基于滑动平均与分段线性回归的时间序列相似性

冯玉伯[1,2] 丁承君[1] 高 雪[1] 朱雪宏[1] 刘 强[1]
(河北工业大学机械工程学院 天津 300130)[1]
(泰华宏业(天津)机器人技术研究院有限责任公司 天津 300130)[2]

摘 要 针对时间序列相似性度量中欧氏距离对异常数据敏感以及 DTW 距离算法效率低的问题,提出基于滑动平均与分段线性回归的时间序列相似性方法。首先,使用初始可变滑动平均算法以及分段线性回归对原始时间序列进行数据变换,并将分段线性回归的参数(截距与距离)集作为时间序列的特征,以实现时间序列的特征提取和数据降维;然后,利用动态时间弯曲距离进行距离计算。该方法在时间序列相似性上与 DTW 算法的性能相近,但是在算法效率上几乎提高了 96%。实验结果验证了该方法的有效性与准确性。

关键词 时间序列,滑动平均,线性回归,动态时间弯曲距离
中图法分类号 TP311.13 **文献标识码** A

Time Series Similarity Based on Moving Average and Piecewise Linear Regression

FENG Yu-bo[1,2] DING Cheng-jun[1] GAO Xue[1] ZHU Xue-hong[1] LIU Qiang[1]
(School of Mechanical Engineering, Hebei University of Technology, Tianjin 300130, China)[1]
(Taihua Hongye (Tianjin) Robot Technology Research Institute Co. Ltd., Tianjin 300130, China)[2]

Abstract Aiming at the problems that the Euclidean distance is sensitive to the anomaly data and the efficiency of the DTW distance algorithm is low, a time series similarity method based on the moving average and the piecewise linear regression was proposed. Firstly, the original variable-averaging algorithm and the piecewise linear regression are used to transform the original time series. The parameters of the piecewise linear regression (intercept and distance) are taken as the characteristics of the time series so that the feature extraction of the time series is realized, and the data is dimensioned. Then it calculated distance using the dynamic time bending distance. The performance of the method is similar to that of DTW algorithm, but the proposed method is almost 96% higher in algorithm efficiency. The experimental results verify the effectiveness and accuracy of the method.

Keywords Time series, Moving average, Linear regression, Dynamic time warping(DTW)

1 引言

时间序列是某个物理量按时间顺序观测得到的一系列观测值,它反映了实体属性随时间变化的特征。现实世界中存在大量的时间序列,如环境监测、物联网、气象研究、网络安全、金融等领域产生了大量的时间序列[1]。时间序列相似性搜索是对一批时间序列数据进行分类、聚类、查询、预测的一项基础任务。而时间序列的近似表示和相似性度量又是序列相似性搜索的关键问题[2],对相似匹配的结果起着决定性作用。

时间序列相似性度量通常采用欧氏距离和动态时间弯曲(Dynamic Time Warping, DTW)距离,欧氏距离不具备分段趋势信息,不能对序列的形态进行表征,此外,欧氏距离对序列的异常数据敏感,轻微的序列异常会导致序列之间的欧氏距离剧烈变动。DTW 是由 Keogh 等[3-4]提出的一种针对时间序列的动态时间弯曲距离方法。该方法对两个任意时间序列的数据构建对应关系,并从对应关系中搜索最优匹配路径,进而更加有效地度量时间序列的相似性。但是该方法的时间复杂度为 O(mn),其中 m,n 是两个时间序列的长度,一定程度上限制了其应用。针对欧氏距离和 DTW 的不足,文献[5]和文献[6]分别提出三元分段趋势模式和七元分段趋势模式,并通过计算模式距离来度量两个等长序列的趋势差异程度,但这种时间序列的相似匹配的精确度依赖于模式划分的颗粒度。文献[7]使用夹角距离来度量时间序列的相似性,具有平移和旋转不变性的优点,但是缺乏序列的分段趋势信息。文献[8]在动态时间弯曲(DTW)距离的基础上,提出了分段 DTW 距离的时间相似性度量方法,与传统的 DTW 距离相比,该方法在保证度量准确性的基础上提高了序列相似性计算的效率。文献[9]提出了分段聚合的动态时间弯曲距离方

本文受天津市科技支撑计划项目(15ZXHLGX00210, 14ZCDZGX00811),天津市科技支撑计划(13ZCZDGX01200),天津市产学研合作项目(14ZCZDSF00025),天津市 863 成果转化项目(13RCHZGX01116),天津市 863 成果转化项目(14RCHZGX00862)资助。
冯玉伯(1974-),男,博士,高级工程师,主要研究方向为物联网、机器学习、数据挖掘;**丁承君**(1973-),男,博士,教授,博士生导师,主要研究方向为移动机器人智能控制、嵌入式计算机系统,E-mail:190532210@qq.com(通信作者);**高 雪**(1991-),女,硕士生,主要研究方向为移动机器人智能控制、嵌入式计算机系统;**朱雪宏**(1987-),博士,主要研究方向为移动机器人智能控制、嵌入式计算机系统;**刘 强**(1993-),男,硕士生,主要研究方向为移动机器人智能控制、嵌入式计算机系统。

法，实现了时间序列的有效降维和特征表示，具有算法过程简单的特点。文献[10]针对时间序列分段线性表示相似度量方法存在的序列长度依赖和多分辨率条件下的潜在识别误差等缺点，提出了一种序列分段线性弧度表示和基于弧度距离的相似度量方法。实现了序列的快速在线分割与相似度量计算。

针对时间序列相似性度量欧氏距离与动态时间弯曲距离的局限性，提出了一种基于滑动平均与分段线性回归的时间序列相似性度量方法。该方法首先对时间序列做初始可变滑动平均处理，以滑动平均处理的时间序列上的局部极值点作为特征点，进而对相邻特征点之间的时间序列做线性回归。将线性回归方程的系数(即截距与斜率)作为时间序列的分段特征表示，并在特征空间中使用 DTW 对时间序列的相似性进行度量。该算法具有对异常数据的不敏感性且有效地实现了数据降维和特征表示，且该算法的效率要远高于 DTW 距离算法。

2　滑动平均与分段线性回归的时间序列相似性

2.1　相关定义与模型

定义 1(时间序列)　一系列有序数据。

$$Y=\{y_j:i=0,1,2,\cdots,n\},\forall\ y_i\in R$$

其中，R 是实数集，i 为采样时间点，一般为等时间间隔。

定义 2(时间序列长度)　时间序列 Y 的元素个数，记为 $|Y|$。

定义 3(子序列)　从原序列 Y 中选取一个片段而形成的新序列，记为 sub(Y)。且满足 $|\text{sub}(Y)|<|Y|$。

初始可变滑动平均模型：给定一个时间序列 Y 和长度为 m 的时间窗口，根据式(1)对时间序列求取算术平均值。

$$s_i=\begin{cases}\dfrac{\sum_{j=0}^{i}y_j}{i+1}, & i\leqslant m-1\\ \dfrac{\sum_{j=0}^{m-1}y_{i-j}}{m}, & i>m-1\end{cases}\tag{1}$$

定义 4　设时间序列为 Y，Y 的初始可变滑动平均为 S，$S=\{s_i,i=0,1,\cdots,n\}$，$|S|=|Y|$，F 为 Y 的特征序列。则：

$$\begin{aligned}&s_0,s_n\in F\\&(s_{i-1}<s_i)\wedge(s_{i+1}<s_i)\wedge(0<i<n)\Rightarrow s_i\in F\\&(s_{i-1}>s_i)\wedge(s_{i+1}>s_i)\wedge(0<i<n)\Rightarrow s_i\in F\\&(s_{i-1}=s_i)\wedge(s_{i+1}<s_i)\wedge(0<i<n)\Rightarrow s_i\in F\end{aligned}\tag{2}$$

使用 $adj(s_i,s_j)=$TRUE 表示 F 中任意两个相邻特征点。

按特征序列对原始时间序列进行子序列划分，则：

$$\begin{aligned}&sub(Y)=\{y_i,y_{i+1},\cdots,y_{i+k}\}\\&y_i\in F\wedge y_{i+k}\in F\wedge adj(y_i,y_{i+k})=\text{TRUE}\end{aligned}\tag{3}$$

时间线性趋势模型，设原始时间序列的特征子序列的确定时间趋势为：

$$u_t=\beta_0+\beta_1 t\tag{4}$$

其中，β_0 和 β_1 分别表示斜率与截距，选取最小二乘法估计 β_0 和 β_1，使得：

$$(b_0,b_1)=\min\left(\sum_{t=1}^{k}[y_t-(b_0+b_1 t)]^2\right)\tag{5}$$

计算式(5)关于 β_0 和 β_1 的偏导数，求解线性方程组。使用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 表示得到的解：

$$\begin{aligned}&\hat{\beta}_1=\frac{\sum_{t=1}^{k}(y_t-\bar{y})(t-\bar{t})}{\sum_{t=1}^{t=n}(t-\bar{t})^2}\\&\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{t}\end{aligned}\tag{6}$$

其中，$\bar{t}=(n+1)/2$ 是 $1,2,\cdots,n$ 的平均数。

2.2　动态时间弯曲距离

动态时间弯曲距离(Dynamic Time Warping，DTW)最先在语音处理方面得到广泛研究[11-12]，并由 Berndt 等最先引入到数据挖掘领域中[13]。

定义 5　时间序列 $X=\{x_i,i=0,1,\cdots,m\}$ 和 $Y=\{y_i,i=0,1,\cdots,n\}$ 之间的动态时间弯曲距离定义为：

$$\begin{aligned}&D_{tw}(\langle\rangle,\langle\rangle)=0\\&D_{tw}(X,\langle\rangle)=D_{tw}(\langle\rangle,Y)=\infty\\&D_{tw}=d(x_0,y_0)+\min\begin{cases}D_{tw}(X,Rest(Y))\\D_{tw}(Rest(X),Y)\\D_{tw}(Rest(X),Rest(Y))\end{cases}\\&d(x,y)=\|x-y\|,x\in X,y\in Y\end{aligned}\tag{7}$$

其中，$Rest(X)=x_2,x_3,\cdots,x_m$，$Rest(Y)=y_2,y_3,\cdots,y_m$。

时间序列 X，Y 的数据点可以组成距离矩阵：

$$D_{m\times n}=\{d(i,j)\}_{m\times n}\tag{8}$$

其中，$0\leqslant i\leqslant m$，$0\leqslant j\leqslant n$，$d(i,j)=\|x_i-y_j\|$。

如图 1 所示，图中 U 对应一条动态时间距离路径，动态时间弯曲距离可使用动态规划方法来计算[14]，其基本思想是寻找一条具有最小弯曲代价(如最小累计距离)的最优路径，算法的时间复杂度为 O($|X|\cdot|Y|$)。

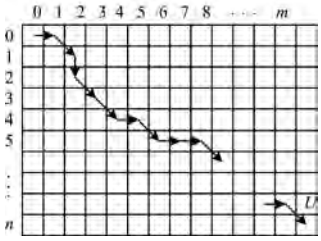

图 1　动态时间弯曲距离路径

3　算法设计

由于动态时间弯曲距离路径算法的时间复杂度为 O($|X|\cdot|Y|$)，当 X，Y 为高维时间序列时，使用 DTW 对 X，Y 进行相似性距离度量，时间消耗太大，不能应用于大规模时间序列的相似性比较与数据挖掘。因此，在使用 DTW 进行时间序列相似性比较之前，通常先对时间序列进行预处理，通过预处理对时间序列进行数据降维，该过程中需要保证近似准确性。

基于滑动平均与线性回归的时间序列相似性的基本思想：

1)使用初始可变滑动平均模型(见式(1))对原始时间序列 X，Y 做滑动平均处理，得到滑动平均序列 S_X 和 S_Y。

2)在滑动平均序列 S_X 和 S_Y 中提取特征序列 F_X 和 F_Y(见式(2))。

3)按特征序列 F_X 和 F_Y 对原始序列 X，Y 进行子序列划分(见式(3))。

4)针对 X，Y 的子序列，求取时间线性趋势模型(见式(6))

5)针对 X，Y 的子序列线性趋势模型，计算 DTW 距离。

3.1　初始可变滑动平均模型算法

初始可变滑动平均模型算法如算法 1 所示。

算法 1 *function VMA*(X,w)

```
输入:原始时间序列 X;滑动窗口大小 w
输出:初始可变滑动平均序列 S
begin:
  S[0]=X[0];#初始化滑动平均序列
  cum=X[0];#初始化累加器
  for i=2 to length(X) do
    if (i<=w) then #当前序列点的位置小于或等于滑动窗口的
                    大小
    begin
      cum=cum+dat[i];
      S[i]=cum/i;#计算初始滑动平均值
      end;
      else #当前序列点的位置大于滑动窗口大小
        begin
          cum=cum+dat[i]-dat[i-w];
          S[i]=cum/w;#计算滑动平均值
    end
  end
  return(S);
end.
```

该算法的时间复杂度为 O(|X|)。

3.2 特征序列提取算法

特征序列提取算法如算法 2 所示。

算法 2 function *getFeatures*(S)

```
输入:初始可变滑动平均时间序列 S
输出:特征点对应的时间点序列
begin
  F=[]; #初始化时间序列特征时间点
  F.append(0); #时间序列的初始时间点是特征时间点
  for i=1 to len(S)-1 do #遍历滑动平均时间序列
    if(S[i-1]<S[i]) and S[i+1]<S[i])
    begin
      F.append(i);
    end
    if(S[i-1]>S[i] and S[i+1]<S[i])
    begin
      F.append(i);
    end
  F.append(len(S));#时间序列的最终时间点是特征时间点
  return(F);
end.
```

该算法的时间复杂度为 O(|S|)。

3.3 时间序列线性趋势算法

时间序列线性趋势算法如算法 3 所示。

算法 3 function *tsLineRegr*(X,F)

```
输入:原始时间序列 X,特征时间点序列 F
输出:分段时间序列线性模型的截距与斜率
begin
  for i=0 to len(F)-1 #遍历所有特征时间点,划分子序列
  begin
    l=X[F[i]:F[i+1]]; #获取子序列特征值
    t=F[[i]:F[i+1]]; #获取子序列对应的特征时间点
    regr.fit(t,l); #对子序列做线性回归
    a,b=regr.coef,regr.intercept; #获取子序列线性回归的斜率
                                   和截距
    Features.append([a,b]); #保存每一个子序列线性回归的斜率
                             与截距
  end
return(Features);
end
```

该算法的时间复杂度为 O(|X|)。

对时间序列 X 应用上述算法,即经由初始可变滑动平均、特征序列提取、时间序列线性趋势算法,可将时间序列 X 映射到以分段直线的截距和斜率为表征的特征空间,这种方法实现了时间序列的数据压缩,减少了时间序列的数据量,同时保留了序列的局部特征和趋势特征,为度量时间序列的相似性提供了支撑。图 2 为随机游走产生的时间序列图及采用滑动窗口为 20 生成的分段线性回归图。原始数据共 1000 个数据点,经由算法处理后,变换为有 93 个(斜率、截距)表征的特征序列,数据压缩率为 81.4%。由图 2 可以看出,特征序列良好地反映了原始序列的局部和整体的变化趋势。同其他算法相比,上述算法可以通过调整滑动窗口的大小得到不同数据量的特征序列,有利于根据实际情况调整数据压缩比和计算速度。

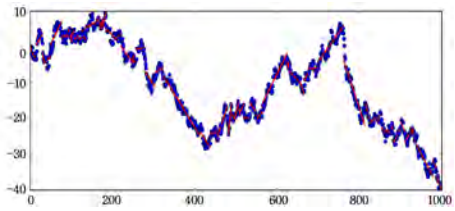

图 2 随机游走时间序列图及其分段线性回归映射

3.4 滑动平均与线性回归的 DTW 距离(MALRDTW)

设 X,Y 为原始时间序列,X_F,Y_F 为特征序列。使用 MALRDTW 度量 X 和 Y 的相似性的算法如下。

算法 4 function *MALRDTW*(X,Y)

```
输入:时间序列 X,Y
输出:时间序列 X,Y 的相似度
begin
  X_MV=VMA(X,w1);Y_MV=VMA(Y,w2);
  XFeatures=getFeatures(X_MV);YFeatures=getFeatures(Y_
  MV);
  XF=tsLineRegr(X,XFeatures);YF=tsLineRegr(Y,YFeatures);
  Dist=DTW(XF,YF);
  return(Dist);
end.
```

在原始时间序列上直接运用 DTW 距离算法计算时间序列的相似性,计算复杂度为 O(|X||Y|);而采用 MALRDTW 距离算法的复杂度为 O(|X_F||Y_F|)。而|X_F||Y_F|的大小依赖于滑动平均窗口的大小,当选择合适的窗口使得数据压缩率达到 80%时,计算复杂度将减少近 90%。

4 实验比较

为了对 MALRDTW 时间序列相似性算法的性能进行评估,对 DTW 距离算法和 MALRDTW 距离算法进行了比较,将算法的运行时间和相对距离比作为衡量指标。之所以采用相对距离比作为衡量指标,主要是因为 DTW 距离算法是在原始序列空间进行累计距离度量,而 MALRDTW 距离算法是在经过初始可变滑动平均和子序列线性回归变换后,在原始序列的子序列线性回归的截距与斜率的特征空间中计算累计距离,因此二者计算的累计距离不是在同一空间下进行的。

由于随机游走模型是布朗运动理想的数学模型，可以应用于互联网链接分析和金融股票市场，故测试数据使用随机游走模型产生 5 条拥有 1000 个数据点的时间序列，对 5 个时间序列进行相似性比较。算法在同一台计算机上运行，使用 Python 为算法的设计软件，对 DTW 和 MALRDTW 算法进行测试，算法的运行时间及相似比较结果如表 1 所列。表 1 的实验结果说明，DTW 与 MALRDTW 动态时间弯曲距离算法在时间序列上进行相似度度量时结果相近，说明二者之间的准确度近似。但对于算法的运行效率而言，MALRDTW 距离算法远高于 DTW 算法，约提高 96%，这一点在高维度时间序列相似性比较中表现得更为突出。

表 1　DTW 距离算法与 MALRDTW 距离算法的实验结果比较

距离度量方法	DTW 距离算法			MALRDTW 距离算法		
	累积距离	相对距离比	CPU 运行时间/s	累积距离	相对距离比	CPU 运行时间/s
1	10.056	0.0000	0.031	506.346	0.0000	0.001
2	56.209	5.5896	0.032	4320.58	8.5329	0.001
3	87.440	8.6953	0.031	4564.18	9.0140	0.001
4	52.130	5.1840	0.035	4250.00	8.3935	0.004
5	140.500	13.9718	0.035	6977.24	13.7796	0.001

结束语　针对时间序列相似性比较中欧氏距离对序列的异常数据敏感和动态时间弯曲距离时间复杂度为$O(mn)$的问题，提出基于滑动平均与分段线性回归的时间序列相似性算法。算法利用初始可变滑动平均算法对原始时间序列进行预处理，消除了由原始时间序列中的异常数据带来的不利影响，同时使得时间序列更加平滑；在滑动平均序列上提取极值特征点，并以特征点所对应的时间点对原始时间序列进行子序列划分，应用线性回归算法对子序列进行处理；将线性回归的截距和斜率作为原始时间序列的特征序列，实现了数据降维处理。

使用滑动平均与分段线性回归处理后的动态时间弯曲距离算法(MALRDTW)取得了与 DTW 算法相近的相似性比较性能，但是在算法效率上明显优于 DTW 距离算法。

参 考 文 献

[1] 李海林，杨丽彬. 时间序列数据降维和特征表示方法[J]. 控制与决策，2013，28(11)：1718-1722.

[2] Al-NAYMAT G，TAHERI J. Effects of dimensionality reduction techniques on time series similarity measurements[C]// IEEE/ACS International Conference on Computer Systems and Applications. Piscataway：IEEE，2008：188-195.

[3] KEOGH E，RATANAMAHATANA C A. Exact indexing of dynamic time warping[J]. Knowledge and Information Systems，2005，7(3)：358-386.

[4] KEOGH E，PAZZANI M. Derivative dynamic time warping[C]// The First SIAM International Conference on Data Mining. Washington：IEEE，2001：1-11.

[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] DONG X L，GU C K，WANG Z O. Study on Time Series Similarity Measurement Based on Morphology [J]. Journal of Electronics & Information Technology，2007，29(5)：1228-1231.

[7] 张鹏，李学仁，张建业，等. 时间序列的夹角距离及相似性搜索[J]. 模式识别与人工智能，2008，21(6)：763-767.

[8] 陆薛妹，胡轶，方建安. 基于分段极值 DTW 距离的时间序列相似性度量[J]. 微计算机信息，2007，23(27)：204-206.

[9] 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.

[10] 朱天，白似雪. 基于模式距离度量的时间序列相似性搜索[J]. 微计算机信息，2007，23(30)：216-217.

[11] RABINER L，JUANG B H. Fundamentals of Speech Recognition[J]. Tsinghua University Press，1993，1(1)：353-356.

[12] VULLINGS H J L M，VERHAEGEN M H G，VERBRUGGEN H B. ECG segmentation using time-warping[M]// Advances in Intelligent Data Analysis Reasoning about Data. Springer Berlin Heidelberg，1997：275-285.

[13] BERNDT D J，CLIFFORD J. Using dynamic time warping to find patterns in time series[C]// KDD workshop. Seattle：AAAI Press，1994：359-370.

[14] BERNDT D J，CLIFFORD J. Finding patterns in time series：a dynamic programming approach[C]// Advances in Knowledge Discovery & Data Mining. Washington：American Association for Artificial Intelligence，1996：229-248.

(上接第 96 页)

[15] KIM S N，KAN M Y. Re-examining automatic keyphrase extraction approaches in scientific articles[C]// Proceedings of the ACL-IJCNLP Workshop on Multiword Expressions. USA：ACL，2009：9-16.

[16] LOPEZ P，ROMARY L. HUMB：Automatic key term extraction from scientific articles in GROBID[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. Sweden：ACM，2010：248-251.

[17] JIANG X，HU Y H，LI H. A ranking approach to keyphrase extraction[C]// 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York：ACM，2009：756-757.

[18] HUANG Z H，XU W，YU K. Bidirectional LSTM-CRF Models for Sequence Tagging(arXiv)(Version1. 0)[OL]. https://arxiv. org/abs/1508. 01991.

[19] BENGIO Y，DUCHARME R，VINCENT P，et al. A neural probabilistic language model[J]. Journal of Machine Learning Research，2003，3(6)：1137-1155.

[20] COLLOBERT R，WESTON J，BOTTOU L，et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research，2011，12(1)：2493-2537.

[21] MIKOLOV T，YIH W T，ZWEIG G. Linguistic regularities in continuous space word representations[C]// NAACL-HLT. USA：ACL，2013：746-751.

[22] LEVY O，GOLDBERG Y，DAGAN I. Improving distributional similarity with lessons learned from word embeddings[J]. Transactions of the Association for Computational Linguistics，2015，75(3)：211-225.

[23] LAMPLE G，BALLESTEROS M，SUBRAMANIAN S，et al. Neural Architectures for Named Entity Recognition (arXiv)(Version3. 0)[OL]. https://arxiv. org/abs/1603. 01360.

[24] LAFFERTY F，MCCALLUM A，PEREIRA F. Conditional Random Fields：Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of ICML-2001. New York：ACM，2001：282-289.