第35卷 第24期 计 算 机 工 程 2009年12月
Vol.35 No.24 Computer Engineering December 2009
·软件技术与数据库· 文章编号：1000—3428(2009)24—0054—03 文献标识码：A 中图分类号：TP391.4

基于参数重要度的多元时间序列相似性查询

毛红保[1]，张凤鸣[1]，冯 卉[2]，吕慧刚[1]
(1. 空军工程大学工程学院，西安 710038；2. 空军工程大学导弹学院，三原 713800)

摘 要：针对多元时间序列的相似性查询问题，给出参数重要度的定义，提出一种基于参数重要度的候选集查询方法。通过对多元时间序列的SVD分解，将奇异值向量和特征矩阵作为多元序列的特征，基于线性空间中的坐标变换原理构造2个多元时间序列的相似性度量模型，实现在候选集上的精确匹配并获得最终的结果集。对飞行数据的相似性查询实验验证了该方法的有效性。

关键词：多元时间序列；相似性查询；参数重要度；特征提取；相似性度量

Similarity Query in Multivariate Time Series Based on Parameter Importance Degree

MAO Hong-bao[1], ZHANG Feng-ming[1], FENG Hui[2], LV Hui-gang[1]
(1. Engineering Institute, Air Force Engineering University, Xi'an 710038;
2. Missile Institute, Air Force Engineering University, Sanyuan 713800)

【Abstract】 Aiming at the problem of multivariate time series similarity search, this paper presents the definition of parameter importance degree and puts forward a candidate sets obtaining method based on it. It extracts singular vector and eigenvector matrix as the features of multivariate time series by SVD, constructs similarity measure modal via coordinate transformation theory in linear space, realizes precise matching on candidate sets and gets ultimate results. Experiments on flight data similarity query show the validity of the method.

【Key words】 multivariate time series; similarity query; parameter importance degree; feature extraction; similarity measure

1 概述

自从文献[1]提出时间序列相似性搜索问题以来，目前的研究主要针对一元时间序列进行，多元时间序列上相似性查询的研究还不多见。但现实世界中大量模式和状态的刻画仅靠单一参数是不够的，在很多领域中多元时间序列(Multivariate Time Series, MTS)普遍存在，研究多元时间序列上的相似性查询非常必要。

将时间序列 $X=\langle x_1,x_2,\cdots,x_n\rangle$ 的每个元素 x_i 看成由 m 个参数组成的向量，便可表示为多元时间序列。长度为 n 的时间序列记为矩阵 $\boldsymbol{X}(\boldsymbol{X}\in R^{n\times m})$。在多元时间序列相似性查询中，代表序列模式的子序列也是一个矩阵而不是向量，图1为2个多元时间序列相似性查询的示意图。

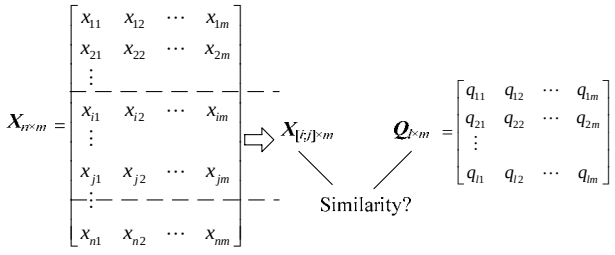

图1 多元时间序列相似性查询示意图

如何从多元时间序列中分割出潜在相似的子序列，并确定这些子序列与查询序列的相似度，是多元时间序列相似性查询的关键问题，这是因为：(1)多元时间序列的数据量大，为避免计算上的时间开销，在相似性查询时应该选取潜在有用的子序列进行相似性匹配。(2)通常多元时间序列的多个参数之间存在一定的关联，这些参数应该作为一个整体看待而不应该割裂开来[2]。(3)由于采样频率或模式持续时间的差异，通常2个相似子序列的长度并不相等[3]。

文献[2]提出一种通过主分量分析(PCA)进行多元时间序列相似性匹配的方法，并在Human Gait(根据人的行走姿态来进行身份识别)等多个数据库上的实验都取得了较好的效果。但它只给出了序列间相似性匹配的模型，而没有给出有效的序列分割和查询方法。文献[3-4]对CyberGlove的动作识别问题进行研究：文献[3]基于滑动窗口技术研究了动作的分割和匹配方法，它需要将所有分割后的序列与标准动作库中的每一个动作分别进行匹配；文献[4]则通过构建树形索引Interval-Tree来组织标准动作库，从而加速动作的识别过程。本文提出了一种基于参数重要度的多元时间序列分割方法，并构造2个多元时间序列的相似性度量模型，从而实现在候选集上的精确匹配并获得最终的结果集。

2 基于参数重要度的候选集查询

2.1 查询算法

考虑子序列匹配问题：给定多元时间序列 $\boldsymbol{X}\in R^{n\times m}$ 和查询序列 $\boldsymbol{Q}\in R^{l\times m}$，需要在某种相似性度量函数 $Sim(\cdot)$ 下查找所有与 $\boldsymbol{Q}$ 相似的子序列 $\boldsymbol{X}_{[i:j]\times m}$ $(1\le i\le j\le n)$。由于相似的子序

作者简介：毛红保(1979－)，男，讲师、博士研究生，主研方向：时间序列分析，智能决策；张凤鸣，教授、博士生导师；冯 卉，助教、硕士；吕慧刚，博士研究生

收稿日期：2009-05-04 E-mail：maohbao@126.com

列长度不一定相等，因此理论上可能的子序列非常多(正比于 n^2)，但通常这些子序列中大多数都是不满足相似性要求的。

查询序列 $Q_{l\times m}$ 代表用户关注的某个有意义的模式，其 m 个参数形成了 m 个一元时间序列，该模式是由这些一元时间序列的变化共同确定的。通常不同的参数对模式形成的贡献会有差异，例如有的参数只需要改变少许就会对模式产生质的影响，有的参数即使发生一定的变化原来的模式仍然成立，即它们对该模式而言重要度是不一样的。如果首先按照其中的一个参数以一元时间序列的方式进行查询，显然重要度高的参数得到的候选集数量会小于重要度低的参数，因为前者具有更强的约束和过滤能力。因此，按参数重要度的高低依次进行一元时间序列的子序列查询和全序列匹配，便能快速得到查询的候选集。该过程见图 2(按第一个参数进行子序列查询，在结果集中按第 2 个参数进行全序列匹配)。

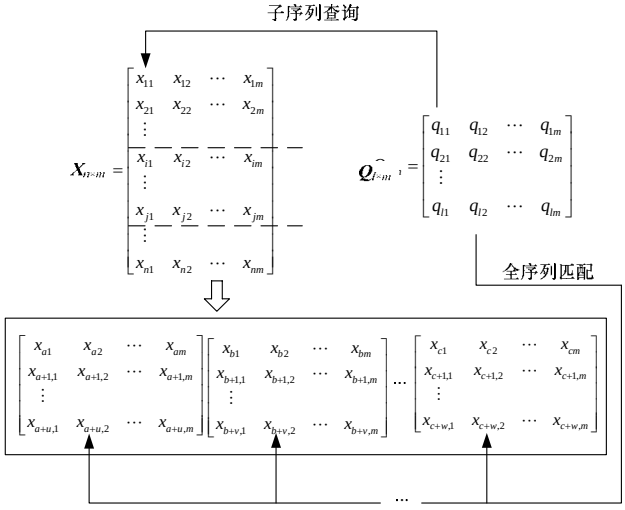

图 2　基于参数重要度的候选集查询过程

但根据上述方法得到的候选集并不是结果集，因为对于一个多元模式，其参数间存在内在的相互关联，即使候选集中所有的 m 个参数都分别满足相似性要求，它们组合在一起也不一定是满足要求的子序列。因此，需要对这些候选集进行更精确的相似性匹配，从而消除误报。

2.2　参数重要度的确定

给定多元时间序列 $\boldsymbol{X}_{n\times m}$ 和查询序列 $\boldsymbol{Q}_{l\times m}$ 后，参数的重要度可以通过 2 种方法确定：

(1)根据经验确定。对于一个有意义的查询模式，有经验的用户一般能够判断出哪些参数的特征是显著的，可以人为给出其中若干个参数的重要度顺序。

(2)根据序列的复杂度确定。该方法是笔者通过大量的工程实验得出的。复杂度即复杂性测度，它是非线性时间序列分析中一个重要的非线性指标。本文根据时间序列与查询序列参数复杂度之间的关系，定义参数查询的重要度指标。因为这种方法无需用户经验，适用于任意模式的相似性查询，并在实验中取得了较好的效果。

定义(参数重要度) 给定多元时间序列 $\boldsymbol{X}\in R^{n\times m}$ 和查询序列 $\boldsymbol{Q}\in R^{l\times m}$，则第 i 个参数 m_i 的查询重要度为

$$I(m_i)=1-\frac{|c(X(m_i))-c(Q(m_i))|}{c(X(m_i))+c(Q(m_i))} \tag{1}$$

其中，$X(m_i)$ 和 $Q(m_i)$ 分别表示 $\boldsymbol{X}$ 和 $\boldsymbol{Q}$ 中参数 m_i 的一元时间序列；$c(x)$ 表示序列 x 的相对复杂度，采用 Lempel-Ziv 算法[5]实现。

从该定义中可以看出，参数 m_i 的查询重要度与该参数对应的时间序列和查询序列的相对复杂度有关，两者相对复杂度越贴近(相对贴近)则该参数的重要度越高(趋近于 1)，否则该参数的重要度越低(趋近于 0)。

飞行数据是一种典型的多元时间序列数据，其中含有很多有意义的局部模式，这些模式的识别对飞机健康状态的监测及飞行动作的评估都具有重要意义。某飞行数据序列中的 2 个参数(倾斜角和俯仰角)及其子模式 P_1, P_2 如图 3 所示。

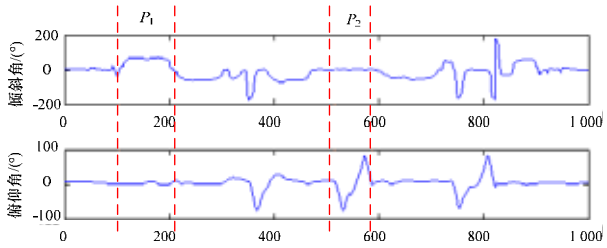

图 3　含有 2 个参数的飞行数据序列及其子模式 P_1, P_2

如果要查询其中的子模式 P_1 或 P_2，按 2.1 节的查询算法，首先应根据其中的一个参数进行子序列查询，并在查询结果中按照另一个参数进行全序列匹配。为了确定参数的查询顺序，各参数在不同序列中的相对复杂度及重要度计算结果如表 1 所示。

表 1　参数的相对复杂度及重要度计算结果

	倾斜角			俯仰角		
	全序列	P_1	P_2	全序列	P_1	P_2
相对复杂度	0.169 4	0.278 9	0.739 9	0.159 5	0.418 4	0.328 8
重要度	-	0.755 8	0.372 6	-	0.551 9	0.653 1

表 1 中的数据表明模式 P_1 的倾斜角参数重要度大于俯仰角参数重要度，模式 P_2 的俯仰角参数重要度大于倾斜角参数重要度。即查询 P_1 时应首先按倾斜角查询，查询 P_2 时应首先按俯仰角查询。这与直观上的观察结论是一致的，但对于参数更多、更复杂的模式而言，依靠直观的判断往往无法凑效，必须依据参数重要度这样的定量指标才能合理地确定参数的查询顺序。

3　多元模式的特征提取与相似性度量

3.1　特征提取

给定查询序列 $\boldsymbol{Q}\in R^{l\times m}$，基于 SVD 的序列特征提取方法分为如下 2 步：

(1)令 $\bar{\boldsymbol{Q}}=\boldsymbol{Q}^{\mathrm{T}}\times\boldsymbol{Q}$，则 $\bar{\boldsymbol{Q}}$ 为 $m\times m$ 的实对称阵；

(2) 对 $\bar{\boldsymbol{Q}}$ 进行 SVD 分解：$\bar{\boldsymbol{Q}}=\boldsymbol{U}\Sigma_{\boldsymbol{Q}}\boldsymbol{U}^{\mathrm{T}}$，其中，$\boldsymbol{U}=[u_1,u_2,\cdots,u_m]$，$\Sigma_{\boldsymbol{Q}}=diag(\sigma_{Q1},\sigma_{Q2},\cdots,\sigma_{Qm})$，并记奇异值向量 $[\sigma_{Q1},\sigma_{Q2},\cdots,\sigma_{Qm}]=\boldsymbol{\sigma}_{\boldsymbol{Q}}$。

执行上面的分解后，将奇异值向量 $\boldsymbol{\sigma}_{\boldsymbol{Q}}$ 和正交阵 $\boldsymbol{U}$(称为特征矩阵)作为多元序列 $\boldsymbol{Q}$ 的特征。进行这样提取的优势在于：

(1)将长度不同的模式序列统一到同一个尺度。通常情况下相似的时序数据模式在多次出现时持续的时间不一样，因此对应的数据序列长度也不一致，通过上述 SVD 特征提取后序列特征的维数只与参数个数 m 有关，而与序列的长度 l 无关，这使得 2 个非等长序列的相似性匹配成为可能。

(2)降低了模式序列的维数。对于一个有意义的模式通常 $l>>m$，因此，特征提取后对序列的降维效果非常显著，大

大减小了相似性匹配时的计算量。

(3)具有消除参数间关联性的作用。SVD 具有与 PCA(主成分分析)等价的提取多元数据主成分的能力，所以它能消除多个参数的表达冗余并保留数据主要特征。

3.2 相似性度量

给定 2 个矩阵 $\boldsymbol{Q}\in\boldsymbol{R}^{l\times m}$ (查询序列)和 $\boldsymbol{P}\in\boldsymbol{R}^{r\times m}$ (候选序列)。设其 SVD 分解为 $\bar{\boldsymbol{Q}}=\boldsymbol{Q}^{\mathrm{T}}\times\boldsymbol{Q}=\boldsymbol{U}\Sigma_{\boldsymbol{Q}}\boldsymbol{U}^{\mathrm{T}}$，$\bar{\boldsymbol{P}}=\boldsymbol{P}^{\mathrm{T}}\times\boldsymbol{P}=\boldsymbol{V}\Sigma_{\boldsymbol{P}}\boldsymbol{V}^{\mathrm{T}}$，在文献[2]的基础上给出如下 2 个多元时间序列模式 $\boldsymbol{Q}\in\boldsymbol{R}^{l\times m}$ 和 $\boldsymbol{P}\in\boldsymbol{R}^{r\times m}$ 的相似性度量公式：

$$Sim_{SVD}(\boldsymbol{Q},\boldsymbol{P})=\sum_{i=1}^{m}w_i\left|\langle u_i,v_i\rangle\right|=\sum_{i=1}^{m}w_i\left|\cos\theta_i\right| \tag{2}$$

其中，$\boldsymbol{U}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_m]$；$\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m]$。$\langle\boldsymbol{u}_i,\boldsymbol{v}_i\rangle$ 表示向量 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 的内积，因为 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 都为单位向量，因此，也可用 $\boldsymbol{u}_i$ 和 $\boldsymbol{v}_i$ 夹角的余弦表示。通过取绝对值，使度量时总是取 2 个向量夹角中的锐角($\cos(\pi-\theta)=-\cos(\theta)$)。$\boldsymbol{w}=[w_1,w_2,\cdots,w_m]$ 为各坐标方向上的权重向量，它由奇异值向量 $[\sigma_{Q1},\sigma_{Q2},\cdots,\sigma_{Qm}]$ 和 $[\sigma_{P1},\sigma_{P2},\cdots,\sigma_{Pm}]$ 共同确定，具体计算过程为

$$w_{Qi}=\sigma_{Qi}\Big/\sum_{j=1}^{m}\sigma_{Qj}\,,\ w_{Pi}=\sigma_{Pi}\Big/\sum_{j=1}^{m}\sigma_{Pj}\qquad(i=1,2,\cdots,m)$$

$$w_i=(w_{Qi}+w_{Pi})\Big/(\sum_{j=1}^{m}w_{Qj}+\sum_{j=1}^{m}\sigma_{Pj})\qquad(i=1,2,\cdots,m) \tag{3}$$

即先将奇异值向量进行归一化处理，然后将两者相结合形成新的归一化向量形成权重。显然 $\sum_{i=1}^{m}w_i=1$，且 $0\le\left|\cos\theta_i\right|\le1$，因此，$Sim_{SVD}(Q,P)\in[0,1]$。

上面的相似性度量公式可理解为：首先，模式 Q 和 P 的相似性表示为各自特征矩阵的向量张成的坐标空间中每个坐标轴方向上相似性的累加；每个坐标轴方向上的相似性由坐标轴本身的夹角和该坐标方向权重(由奇异值)的乘积构成。通过 SVD 分解，将 2 个非等长模式的相似性转换为 R^m 空间中各坐标方向含权重的 2 个基底的相似性问题，体现了多元模式相似的实质。

有了任意 2 个多元模式精确的相似性度量模型，在给定查询序列 $\boldsymbol{Q}$ 和阈值 ε 的情况下，就可以对基于参数重要度获得的候选集进行过滤从而得到最终的查询结果。

4 实验分析

笔者基于飞行数据的飞行动作识别与评估为背景，验证本文提出的多元时间序列相似模式查询方法的有效性。从一个架次的飞行数据中识别某个飞行动作是否存在，并评价其完成情况，具有很强的工程应用价值。图 4 标示出了一个架次的飞行数据中 3 个含有飞行动作的数据段(图中标号分别为 1,2,3)。其中，所标动作依次为：水平“8”字，60°盘旋，急上升转弯。

在图 4 所示数据的基础上，将图中标示的第 2 个动作“60°盘旋”作为标准查询序列数据(t=813~1 016)，执行如下的数据预处理工作：将航向角的变化区间由[0°,360°]转化为 $(-\infty,+\infty)$，消除 0°与 360°之间的突变；对所有参数进行去均值和去标准差处理。然后根据 2.2 节给出的参数重要度计算方法，得各参数查询顺序为<航向角,倾斜角,速度,高度>(俯仰角参数对该飞行动作而言无显著特征，因此，在一元查询时不考虑)。根据该参数顺序获得的查询结果集含 3 个飞行动作，分别为 t=1 107~1 215(A)，t=1 362~1 459(B)，t=1 683~1 791(C)。其中一元子序列匹配基于 DTW 距离度量，采用 Optimized Naive-Scan 查询策略[6]；一元全序列匹配采用 DTW 距离度量。一元匹配时需要各参数的查询阈值，本文对阈值的确定基于实践经验得出，分别为 ε(航向角) =0.5，ε(倾斜角)=8，ε(速度) =35，ε(高度)=40。

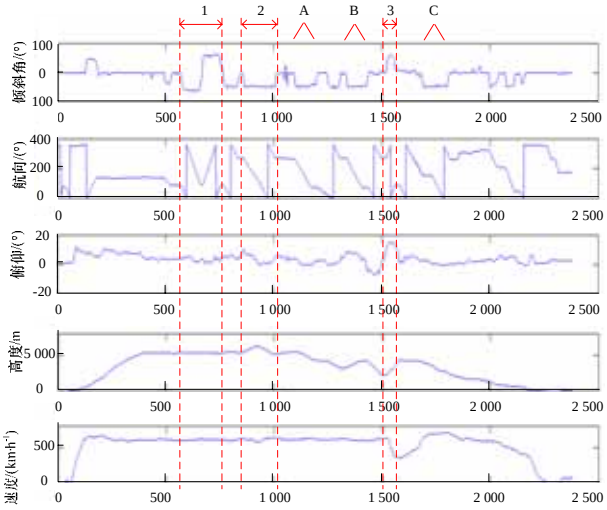

图 4 包含多种飞行动作的飞行数据曲线

在查询到的 3 个“60°盘旋”动作中，根据 3.2 节给出的相似性度量模型，它们与查询序列的相似度分别为 $sim(A)=0.718$，$sim(B)=0.913$，$sim(C)=0.866$。该度量结果与直观上的动作评估结论是一致的。

5 结束语

多元时间序列的相似性查询与一元时间序列的相似性查询既有联系又有区别。两者都会出现相似模式在长度上的伸缩变化，但多元时间序列需要克服多个参数间的相关性，并解决 2 个多元模式的相似性度量问题。本文的研究工作对于多元时间序列的相似性查询问题具有一定的理论意义和应用价值。

参考文献

[1] Keogh E, Kasetty S. On the Need for Time Series Data Mining Benchmarks: A Survey and Empirical Demonstration[C]//Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Alberta, Canada: [s. n.], 2002.

[2] Yang K, Shahabi C. A PCA-based Similarity Measure for Multivariate Time Series[C]//Proc. of MMDB'04. Washington D. C., USA: ACM Press, 2004.

[3] Li Chuanjun, Zhai Peng, Zheng Siqing, et al. Segmentation and Recognition of Multi-attribute Motion Sequences[C]//Proc. of MM'04. New York, USA: ACM Press, 2004.

[4] Li Chuanjun, Pradhan G, Zheng Siqing, et al. Indexing of Variable Length Multi-attribute Motion Data[C]//Proc. of MMDB'04. Washington D. C., USA: ACM Press, 2004.

[5] Lempel A, Ziv J. On the Complexity of Finite Sequences[J]. IEEE Transactions on Information Theory, 1976, 22(1): 75-81.

[6] Kim M S, Kim S W, Shin M. Optimization of Subsequence Matching Under Time Warping in Time-series Databases[C]//Proc. of the 20th Annual ACM Symposium on Applied Computing. New Mexico, USA: ACM Press, 2005.

编辑 金胡考