究发现,两条序列的距离矩阵D中存在一些间距不等、颜色深浅不同的横竖条带状平行线(三维图中是平面),这些平行线构成了D中大小不一的矩形,而弯曲路径W必定经过D的对角线上的各小矩形对角点,由此用这些小矩形将W分成若干段,除起止段外每段的W完全位于所在的矩形中,如图7。

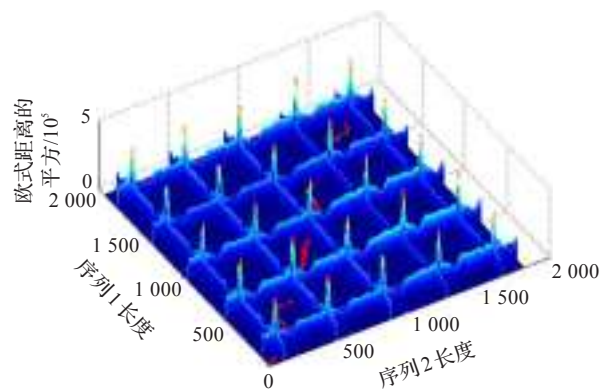


图7 仰角为172°方位角为32°时的D与W

这些平行线正是序列中信号特异性较大的位置也即各QRS波群位置,W经过对角线上矩阵的对角点从事实上证明了最大特征匹配的正确性。由此只需计算D中位于对角线上的矩阵内的值,而忽略其他部分,从而减少运算时间。对于起止段的约束如下:

W起点 $w_1=(1,1)$,起点段的右上角点为两序列的第一个R波所在位置构成的横纵坐标。

W终点 $w_k=(m,n)$,终点段的左下角点为两序列最后一个R波所在位置构成的横纵坐标。如图8自适应窗口为图中白色部分,红色曲线为动态弯曲路径。匹配效果如图9。

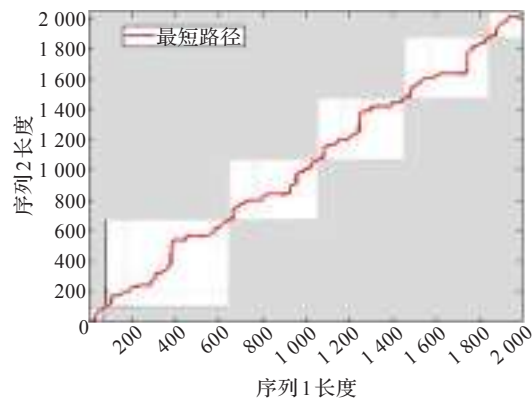


图8 自适应约束窗口

4 实验结果

为验证算法有效性,本文选取中国心血管数据库(CCDD)中室性心律失常部分ECG信号实验,从最短路径的计算精度及算法消耗时间方面对传统算法与本文提出算法进行比较。如表1所示。 $Len(t)$ 代表t长度, Dis 代表传统算法的最短DTW距离, Th_Dis 代表本文

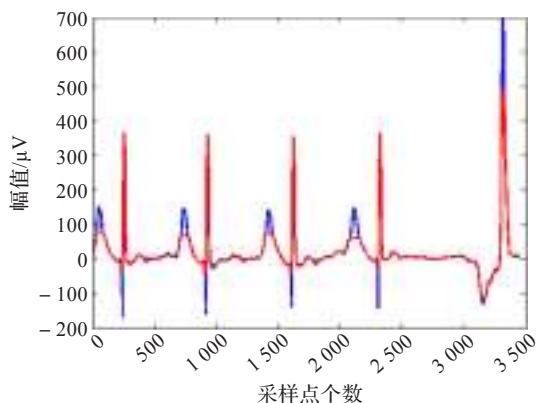


图9 匹配效果图

表1 实验性能比较

Len(t1)	Len(t2)	Dis	T/s	Th_Dis	Th_T/s
499	542	1.267 7	2.503	0.888 00	0.263
817	844	0.872 3	4.194	0.548 90	0.801
1 089	1 126	0.747 1	11.337	0.496 10	0.950
1 485	1 516	0.594 7	25.826	0.398 10	1.272
1 994	2 042	0.487 5	56.746	0.321 90	1.859
2 266	2 324	0.430 7	93.402	0.312 70	1.995
2 831	2 850	0.372 3	136.099	0.263 20	2.732
3 103	3 132	0.339 1	198.387	0.258 90	2.956
3 525	3 548	0.304 5	304.011	0.240 31	3.186

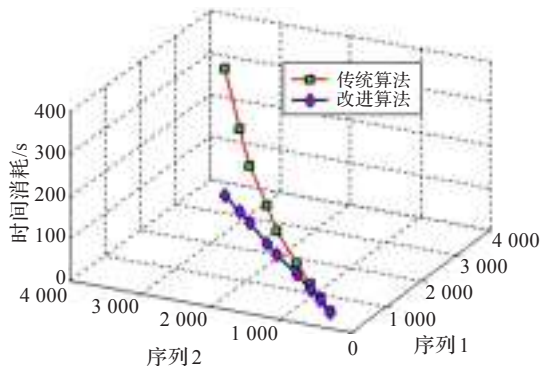


图10 时间消耗对比图

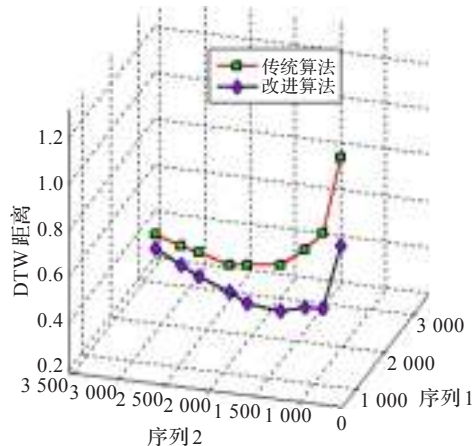


图11 DTW距离对比图

算法的最短DTW距离, T 与 Th_T 分别代表传统算法与本文算法的时间消耗。

实验表明改进算法的准确度更高,随着序列长度增加,算法的时间消耗优势更加明显。

5 结束语

ECG信号由于维数高、数据量大、变化多,直接进行相似性度量不能满足实际应用需求。本文对时间序列相似性度量的传统算法进行了深入研究,在此基础上提出最大特征优先匹配原则和自适应约束窗口法,进一步提高了运算精度的同时减少了算法时间复杂度,为时间序列的分类、检索提供了新的思路。

参考文献:

- [1] 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.
- [2] Danielsson P.Euclidean distance mapping[J].Computer Graphics and Image Processing,1980,14:227-248.
- [3] Berndt D,Clifford J.Using dynamic time warping to find patterns in time series[C]//AAAI-94 Workshop on Knowledge Discovery in Databases,Seattle,WA,USA,1994:229-248.
- [4] Huang B,Kinsner W.ECG frame classification using dynamic time warping[C]//Electrical and Computer Engineering,IEEE CCECE 2002,Canadian Conference,2002,2:1105-1110.
- [5] Rakthanmanon T,Campana B,Mueen A.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.New York,NY,USA:ACM,2012:262-270.
- [6] 李海林,郭崇慧,杨丽彬.基于分段聚合时间弯曲距离的时间序列挖掘[J].山东大学学报:工学版,2011,41(5):57-62.
- [7] Eamonn J,Michael J.Derivative dynamic time warping[C]//The First SIAM International Conference on Data Mining.Washington:IEEE,2001:1-11.
- [8] Yun-Chi Yeha C,Wang Wen-June.QRS complexes detection for ECG signal:The difference operation method[J].Computer Methods and Programs in Biomedicine,2008,91:245-254.
- [9] 杨小冬,宁新宝,何爱军,等.基于多尺度的人体ECG信号质量指数谱分析[J].物理学报,2008,57(3):1514-1521.
- [10] 张嘉伟.心电图形态特征的识别及其在分类中的作用研究[D].上海:华东师范大学,2012.
- [11] Osowski S,Hoai L T,Markiewicz T.SVM-based expert system for reliable heartbeat recognition[J].IEEE Transactions on Biomedical Engineering,2004,51(4):582-589.
- [12] Fisch D,Gruber T,Sick B.SwiftRule:mining comprehensible classification rules for time series analysis[J].IEEE Transactions on Knowledge and Data Engineering,2011,23(5):774-787.
- [13] 董军,徐森,詹聪明,等.心电图识别与分类:方法、问题和新途径[J].生物医学工程学杂志,2007,24(6):1224-1229.
- [14] Davis D.Quick and accurate 12-lead ECG interpretation[M].李立志,史大卓,译.北京:科学技术文献出版社,2004:421-424.
- [15] 陈乾,胡谷雨.一种新的DTW最佳弯曲窗口学习方法[J].计算机科学,2012,39(8):191-195.