软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2013,24(11):2642−2655 [doi: 10.3724/SP.J.1001.2013.04464]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

辅助信息自动生成的时间序列距离度量学习*

邹朋成[1], 王建东[1], 杨国庆[2], 张 霞[1], 王丽娜[1]

[1](南京航空航天大学 计算机科学与技术学院,江苏 南京 210016)

[2](中国民航信息技术科研基地,天津 300300)

通讯作者: 王建东, E-mail: aics@nuaa.edu.cn

摘 要: 对于时间序列聚类任务而言,一个有效的距离度量至关重要.为了提高时间序列聚类的性能,考虑借助度量学习方法,从数据中学习一种适用于时序聚类的距离度量.然而,现有的度量学习未注意到时序的特性,且时间序列数据存在成对约束等辅助信息不易获取的问题.提出一种辅助信息自动生成的时间序列距离度量学习(distance metric learning based on side information autogeneration for time series,简称 SIADML)方法.该方法利用动态时间弯曲(dynamic time warping,简称DTW)距离在捕捉时序特性上的优势,自动生成成对约束信息,使习得的度量尽可能地保持时序之间固有的近邻关系.在一系列时间序列标准数据集上的实验结果表明,采用该方法得到的度量能够有效改善时间序列聚类的性能.

关键词: 度量学习;动态时间弯曲;辅助信息自动生成;时间序列聚类

中图法分类号: TP181 **文献标识码**: A

中文引用格式: 邹朋成,王建东,杨国庆,张霞,王丽娜.辅助信息自动生成的时间序列距离度量学习.软件学报,2013,24(11):2642−2655. http://www.jos.org.cn/1000-9825/4464.htm

英文引用格式: Zou PC, Wang JD, Yang GQ, Zhang X, Wang LN. Distance metric learning based on side information autogeneration for time series. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2642−2655 (in Chinese). http://www.jos.org.cn/1000-9825/4464.htm

Distance Metric Learning Based on Side Information Autogeneration for Time Series

ZOU Peng-Cheng[1], WANG Jian-Dong[1], YANG Guo-Qing[2], ZHANG Xia[1], WANG Li-Na[1]

[1](College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

[2](Information Technology Research Base of Civil Aviation Administration of China, Tianjin 300300, China)

Corresponding author: WANG Jian-Dong, E-mail: aics@nuaa.edu.cn

Abstract: An effective distance metric is essential for time series clustering. To improve the performance of time series clustering, various methods of metric learning can be applied to generate a proper distance metric from the data. However, the existing metric learning methods overlook the characteristics of time series. And for time series, it is difficult to obtain side information, such as pairwise constraints, for metric learning. In this paper, a method for distance metric learning based on side information autogeneration for time series (SIADML) is proposed. In this method, dynamic time warping (DTW) distance is used to measure the similarity between two time series and generate pairwise constraints automatically. The metric which is learned from the pairwise constraints can preserve the neighbor relationship of time series as much as possible. Experimental results on benchmark datasets demonstrate that the proposed method can effectively improve the performance for time series clustering.

Key words: metric learning; dynamic time warping; side information autogeneration; time series clustering

时间序列聚类在许多领域都有广泛应用,而对大量时序聚类任务而言,距离度量至关重要.一方面是由于聚类算法本身对度量非常敏感,另一方面是由于时序数据随时间动态变化的特性,导致常用的一些度量在处理时

* 基金项目: 国家自然科学基金(61139002)
收稿时间: 2013-01-06; 修改时间: 2013-07-12, 2013-08-02; 定稿时间: 2013-08-27

间序列数据时效果不佳.目前,时间序列最常用的度量方法有 L_p 范数距离(如欧几里德距离)和动态时间弯曲(dynamic time warping,简称 DTW)距离等.但欧氏距离不支持时间序列的动态特性,DTW 距离虽然考虑了时间序列的线性漂移和动态弯曲的特性,但由于不满足三角不等式,它并非严格的距离度量,影响了聚类性能.因此,为了提高时间序列聚类性能,需要寻求更适合的时间序列距离度量.

然而,寻找一种合适的度量并非易事,对于特定的时间序列数据集,它所处的特定度量空间无法准确获知,只能利用先验信息尽可能准确地估计这个度量.为了弥补传统距离度量的不足,研究人员已经提出了一系列距离度量学习算法,其中最常见的是学习优化一种马氏距离(Mahalanobis)度量.一类是利用带标签信息的数据进行学习,然而在很多的实际问题中,想要获取样本的类别标签代价非常大,因此,另一类学习算法转而采用一些更容易获取的数据信息,称为辅助信息(如成对约束:must-link 和 cannot-link).在很多情况下,判断两个样本是否相似要比直接获取样本类别容易得多,所以有许多度量学习算法都使用了成对约束.针对聚类任务的度量学习,最早由 Xing 等人[1]提出了一种使用成对约束学习马氏距离度量的凸优化方法.随后的改进方法包括相关成分分析(RCA)[2,3]、判别组成分析(DCA)[4]、基于迹之比优化的度量学习[5]等.为了充分利用无标签样本,在使用成对约束的同时,无标签的样本被考虑用于促使所学度量保持数据原本的拓扑结构[6−8].进一步地,距离度量学习算法经过核化扩展,用于非线性的度量学习[9,10].目前,距离度量学习已被应用到许多实际场景中,如人脸识别[11]、医学图像检索[12]、行为识别[13]等等.

但到目前为止,已有的度量学习方法仍需由用户提供辅助信息,而对于时间序列聚类等无监督问题,人为地给出成对约束等辅助信息十分困难,仍需付出很大代价.此外,已有的度量学习方法主要是针对静态的特征数据,还很少有针对时间序列设计的度量学习方法.最近,Prekopcsák 和 Lemire[14]提出了一种基于类别的时间序列马氏距离,但也仅适用于已知样本类别的情形.能否在缺乏监督信息的情况下,通过挖掘数据先验信息自动生成辅助信息,指导度量学习过程,是一个值得考虑的问题.

针对上述问题,本文提出一种辅助信息自动生成的时间序列距离度量学习(distance metric learning based on side information autogeneration for time series,简称 SIADML)方法.该方法在无标签信息的时间序列数据上,利用 DTW 方法挖掘时间序列样本间的近邻信息,自动生成一系列成对约束,作为辅助信息指导距离度量学习,免除了人力标注的代价.并且,辅助信息自动生成阶段具备较强的适应性,可独立于度量学习阶段,使生成的辅助信息方便地嵌入到不同的度量学习过程中.该方法促使度量学习过程能够充分利用时间序列的先验信息,约束新度量尽可能地保持时序样本间的近邻关系,同类样本中原本相似程度较高的样本对在新度量下依然保持高相似度.且所学度量满足距离度量的性质,能够更好地适用于时间序列聚类.最终,该方法在一系列时间序列标准数据集上进行了实验,与传统度量及人为提供辅助信息的度量学习方法比较,所学度量能够有效地提高时间序列聚类性能,验证了该方法的有效性.

1 相关工作

1.1 动态时间弯曲距离

动态时间弯曲(DTW)方法是一种常用的序列间相似度计算方法,最早由 Shakoes 和 Chiba[15]提出.其主要特点是考虑了时序事件中各阶段持续事件长度不一、时序间相位延迟等情形,因而度量结果贴近人的直觉观察.DTW 距离用于一阶最近邻分类,已经在大量时间序列数据集上得到了非常好的分类效果[16,17].

其不足之处是不满足三角不等式,即 $DTW(x,y)+DTW(y,z)\geqslant DTW(x,z)$.这一不满足距离度量的性质[18],导致由 DTW 距离得到的相似性不具备传递性,影响了在聚类中的应用.为此,研究者们对 DTW 进行了扩展,如 Shimodaira 等人[19]和 Cuturi 等人[20,21]利用 DTW 构造适用于时间序列的核度量,在一定程度上可弥补 DTW 的不足,并应用到人体运动行为分析、面部事件识别等复杂时序聚类问题中[22,23].但由于 DTW 距离计算需使用动态规划,复杂度较高,致使相应的时序聚类算法效率较低.

与许多其他度量不同,DTW 能够度量不同长度序列间的相似度.Ratanamahatana 和 Keogh[24]指出,直接比较两个不等长序列,和把这两个序列转换成等长再比较,两者在比较结果精度上的差异统计不显著.因此,在本文

的时间序列度量研究中,只需考虑等长序列间的度量.对于长度不相等的序列,可以先利用线性插值等方法将其转换成等长序列.

本文借助 DTW 获取相似样本的优势挖掘时间序列数据的信息,转而指导学习一种新的度量,提高时间序列聚类算法的效率和性能.

1.2 Mahalanobis距离度量学习

给定两个时间序列 $X_1=(x_1,x_2,\ldots,x_n)^T$ 和 $X_2=(x_1',x_2',\ldots,x_n')^T$,$X_1$ 和 X_2 关于矩阵 A 的马氏距离为

$$d_A(X_1,X_2)=(X_1-X_2)^TA(X_1-X_2) \tag{1}$$

矩阵 A 表示 X_1 和 X_2 的协方差矩阵的逆.但由于时间序列的高维特性,实际处理过程中协方差矩阵常常不可逆.Zoltán 等人[14]对此提出协方差矩阵求逆的方案,计算近似的马氏距离,处理时序分类任务.本文提出的时间序列度量学习与此有很大的不同,因为无需样本类标签,故能直接用于无监督时间序列聚类.

本文提及的距离度量学习已对上述距离做了扩展,矩阵 A 不再局限为协方差矩阵的逆,而只需是半正定矩阵,以保证 d_A 的度量性.A 包含了对原数据样本的线性变换,对 A 做 Cholesky 分解 $A=W^TW$,则

$$d_A(X_1,X_2)=(X_1-X_2)^TW^TW(X_1-X_2)=(W(X_1-X_2))^T(W(X_1-X_2)) \tag{2}$$

确定 A 或者 W,就能得到对应的度量函数.目前,基于成对约束的度量学习已受到越来越多的关注.成对约束包含 must-link 和 cannot-link,分别表示同类的样本对和不同类的样本对.Xing 等人[1]提出了一种使用成对约束学习度量矩阵的凸优化方法,能够有效地求得矩阵 A,但当数据维度较高时,计算效率很低.Xiang 等人[5]在 DCA[4]的基础上提出了一种使用迹之比构造的目标函数求解距离度量中的线性变换 W,模型如下:

$$W^*=\arg\max_{W^TW=I}\frac{tr(W^T\hat{S}_bW)}{tr(W^T\hat{S}_wW)} \tag{3}$$

其中,tr 表示矩阵的迹,$\hat{S}_b$ 是 cannot-link 约束集的协方差矩阵,$\hat{S}_w$ 是 must-link 约束集的协方差矩阵.引入约束 $W^TW=I$,是为了防止优化目标的解退化.最终使用一种二分搜索的方法迭代得到了最优解,求解过程不需要求协方差矩阵的逆,避免了奇异矩阵求逆的问题;并且提高了计算效率,有利于处理高维的时序数据.

Baghshah 等人[6]在 Xiang 等人[5]的工作基础上,除了使用 must-link 和 cannot-link 约束集以外,还考虑了大量无标签样本数据在输入空间的拓扑结构,将局部线性嵌入(locally linear embedding,简称 LLE)[25]的思想融入其中度量学习算法中,使样本在映射过后的空间中能够尽可能地保持数据的局部线性关系.

2 辅助信息自动生成的策略设计

时间序列不同于静态数据,它在输入空间中的每个维度是同一属性在不同时刻的表征,各个维度间有直观的相关性,因而用一般的度量方法(如欧氏距离)来度量时间序列数据并不一定合适.若构造一个恰当的距离矩阵 A,则通过对应的线性变换 W,将时间序列样本映射到新的向量空间中.时间序列在新空间下的属性是通过原属性的组合变换得到的,已经考虑了时间序列的特性.也就是说,通过学习矩阵 A,可以找到一个能够更恰当地表达序列数据间关系的新空间向量表示.此外,新的向量表示可对时间序列样本实现降维,有助于进一步提高时间序列聚类的效率.

现有度量学习方法往往需要由用户提供辅助信息,如成对约束.而在无监督的时间序列聚类实际问题中,让用户给定辅助信息需要付出很大的代价,且提供的辅助信息数量有限.为此,本文首先给时间序列度量学习提供了一种辅助信息自动生成的方法.该方法可在无标签信息的时间序列数据上,依据时间序列的动态弯曲特性,自动生成一些成对约束信息.

设 $X=\{x_i|i=1,2,\ldots,n\}$ 是一个时间序列数据集,样本个数为 n,长度为 d.令 $S=\{(x_i,x_j)\}$ 表示相似约束集,x_i 和 x_j 是一对相似样本.类似地,令 $D=\{(x_i,x_k)\}$ 表示相异约束集,x_i 和 x_k 是一对相异的样本.

为了自动生成度量学习所需的成对约束信息,本文考虑用 k 近邻方法寻找样本间的相似关系,将获得的近邻样本对构成相似样本约束集,而相距最远的样本对构成相异样本约束集.前人的研究工作表明,在不依赖距离

度量的情况下,一阶最近邻规则的渐近误差率的上界是 2 倍的贝叶斯分类误差[26].而在实际中,有限的训练样本数和样本的高维特性会导致最近邻方法性能下降,选择一个较好的度量十分重要.在时间序列数据上,DTW 距离在捕获时间序列平移和伸缩等特性时优势明显,而它在进行最近邻分类时取得的高分类精度印证了这一点.因此,用 DTW 作为度量寻找时间序列样本的近邻更有效.

辅助信息自动生成的详细过程如算法 1 所示.首先通过计算时间序列样本间的 DTW 距离,得到 DTW 距离矩阵δ.其中,δ_{ij}表示样本 x_i 和 x_j 之间的 DTW 距离,且$\delta_{ij}=\delta_{ji}$.然后,在 DTW 距离矩阵δ中为每一个时间序列样本搜索其一阶最近邻,使得每个样本和它的一阶最近邻样本形成一系列的时间序列相似样本对.这些样本对就作为一系列的相似成对约束添加到 S 集合中,最终构成了相似约束集.

算法 1. 辅助信息自动生成算法.

输入:时间序列数据集 $X=\{x_i|i=1,2,\ldots,n\}$,阈值 t.

输出:相似约束集 S 和相异约束集 D.

1. 计算 X 的 DTW 距离矩阵δ.$\delta_{ij}=\delta_{ji}=DTW(x_i,x_j)$.
2. 初始化相似约束集 S 和相异约束集 D:$S=\varnothing$,$D=\varnothing$.
3. 构造相似约束集 S:

   For i=1 to n

   $j=\text{argmin}_{i\neq j}\{\delta_{i1},\ldots,\delta_{ij},\ldots,\delta_{in}\}$, $S\leftarrow S\cup(x_i,x_j)$

   endfor
4. 构造相异约束集 D:

   While $count\_num\geqslant nt$:

   For 矩阵行号 i=1 to n

   $j=\text{argmax}_{i\neq j}\{\delta_{i1},\ldots,\delta_{ij},\ldots,\delta_{in}\}$,$D\leftarrow D\cup(x_i,x_j)$;

   endfor

   统计各样本在一阶最不近邻中出现的次数 $count[j]$,$count\_num\leftarrow\max(count[j])$;

   如果 $count\_num\geqslant nt$,$D\leftarrow\varnothing$,则计算最不近邻时剔除第 j 个样本.

   endwhile
5. 结束.

与相似约束集的生成方式类似,再从距离矩阵δ中搜索每个时间序列样本的一阶最不近邻,即与特定时间序列样本的DTW距离最大的那个样本,将该样本和它的一阶最不近邻作为一对相异成对约束添加到D集合中,最终构成相异成对约束集.值得注意的是,相异约束集的构造与相似约束集并不完全一致.在寻找各个样本的一阶最不近邻时,需要考虑孤立样本点的存在,并避免孤立点对相异成对约束集有效性的影响.例如,若某个时间序列样本远离其他所有的样本,则该样本将出现在所有的相异成对约束中,使得相异成对约束仅仅指导该样本在新度量下依然保持孤立,而忽略其他样本间的距离关系,严重削弱了相异约束的作用.为此设置了阈值 t($0<t\leqslant 1$),若样本 x_i 在一阶最不近邻中出现的次数与一阶最不近邻对总数之比大于 t,则视样本 x_i 为孤立样本.剔除样本 x_i,再重新计算各样本的一阶最不近邻.t取值越大,剔除的样本孤立程度就越高.当取 t=1 时,需要剔除的样本为其余每个样本的最不近邻;当 t=0.5 时,意味着需要剔除的某个样本是样本总数中 50%样本的最不近邻.因此,t 取值越小,需要剔除的样本孤立程度就越小,而剔除的样本数量也就越多,剩余的相异约束信息也会逐渐减小.在算法实际运行过程中,依据实际数据集的情况调节 t 的取值,使得在去除孤立程度较高样本的同时,控制相异约束信息的减少.

这里,集合 S 和 D 是利用 DTW 距离在无标签数据上自动获得的,不同于利用类别标签直接构造由同类样本对和不同类样本对形成的约束集,其正确率依赖于 DTW 最近邻分类的准确性.

由于 DTW 复杂度较高,直接增加了辅助信息生成过程的时间代价,从算法实用性角度考虑,有必要进一步提高算法效率.通过分析算法 1 发现,最终生成的辅助信息仅需寻找每个样本的最近邻和最不近邻样本,及对应

于这些样本对的DTW距离.依据该分析,辅助信息自动生成过程可看成是基于DTW的时间序列相似搜索问题.目前,已经有大量的研究工作用于提高 DTW 相似搜索的效率,一类主要的方法是边界距离过滤,即寻找一个低计算成本、相对真实 DTW 距离更小的下界函数 DTW_{LB} 以过滤掉满足 $DTW_{LB}\geqslant\varepsilon$ 的时间序列,或一个低计算成本、稍大于真实 DTW 距离的上界函数 DTW_{UB},以筛出那些满足 $DTW_{UB}\leqslant\varepsilon'$ 的时间序列(其中,ε 和 ε' 分别表示当前已找到的最近邻和最远邻所对应的距离)[27].已经提出来的 DTW 距离的上下界距离函数包括 LB_Kim[28], LB_Keogh[18],LB_Z[29],UB_Z[29]等.因而,在自动生成辅助信息过程中,引入已有的边界函数完全可以避免计算整个的 DTW 矩阵,有效降低时间复杂度.本文采用边界函数 LB_Z 和 UB_Z,其计算复杂度为 $O(d)$,优于 DTW 的 $O(d^2)$,边界函数具体计算公式见文献[29],本文不再赘述.改进后的算法见算法 2,通过引入更紧致的边界函数能够尽可能地减少 DTW 的计算次数,且不会影响最终找到的成对样本,因而算法 1 和算法 2 最终获得的相似约束集 S 和相异约束集 D 是一致的.

算法 2. 改进的辅助信息自动生成算法.

输入:时间序列数据集 $X=\{x_i|i=1,2,\ldots,n\}$,阈值 t.

输出:相似约束集 S 和相异约束集 D.

1. 初始化 X 的 DTW 距离矩阵 δ:$\delta_{ij}=0$,初始化相似约束集 S 和相异约束集 D:$S=\varnothing$,$D=\varnothing$.
2. 构造相似约束集 S:
   For i=1 to n
   初始化到样本 x_i 的最小距离:$min\_dist(i)\leftarrow Inf$;
   For j=1 to n and $j\neq i$
   计算样本 x_i 和 x_j 的 DTW 距离下界:$LB\_dist\leftarrow LB\_Z(x_i,x_j)$;
   如果 $LB\_dist<min\_dist(i)$
   计算 x_i 和 x_j 的真实 DTW 距离:$true\_dist\leftarrow DTW(x_i,x_j)$,$\delta_{ij}\leftarrow true\_dist$;
   若 $true\_dist<min\_dist(i)$,则更新 $min\_dist(i)\leftarrow true\_dist$,$index\_min(i)\leftarrow j$.
   endfor
   $S\leftarrow S\cup(x_i,x_{index\_min(i)})$
   endfor
3. 构造相异约束集 D:
   while $count\_num\geqslant nt$:
   For i=1 to n
   初始化到样本 x_i 的最大距离:$max\_dist(i)\leftarrow 0$
   For j=1 to n and $j\neq i$
   计算样本 x_i 和 x_j 的 DTW 距离上界:$UB\_dist\leftarrow UB\_Z(x_i,x_j)$;
   如果 $UB\_dist>max\_dist(i)$
   计算 x_i 和 x_j 真实 DTW 距离:$true\_dist\leftarrow DTW(x_i,x_j)$,$\delta_{ij}\leftarrow true\_dist$;
   若 $true\_dist>max\_dist(i)$,则更新 $max\_dist(i)\leftarrow true\_dist$,$index\_max(i)\leftarrow j$.
   endfor
   $D\leftarrow D\cup(x_i,x_{index\_max(i)})$
   endfor
   统计各样本在一阶最不近邻中出现的次数 $count[j]$,$count\_num\leftarrow\max(count[j])$;
   如果 $count\_num\geqslant nt$,$D\leftarrow\varnothing$,则计算最不近邻时剔除第 j 个样本.
   endwhile
4. 结束.

3 时间序列距离度量学习

自动生成的辅助信息将指导学习新的时序度量,所学度量也将尽可能地保持时间序列间固有的近邻关系.该方法可以作为时间序列聚类的一个初始化步骤,将习得的新度量直接应用于聚类.该方法的内容主要分为两个部分:第 1 部分是基于时间序列近邻保持的度量学习目标函数设计,第 2 部分是关于距离矩阵 A 的目标函数求解.

3.1 时间序列近邻保持

利用算法 1 或算法 2 生成的成对约束信息,即可用于学习新的距离度量.这些辅助信息不仅包含了哪些样本对相似和哪些样本对不相似,而且可以通过 DTW 距离进一步了解这些样本对之间的相似程度,指导学习新度量,促使相似度较高的样本对在新度量下依然保持高相似度,反之亦然.为此,本文设计了如下的目标函数:

$$W^* = \arg\max_W \frac{\sum_{(x_i,x_j)\in D} c'_{ij} \| W^T x_i - W^T x_j \|^2}{\sum_{(x_i,x_j)\in S} c_{ij} \| W^T x_i - W^T x_j \|^2} \tag{4}$$

其中,优化目标 W 就是距离度量学习所需的线性变换矩阵.W 将输入空间的时间序列样本 x 映射到新的向量空间,即 $y=W^Tx$,度量矩阵 $A=WW^T$.分母部分是带权的相似成对约束的距离和,类似地,分子部分是带权的相异成对约束的距离和.优化目标是使相似的样本对之间的距离和尽可能地小,而同时使相异的样本对之间的距离和尽可能大.

引入权值 c_{ij} 和 c'_{ij} 是为了在新空间中尽可能地保持样本间原有的近邻程度.DTW 距离越小,表示两个样本相似程度越高.例如,若 $\delta_{ij}<\delta_{mn}$,则优化目标期望 $d_W(x_i,x_j)<d_W(x_m,x_n)$.权值 c_{ij} 可以通过公式(5)计算得到:

$$c_{ij} = \exp\left(-\delta_{ij} \Big/ \sum_{(x_s,x_t)\in S} \delta_{st}\right),\ \forall (x_i,x_j)\in S \tag{5}$$

δ_{ij} 是样本 x_i 和 x_j 间的距离,δ_{ij} 越小,对应的 c_{ij} 越大,指数函数的使用使得权重值对近邻程度的差异更加敏感.最大化目标函数(4),促使在目标空间中 $d_W(x_i,x_j)$ 尽可能地小,保持样本 x_i 和 x_j 之间较高的相似度.

类似地,权值 c'_{ij} 可以通过公式(6)计算得到.对于差异较大的样本对,其对应的 c'_{ij} 较大.

$$c'_{ij} = \exp\left(\delta_{ij} \Big/ \sum_{(x_s,x_t)\in D} \delta_{st}\right),\ \forall (x_i,x_j)\in D \tag{6}$$

为了求解目标函数,需要再添加一项正交约束 $W^TW=I$,防止解的退化.对目标函数做如下变形:

$$\begin{aligned} W^* &= \arg\max_{W^TW=I} \frac{\sum_{(x_i,x_j)\in D} c'_{ij} \| W^T x_i - W^T x_j \|^2}{\sum_{(x_i,x_j)\in S} c_{ij} \| W^T x_i - W^T x_j \|^2} \\ &= \arg\max_{W^TW=I} \frac{\sum_{(x_i,x_j)\in D} c'_{ij} (x_i-x_j)^T WW^T (x_i-x_j)}{\sum_{(x_i,x_j)\in S} c_{ij} (x_i-x_j)^T WW^T (x_i-x_j)} \\ &= \arg\max_{W^TW=I} \frac{tr\left(W^T\left(\sum_{(x_i,x_j)\in D} c'_{ij}(x_i-x_j)(x_i-x_j)^T\right)W\right)}{tr\left(W^T\left(\sum_{(x_i,x_j)\in S} c_{ij}(x_i-x_j)(x_i-x_j)^T\right)W\right)} \\ &= \arg\max_{W^TW=I} \frac{tr(W^T\cdot FS_D\cdot W)}{tr(W^T\cdot FS_S\cdot W)} \end{aligned} \tag{7}$$

其中,FS_S 表示 S 中样本的带权协方差矩阵,FS_D 表示 D 中样本的带权协方差矩阵,它们的具体形式如下:

$$FS_S = \sum_{(x_i,x_j)\in S} c_{ij}(x_i-x_j)(x_i-x_j)^T \tag{8}$$

$$FS_D = \sum_{(x_i,x_j)\in D} c'_{ij}(x_i-x_j)(x_i-x_j)^T \tag{9}$$

3.2 目标函数的求解

目标函数(10)是一个矩阵迹之比的优化求解问题,且带权协方差矩阵 FS_S 和 FS_D 是半正定的,满足 Xiang 等

人在文献[5]中提出来的算法求解条件,因此可采用 Xiang 等人提出的算法求解该优化问题.其算法针对目标优化问题求解考虑了两种情况.第 1 种情形是 $d'>d-r$(d'是新的向量空间的维度,r 是分母中协方差矩阵的秩),该情形下可确保目标函数的分母不为 0.采用二分搜索方法,通过预先计算得到目标函数的上下界,反复迭代可得到最优解.第 2 种情形是 $d'\leqslant d-r$,该情形下,目标函数的分母可能为 0,此时,算法对样本进行了零空间转换,然后对目标函数的分子求最优.

然而,本文在处理时间序列数据时发现,由于时序数据的高维度特性,协方差矩阵 FS_S 往往是一个低秩矩阵,遭遇目标函数分母为 0 的情形.若直接使用 Xiang 的原始算法,会导致分母上最近邻约束作用失效.因此,本文在 Xiang 的算法基础上,为了避免目标函数分母为 0,在算法中为分母增加了正则化项,即

$$\widehat{FS}_S = FS_S + \lambda I_d \tag{10}$$

其中,$\lambda>0$,I_d 是 d 阶单位矩阵.此时,$\widehat{FS}_S$ 是一个正定矩阵,确保优化过程中相似约束发挥作用.此时,目标函数优化只需考虑 $d'<d-r$ 的情形,具体算法见算法 3.该算法可以有效地避免矩阵求逆问题.在许多度量学习的优化问题中都需要对协方差矩阵求逆,而由于时序数据的高维度特性,矩阵求逆计算代价大;同时,时序数据构成的协方差矩阵往往是一个低秩矩阵,奇异矩阵求逆困难.辅助信息自动生成的时间序列距离度量学习(SIADML)的完整算法见算法 4.

算法 3. 矩阵迹之比问题求解算法.

输入:$FS_D,FS_D\in\mathbb{R}^{d\times d}$,维度参数 d',正则化参数λ,误差常数ε.

输出:矩阵 $W^*\in\mathbb{R}^{d\times d'}$.

1. 令 $S_1\leftarrow FS_D,S_2\leftarrow FS_S+\lambda I_d$.
2. 找到矩阵 S_1 最大的前 d'个特征值$\alpha_1,\ldots,\alpha_{d'}$和矩阵 S_2 最小的前 d'个特征值$\beta_1,\ldots,\beta_{d'}$.
3. $\lambda_1\leftarrow tr(S_1)/tr(S_2),\lambda_2\leftarrow\sum_{i=1}^{d'}\alpha_i\Big/\sum_{i=1}^{d'}\beta_i,\lambda\leftarrow(\lambda_1+\lambda_2)/2$.
4. while $\lambda_1-\lambda_2>\varepsilon$时,执行:
   计算矩阵 $S_1-\lambda S_2$ 的最大的前 d'个特征值的和 $g(\lambda)$;
   若 $g(\lambda)>0$,则更新$\lambda_1\leftarrow\lambda$;否则,更新$\lambda_2\leftarrow\lambda$;
   更新$\lambda\leftarrow(\lambda_1+\lambda_2)/2$.
   Endwhile
5. $W^*=[\mu_1,\ldots,\mu_{d'}]$,其中,$\mu_1,\ldots,\mu_{d'}$是矩阵 $S_1-\lambda S_2$ 最大的前 d'个特征值所对应的特征向量.
6. 结束.

算法 4. 辅助信息自动生成的时间序列距离度量学习(SIADML)算法.

输入:时间序列数据集 $X=\{x_i|i=1,2,\ldots,n\}$.

输出:Mahalanobis 距离矩阵 A^*.

1. 根据算法 1 或算法 2 生成相似约束集 S 和相异约束集 D(算法 1 或算法 2 的输入为:阈值 t);
2. 根据公式(5)和公式(6)计算相似成对样本的距离权重 c_{ij} 和相异成对样本的距离权重 c'_{ij};
3. 根据公式(8)和公式(9)计算 S 中样本的带权协方差矩阵 FS_S 和 D 中样本的带权协方差矩阵 FS_D;
4. 根据算法 3 计算得到线性变换矩阵 W^*(算法 3 的输入为:$FS_D,FS_S,d',\lambda,\varepsilon$);
5. 计算 Mahalanobis 距离矩阵 $A^*=W^*W^{*T}$.
6. 结束.

3.3 算法复杂度分析

SIADML 算法复杂度主要分为辅助信息生成和矩阵迹之比优化问题求解两部分.

设时间序列样本数为 n,序列长度为 d.算法 1 中由于 DTW 计算复杂度为 $O(d^2)$,则构造距离矩阵的计算复杂度为 $O(n^2d^2)$,对每个样本寻找最近邻的复杂度为 $O(n)$,因而构造成对约束集的计算复杂度为 $O(n^2d^2)$.而在改进的算法 2 中,可以避免计算整个 DTW 距离矩阵,由于边界函数的计算复杂度仅为 $O(d)$,当样本量较大时,计算

单个样本最近邻的平均复杂度趋于 $O(nd)$,因而构造成对约束集的计算复杂度可近似为 $O(n^2d)$.

设相似约束集和相异约束集中的样本对数分别为$|S|$和$|D|$,易证$\lfloor(n+1)/2\rfloor\leqslant|S|\leqslant(n-1)$,$\lfloor(n+1)/2\rfloor\leqslant|D|\leqslant(n-1)$,则带权协方差矩阵的计算复杂度为 $O(nd)$.由于对称矩阵特征分解,计算前 d'个特征值对应的特征向量的计算复杂度是 $O(d'd^2)$,矩阵迹之比优化问题的计算复杂度是 $O(ld'd^2)$,其中,l 表示算法 3 中步骤 4 的迭代次数.因此,改进后的算法整体复杂度近似为 $O(n^2d+ld'd^2)$.由于对各样本计算最近邻和最不近邻的过程很容易并行,因而在样本量 n 较大时,可进一步通过多核或者 GPU 等进行并行化处理,进一步提升效率.然而,本文关注的重点是辅助信息自动生成的策略设计,故本文不再对算法效率提升做进一步探讨.

4 实验及其分析

为了验证 SIADML 方法的有效性,本文在一系列时间序列标准数据集上进行了聚类实验.首先将本文 SIADML 方法与无辅助信息即没有经过学习的一般度量方法进行比较;其次,将 SIADML 方法与人为给定辅助信息的度量学习方法进行比较,并分析增加自动生成的辅助信息对度量学习结果的影响.

4.1 SIADML与一般度量方法(未使用辅助信息)的比较

4.1.1 实验设置

在实验中,为了验证 SIADML 的性能,选取了 4 种依据不同度量的聚类算法做对比,参与比较的算法如下:

(1) 基于欧氏距离的 K-means 聚类算法(ED).

(2) 基于 DTW 距离的 K-means 聚类算法(DTW).

(3) 基于高斯核(RBFK)的 K-means 聚类算法,利用核函数将时间序列样本映射到高维特征空间,然后用 K-means 对变换后的时序数据进行聚类.

(4) 基于 TGAK 核[21]的 K-means 算法(TGAK).TGAK 核是利用 DTW 特性构造的一种核函数.

在实验中,采用当前常用的聚类评价指标 Rand Index(RI)来衡量各种聚类算法的性能.它使用聚类结果与样本的真实类标号来估计算法的聚类性能[1,5,6],具体计算公式如下:

$$RI=(N_S+N_D)/N_P \tag{11}$$

其中,N_S 表示属于同一类的样本对在聚类后仍被分到同一簇中的样本对数,N_D 是属于不同类的样本对在聚类后仍被分到不同簇中的样本对数,N_P 是总的样本对数,且 $N_P=n(n-1)/2$.RI 的值越大,表示聚类性能越好.

4.1.2 实验数据集

本文实验中用到的数据集全部来源于 UCR 时间序列数据库[30].该数据库中包含了许多领域的不同数据集,并且已经预先定义了训练集和测试集,方便不同方法之间进行比较.本文共选取了 12 个数据集,对上述几种算法的聚类性能进行测试,这 12 个时间序列数据集的特征描述见表 1.

Table 1 Character description of the time series datasets from UCR

表 1 UCR 时间序列数据集特征描述

数据集	类数	训练集样本数	测试集样本数	时间序列长度
Synthetic Control	6	300	300	60
CBF	3	30	900	128
Trace	4	100	100	275
Two Patterns	4	1 000	4 000	128
Fish	7	175	175	463
Coffee	2	28	28	286
ECGFiveDays	2	23	861	136
FacesUCR	14	200	2 050	131
Symbols	6	25	995	398
ItalyPowerDemand	2	67	1 029	24
Adiac	37	390	391	176
OliveOil	4	30	30	570

4.1.3 实验结果

算法运行在 2.5GHz CPU,8G 内存 Windows 系统的 MATLAB R2010b 环境下.对于每个数据集,在运行算法

之前都按样本进行了 0-1 规范化处理.每个算法均重复实验 100 次,结果取这 100 次实验的平均值.

在公式(10)中引入正则化项来保证协方差矩阵的正定性,因而在比较各种度量在聚类中的有效性之前,本文首先研究正则化参数λ对本文方法聚类效果的影响.在实验中,λ的取值范围是区间$[10^{-10},10^{2}]$,相应的实验如图 1 所示.本文选取其中 6 个数据集做了实验,从实验结果可以看出:当λ的取值范围在区间$[10^{-10},10^{-3}]$时,聚类性能较好且保持平稳;而当λ的取值范围在$[10^{-3},10^{2}]$时,聚类性能下降明显.

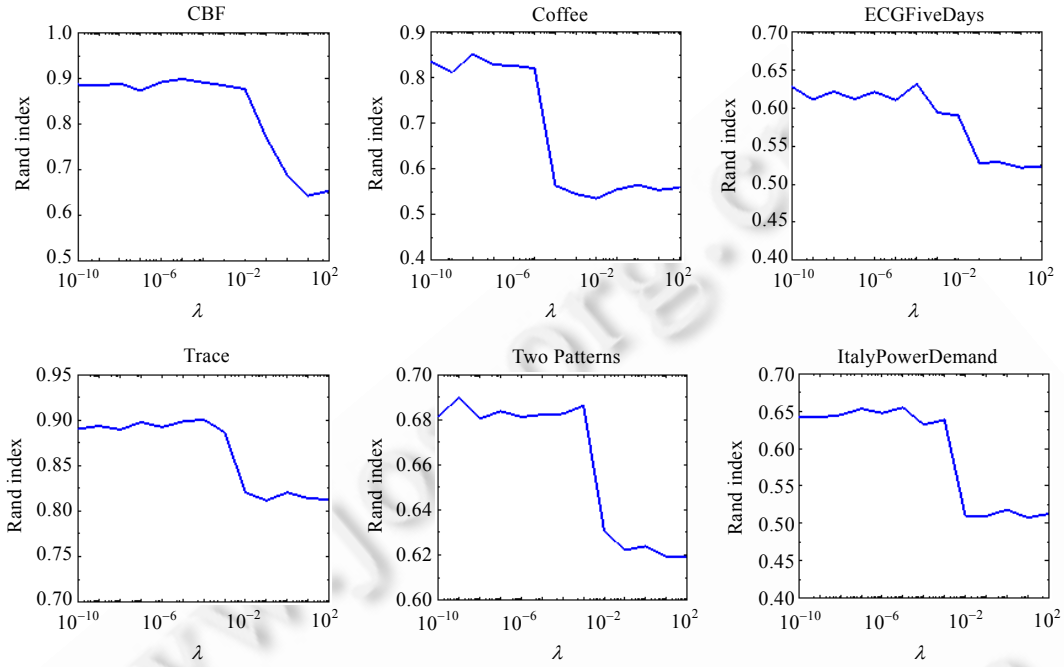


Fig.1 Clustering performance of SIADML by choosing different regularization term λ

图 1 正则化项λ选取不同值时,SIADML 方法的聚类性能

为了比较各种度量的聚类性能,本文在每个数据集的训练数据和测试数据上分别进行了实验.对于本文提出的方法,首先是利用训练数据集学习得到度量,然后将所得度量应用到训练集和测试集上进行聚类.在实验中,设置最近邻阶数 k=1,孤立点阈值为 t=0.8,正则化项为$\lambda=10^{-6}$,而由于时间序列数据维度较高,在度量学习中将维度参数统一设为原数据维度的 1/10,即 $d'=d/10$.对高斯核函数,参数设为σ=0.70711.对 TGAK 核函数,参数设为σ=0.70711,T=0.25.对于聚类簇个数的选择,实验中均按照真实数据集实际的类别数进行取值.

为了验证用 DTW 作为度量寻找辅助信息的有效性,实验中根据样本的真实标签对生成约束信息的正确性进行了统计,对比了利用欧氏距离生成的约束信息,并给出了真实的成对约束总数.结果见表 2,其中,加粗部分表示生成的约束信息质量较高的情况.从表中可以看出,利用 DTW 方法获取的约束信息在大多数的数据集上都要优于传统的欧氏距离,并且约束信息的准确率较高,尤其是相似约束,这有助于学习得到更好的度量.虽然自动生成的辅助信息只占约束信息总数的一小部分,但需要指出的是,实际获取的约束信息质量是有差异的,约束并不是越多越好,目标是用尽可能少的约束得到尽可能好的距离结果.

聚类结果见表 3 和表 4,其中分别给出了各种度量下 K-means 算法的聚类精度.表 3 显示的是各种度量在训练数据集上的聚类结果.由表 3 可以看出,本文提出的 SIADML 方法的聚类精度在大多数数据集上都相对优于其他方法.为了进一步分析本文度量学习方法的泛化性能,将 SIADML 方法在训练集上所学度量用在相应的测试集上进行聚类测试,而其他方法不需要进行学习,即直接在测试数据集上进行聚类.实验结果见表 4,所学度量在测试数据集上的聚类性能依然优于其他方法.

Table 2 Total number/correct number of side information by autogeneration

表 2 自动生成的辅助信息总数及正确的个数

数据集	Euclidean		DTW		依据真实标签的约束总数	
	相似约束 (正确数/总数)	相异约束 (正确数/总数)	相似约束 (正确数/总数)	相异约束 (正确数/总数)	相似约束	相异约束
Synthetic Control	201/260	282/295	**250/254**	**294/294**	7 350	37 500
CBF	18/20	28/28	**26/26**	28/29	139	296
Trace	71/71	98/98	**87/87**	98/98	1 231	3 719
Two Patterns	630/707	904/919	**883/883**	**973/989**	124 837	374 663
Fish	**102/139**	**162/174**	101/146	160/174	2 122	13 113
Coffee	18/18	16/27	**19/19**	**23/27**	182	196
ECGFiveDays	9/16	16/22	**13/17**	16/22	127	126
FacesUCR	105/158	184/199	**129/158**	**192/199**	1 743	18 157
Symbols	14/18	21/22	**16/17**	**24/24**	50	250
ItalyPowerDemand	43/46	37/66	**48/50**	**45/66**	1 089	1 122
Adiac	**171/298**	389/389	162/309	389/389	1 865	73 990
OliveOil	16/21	27/29	**18/20**	27/29	122	313

Table 3 Clustering performance of different metrics on training set

表 3 各种度量在训练集上的聚类性能

数据集	Rand Index±Std. (%)				
	ED	DTW	RBFK	TGAK	SIADML
Synthetic Control	85.45±1.39	85.74±2.13	86.34±0.74	86.60±1.54	**86.99**±0.72
CBF	64.05±3.98	65.90±3.74	64.49±4.53	65.58±2.40	**95.22**±9.48
Trace	82.98±4.34	83.62±2.33	83.73±3.79	83.52±3.50	**90.85**±4.03
Two Patterns	62.87±0.13	62.89±1.12	62.89±0.49	62.90±0.60	**68.33**±1.97
Fish	78.12±1.62	77.50±0.62	**78.84**±1.21	78.56±0.57	77.61±2.17
Coffee	54.76±4.71	56.44±3.77	54.76±0.84	55.75±2.50	**81.69**±11.3
ECGFiveDays	55.26±2.11	59.01±4.22	56.56±3.82	55.34±2.29	**63.08**±4.39
FacesUCR	86.50±0.86	74.38±1.71	**86.65**±0.17	86.43±0.97	86.46±0.43
Symbols	85.10±5.59	84.63±0.05	83.32±2.12	85.83±7.78	**87.78**±1.84
ItalyPowerDemand	49.81±0.10	49.72±0.11	49.93±0.37	49.81±0.74	**73.55**±7.52
Adiac	95.59±0.10	89.86±0.71	95.32±0.07	**95.75**±0.08	94.43±0.37
OliveOil	77.13±4.24	76.64±3.71	76.40±4.80	75.97±4.34	**77.43**±4.13

Table 4 Clustering performance of different metrics on testing set

表 4 各种度量在测试集上的聚类性能

数据集	Rand Index±Std. (%)				
	ED	DTW	RBFK	TGAK	SIADML
Synthetic Control	85.14±1.50	86.14±2.12	86.08±0.83	85.87±1.33	**86.73**±0.81
CBF	67.52±3.77	65.66±2.10	68.05±1.75	69.61±2.12	**86.86**±8.71
Trace	81.72±4.95	81.77±1.87	84.36±2.64	84.85±3.62	**93.27**±5.14
Two Patterns	62.88±0.15	62.85±0.87	63.02±0.87	63.10±0.66	**66.64**±1.69
Fish	76.16±0.86	75.57±0.44	**76.22**±1.71	76.10±0.57	75.66±1.80
Coffee	54.48±4.27	52.42±2.42	55.58±1.23	54.53±1.72	**82.94**±9.40
ECGFiveDays	53.46±2.48	55.52±3.75	53.09±2.76	53.99±1.32	**60.27**±5.22
FacesUCR	**87.22**±0.63	86.40±0.70	86.85±0.58	86.98±1.44	86.95±0.55
Symbols	85.44±6.74	84.82±0.08	83.30±0.87	85.47±7.46	**87.08**±2.04
ItalyPowerDemand	50.30±0.12	49.79±0.20	50.53±0.42	49.66±0.58	**65.40**±6.48
Adiac	95.64±0.14	94.91±0.09	95.04±0.92	**95.76**±0.46	94.75±0.51
OliveOil	76.75±4.00	74.26±2.30	74.88±4.71	73.60±4.83	**77.24**±3.40

根据表 3 和表 4,对聚类结果做进一步的分析:

(i) 首先,SIADML 方法在 9 个数据集上得到了最好的聚类性能,特别是在 CBF,Trace,Coffee 等数据集上,聚类性能得到了非常明显的提升.从表 2 中可以看出:在这些数据集上自动获取的约束信息正确率较高,促进了有效度量的学习;且通过进一步可视化分析发现,这几个数据集的样本在新空间中不同类样本所在簇间隔明显,易于聚类.在剩余的 3 个数据集上,该方法获得的聚类精度也与精度最高的方法相近.这主要归功于 SIADML 方法在无类别信息的时间序列数据上利用 DTW 距离有效构造了一系列成对约束,再用这些成对约束指导度量学习.因此,本文提出的 SIADML 方法是有效的.

(ii) 将 DTW 方法直接用于聚类效果很不好,在各个数据集上均未能提升聚类性能,甚至差于欧氏距离的聚类结果.主要原因是 DTW 不满足三角不等式,相似性不能传递,导致聚类结果很不稳定,且在聚类过程中迭代次数较多,使用 DTW 方法迭代计算簇中心与样本间的距离计算量非常大.因而,利用 DTW 距离的优势,转而指导学习一种满足度量条件且适用于时间序列的距离函数,正是本文方法的重要意义之一.

(iii) 将两种核化聚类方法进行对比,是因为核函数反映的是数据样本在高维特征空间的内积,本质上也是一种度量.实验中采取了两种核函数:一种是常用的高斯核函数,另一种是结合 DTW 距离和高斯核的核函数.从实验结果中可以看出,由于未考虑时间序列数据间的近邻关系,两种核函数并不能为时间序列聚类带来性能的提升.

(iv) 在实验结果中,SIADML 方法在少量训练数据中习得的度量在测试集上依然取得了显著的聚类性能提升.这一方面说明了 SIADML 方法具有较好的泛化能力,另一方面也启发我们在利用 SIADML 方法时,通过对样本数量大的数据集进行抽样,减少 DTW 距离计算的开支,保证该方法在大规模数据集下的实用性.

4.2 SIADML与人为给定辅助信息的度量学习方法比较

在现有的基于成对约束的度量学习方法中,辅助信息依赖人为提供,而本文中采用的辅助信息是利用时间序列特性自动生成的,无须人力参与,可以极大地减少辅助信息的获得代价.本文进一步将 SIADML 方法与其他几种人为提供 must-link 和 cannot-link 的方法在时间序列数据集上进行聚类性能对比.

如图 2 所示,以 ED-Kmeans 为基准,SIADML 方法与 Xing's[1],DCA[4],Xiang's[5],Baghshah's(B's)[6]这 4 种人为给定成对约束的度量学习方法在时间序列数据集上进行了聚类比较.

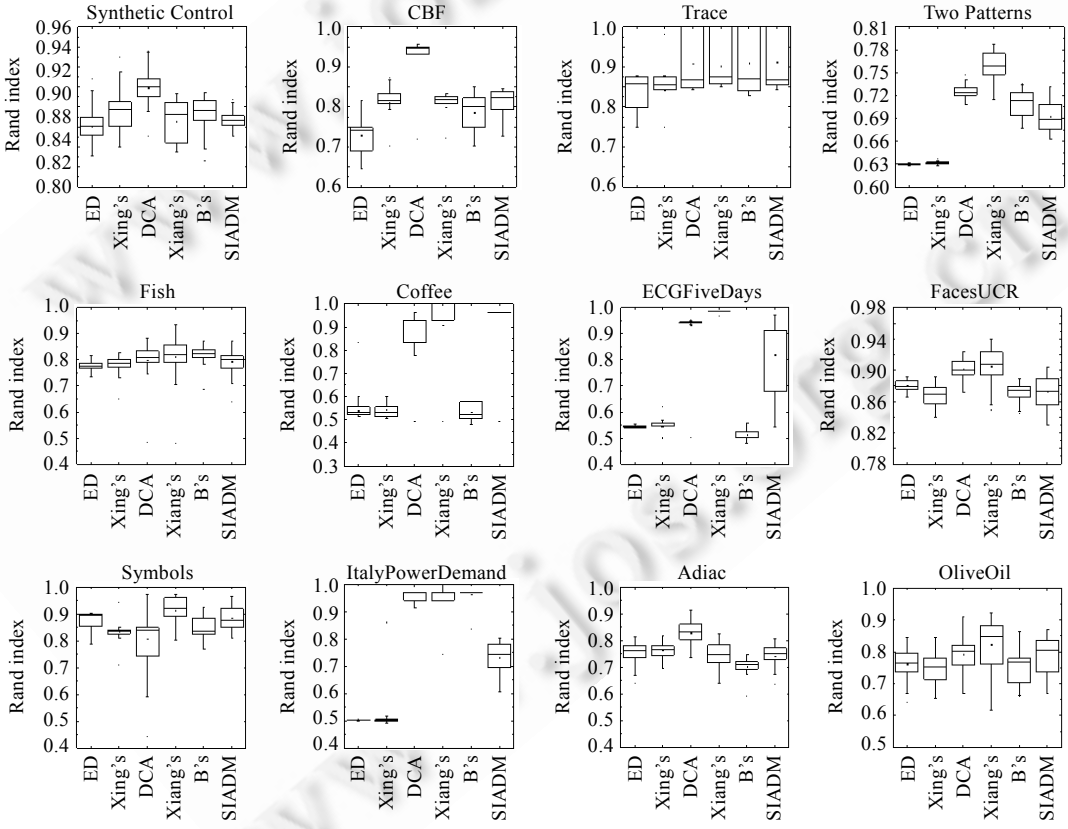


Fig.2 Comparison of the time series clustering: Side information given by people or autogeneration

图 2 时间序列聚类对比:人为给定辅助信息和自动生成辅助信息

如图 2 所示,人为给定的成对约束均是根据样本的真实类别标号随机产生,且约束数量与 SIADML 方法自动生成的辅助信息数量一致.聚类结果中,SIADML 表现出较好的竞争力,在多个数据集上显出与人为给定约束的度量学习接近的性能提升,且明显优于 Xing's 方法. DCA 和 Xiang's 的方法由于全部采用了真实的约束信息,因而聚类性能更好.Baghshah's 方法虽然在 Xiang's 基础上考虑了数据拓扑结构,但 LLE 方法并不总适用于时间序列,如 Coffee,ECGFiveDays 数据集.由此可得,虽然 DTW 距离自身并非严格度量,但在无标签信息的情况下,利用它自动生成辅助信息进行度量学习,也能提升算法性能,并接近于人为提供辅助信息的度量学习结果.而且辅助信息自动生成的阶段相对独立,因而易于嵌入到不同的度量学习算法中.只要自动生成的辅助信息有效,随着具体度量学习算法性能的增强,利用这些辅助信息也能随之获得更好的度量.

在上述实验中,SIADML 方法均是通过寻找一阶最近邻自动获取辅助信息,自然可以联想到通过扩大最近邻阶数 k,获取更多数量的辅助信息.因此,本文又进一步通过增大最近邻阶数 k 的取值,分析所学度量对聚类性能的影响.实验中选取了 CBF,Two Patterns,ECGFiveDays 和 ItalyPowerDemand 这 4 个各类别样本数较大的数据集,实验结果如图 3 所示.在实验中,随着近邻阶数 k 的增大,获得的成对样本对数也随之增加,但聚类性能并未随着辅助信息数量的增加而提高.一个可能的原因是,按照本文方法生成辅助信息时,随着近邻阶数 k 的增大,所获得的成对样本准确性下降,影响了度量学习的效果,因而 k 的取值不宜过大.

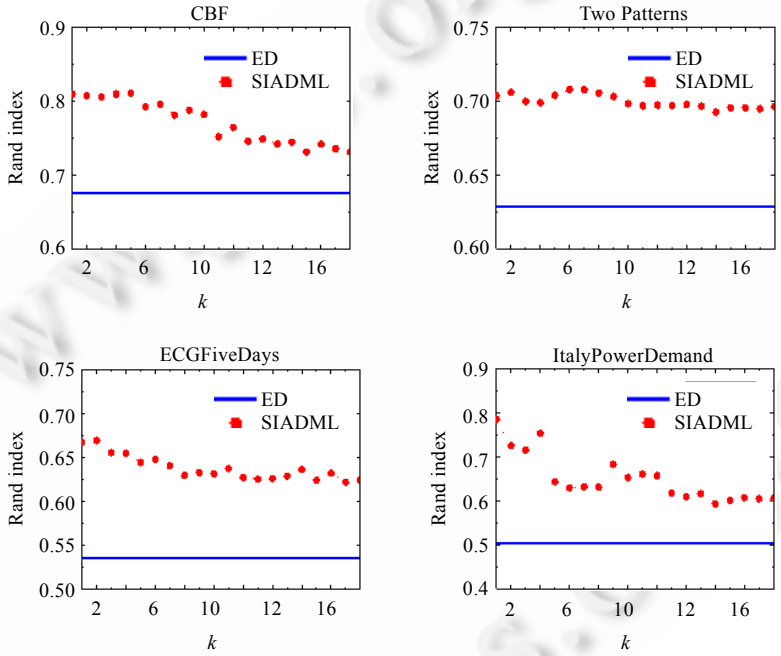


Fig.3 Time series clustering performance by different numbers of side information

图 3 辅助信息数量对时间序列聚类性能的影响

5 总 结

本文将 DTW 方法引入基于成对约束的度量学习算法中,提出了一种辅助信息自动生成的时间序列距离度量学习方法.该方法利用 DTW 构造相似成对约束和相异成对约束,指导学习新的时间序列距离度量.最终,问题模型可以转化成矩阵的迹之比优化问题进行有效求解.该方法能够在无监督环境下自动生成度量学习所需的辅助信息,并且考虑了时间序列样本间的近邻程度.该方法在一系列 UCR 时间序列标准数据集上进行了聚类实验.实验结果表明,该方法能够有效地提高时序聚类性能.

在下一步工作中,为了进一步提高自动生成辅助信息的可靠性,将考虑引入主动学习策略(active

learning)[31],在预生成的辅助信息基础上,对不确定性较高的约束信息进行查询,判断其是 must-link 还是 cannot-link.本文方法将尝试应用到实际的时序聚类任务中,为运动行为分析、面部事件识别等复杂时间序列聚类性能的提高提供可能的度量方案.此外,本文提出的辅助信息自动生成方法也将拓展应用到其他时间序列弱监督学习任务中.

References:

[1] Xing EP, Andrew YN, Jordan M, Russell S. Distance metric learning with application to clustering with side-information. In: Advances in Neural Information Processing Systems 15. The MIT Press, 2002. 505−512.

[2] Shental N, Hertz T, Weinshall D, Pavel M. Adjustment learning and relevant component analysis. In: Proc. of the European Conf. on Computer Vision. Berlin, Heidelberg: Springer-Verlag, 2002. 776−792. [doi: 10.1007/3-540-47979-1_52]

[3] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning a Mahalanobis metric from equivalence constraints. Journal of Machine Learning Research, 2005,6(6):937−965.

[4] Hoi SCH, Liu W, Lyu MR, Ma WY. Learning distance metrics with contextual constraints for image retrieval. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society, 2006. 2072−2078. [doi: 10.1109/CVPR.2006.167]

[5] Xiang S, Nie F, Zhang C. Learning a Mahalanobis distance metric for data clustering and classification. Pattern Recognition, 2008, 41(12):3600−3612. [doi: 10.1016/j.patcog.2008.05.018]

[6] Baghshah MS, Shouraki SB. Semi-Supervised metric learning using pairwise constraints. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence. AAAI Press, 2009. 1217−1222.

[7] Wang QY, Yuen PC, Feng GC. Semi-Supervised metric learning via topology preserving multiple semi-supervised assumptions. Pattern Recognition, 2013,46(9):2576−2587. [doi: 10.1016/j.patcog.2013.02.015]

[8] Zhang Z, Chow TWS, Zhao MB. Trace ratio optimization-based semi-supervised nonlinear dimensionality reduction for marginal manifold visualization. IEEE Trans. on Knowledge and Data Engineering, 2013,25(5):1148−1161. [doi: 10.1109/TKDE.2012.47]

[9] Jun W, Do H, Woznica A, Kalousis A. Metric learning with multiple kernels. In: Advances in Neural Information Processing Systems 24. 2011. 1170−1178.

[10] Jain P, Kulis B, Davis JV, Dhillon IS. Metric and kernel learning using a linear transformation. Journal of Machine Learning Research, 2012,13:519−547.

[11] Guillaumin M, Verbeek J, Schmid C. Is that you? Metric learning approaches for face identification. In: Proc. of the Int'l Conf. on Computer Vision. IEEE Computer Society, 2009. 498−505. [doi: 10.1109/ICCV.2009.5459197]

[12] Yang L, Jin R, Mummert L, Sukthankar R, Goode A, Bin Z, Hoi SC, Satyanarayanan M. A boosting framework for visuality-preserving distance metric learning and its application to medical image retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010,32(1):30−44. [doi: 10.1109/TPAMI.2008.273]

[13] Tran D, Sorokin A. Human activity recognition with metric learning. In: Proc. of the European Conf. on Computer Vision. Berlin, Heidelberg: Springer-Verlag, 2008. 548−561. [doi: 10.1007/978-3-540-88682-2_42]

[14] Prekopcsák Z, Lemire D. Time series classification by class-specific Mahalanobis distance measures. Advances in Data Analysis and Classification, 2012,6(3):185−200. [doi: 10.1007/s11634-012-0110-6]

[15] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. on Acoustics, Speech and Signal Processing, 1978,26(1):43−49. [doi: 10.1109/TASSP.1978.1163055]

[16] Xi X, Keogh E, Shelton C, Wei L, Ratanamahatana CA. Fast time series classification using numerosity reduction. In: Proc. of the Int'l Conf. on Machine Learning. ACM, 2006. 1033−1040. [doi: 10.1145/1143844.1143974]

[17] Rakthanmanon T, Campana BJ, Mueen A, Batista G, Westover B, Qiang Z, Zakaria J, Keogh E. Searching and mining trillions of time series subsequences under dynamic time warping. In: Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2012. 262−270. [doi: 10.1145/2339530.2339576]

[18] Keogh E. Exact indexing of dynamic time warping. In: Proc. of the Int'l Conf. on Very Large Databases. 2002. 406−417. [doi: 10.1007/s10115-004-0154-9]

[19] Shimodaira H, Noma K, Nakai M, Sagayama S. Dynamic time-alignment kernel in support vector machine. In: Advances in Neural Information Processing Systems 15. The MIT Press, 2002. 921−928.

[20] Cuturi M, Vert JP, Birkenes O, Matsui T. A kernel for time series based on global alignments. In: Proc. of the Int'l Conf. on Acoustics, Speech and Signal Processing. IEEE Computer Society, 2007. 413−416. [doi: 10.1109/ICASSP.2007.366260]

[21] Cuturi M. Fast global alignment kernels. In: Proc. of the Int'l Conf. on Machine Learning. ACM, 2011. 929−936.

[22] Zhou F, De la Torre F, Cohn JF. Unsupervised discovery of facial events. In: Proc. of the Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society, 2010. 2574−2581. [doi: 10.1109/CVPR.2010.5539966]

[23] Zhou F, De la Torre F, Hodgins J. Hierarchical aligned cluster analysis for temporal clustering of human motion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,35(3):582−596. [doi: 10.1109/TPAMI.2012.137]

[24] Ratanamahatana CA, Keogh E. Three myths about dynamic time warping data mining. In: Proc. of the SIAM Int'l Conf. on Data Mining. SDM, 2005. 506−510.

[25] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326. [doi: 10.1126/science.290.5500.2323]

[26] Domeniconi C, Gunopulos D. Adaptive nearest neighbor classification using support vector machines. In: Advances in Neural Information Processing Systems 14. The MIT Press, 2001. 665−672.

[27] Feng YC, Jiang T, Li GH, Zhu H. Underlying techniques of efficient similarity search on time series. Chinese Journal of Computers, 2009,32(11):2107−2122 (in Chinese with English abstract). [doi: 10.3724/sp.j.1016.2009.02107]

[28] Kim S, Park S, Chu W. An index-based approach for similarity search supporting time warping in large sequence databases. In: Proc. of the Int'l Conf. of Data Engineering. IEEE Computer Society, 2001. 607−614. [doi: 10.1109/ICDE.2001.914875]

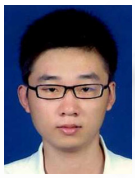
[29] Zhou M, Wong MH. Efficient online subsequence searching in data streams under dynamic time warping distance. In: Proc. of the Int'l Conf. of Data Engineering. IEEE Computer Society, 2008. 686−695. [doi: 10.1109/ICDE.2008.4497477]

[30] Keogh E, Zhu Q, Hu B, Hao Y, Xi X, Wei L, Ratanamahatana CA. The UCR time series classification/clustering homepage. 2011. http://www.cs.ucr.edu/~eamonn/time_series_data/

[31] Settles B. Active learning literature survey. Technical Report, 1648, University of Wisconsin-Madison, 2009.

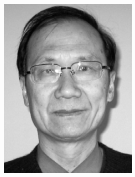
附中文参考文献:

[27] 冯玉才,蒋涛,李国徽,朱虹.高效时序相似搜索技术.计算机学报,2009,32(11):2107−2122. [doi: 10.3724/sp.j.1016.2009.02107]

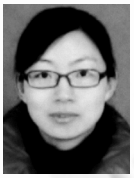


邹朋成(1989－),男,江苏常州人,博士生,CCF 学生会员,主要研究领域为机器学习,数据挖掘.
E-mail: zou-pc@nuaa.edu.cn

张霞(1981－),女,博士生,讲师,主要研究领域为数据挖掘.
E-mail: zhangxia@njcit.cn



王建东(1945－),男,教授,博士生导师,主要研究领域为数据挖掘,机器学习与知识工程,人工智能.
E-mail: aics@nuaa.edu.cn



王丽娜(1979－),女,博士生,讲师,主要研究领域为数据挖掘,人工智能.
E-mail: wangllna@163.com

杨国庆(1949－),男,教授,博士生导师,主要研究领域为模式识别与图像处理,民航信息管理系统.
E-mail: fred9405207@126.com