文章编号: 1001-0920(2018)04-0649-08 DOI: 10.13195/j.kzyjc.2017.0150

基于分量属性近邻传播的多元时间序列数据聚类方法

李海林[1†], 王 成[2], 邓晓懿[1]

(1. 华侨大学 工商管理学院，福建 泉州 362021；2. 华侨大学 计算机科学与技术学院，福建 厦门 361021)

摘 要: 鉴于传统方法不能直接有效地对多元时间序列数据进行聚类分析,提出一种基于分量属性近邻传播的多元时间序列数据聚类方法. 通过动态时间弯曲方法度量多元时间序列数据之间的总体距离,利用近邻传播聚类算法分别对数据之间的总体距离矩阵和分量近似距离矩阵进行聚类分析,综合考虑这两种视角下序列数据之间的关联关系,使用近邻传播方法对反映原始多元时间序列数据的综合关系矩阵实现较高质量的聚类. 数值实验结果表明,与传统聚类方法相比,所提出方法不仅能够有效地反映总体数据特征之间的关系,而且通过重要分量属性序列之间的关联关系分析能够提高原始时间序列数据的聚类效果.

关键词: 多元时间序列；聚类分析；近邻传播；动态时间弯曲；分量属性

中图分类号: TP273 **文献标志码:** A

Multivariate time series clustering based on affinity propagation of component attributes

LI Hai-lin[1†], *WANG Cheng*[2], *DENG Xiao-yi*[1]

(1. College of Business Administration，Huaqiao University，Quanzhou 362021，China；2. College of Computer Sciences and Technology，Huaqiao University，Xiamen 361021，China)

Abstract: In view of the problem that the traditional methods can not be directly effective on such data clustering analysis, a clustering method of multivariate time series data based on component attributes affinity propagation is proposed. The overall distance between multivariate time series data can be measured by dynamic time warping. The clustering analysis of the overall distance matrix and component approximate distance matrix is processed by using the affinity propagation clustering algorithm, considering the relationship between two sequence data from the two kinds of perspectives. The synthetical relationship matrix of the raw multivariate time series data is used for clustering by using the affinity propagation method. The numerical experiment results how that, compared with the traditional clustering methods, the proposed method not only can effectively reflect the relationship of the overall data characteristics, but also improve the clustering effect of the original time series data through the analysis of the relationship between the important component attributes.

Keywords: multivariate time series；clustering analysis；affinity propagation；dynamic time warping；component attribute

0 引 言

时间序列是一组数据按时间先后顺序排列的对象,根据属性维度通常可分为单变量时间序列与多变量(多元)时间序列. 多元时间序列是数据分析领域中较为复杂的数据类型之一,广泛存在于金融、经济、工业工程与电子信息等领域,例如金融股票市场中的交易数据,经济数据分析领域中GDP和CPI指数,工业工程中的机器运行指标和参数等. 多元时间序列不仅具有时间维度长、属性变量多和数据体量大等高维特征,而且还伴有不确定、动态和概念飘移等现象[1-3],这些特征使得传统技术与方法不能直接有效地应用于多元时间序列的数据挖掘与分析. 特别地,大数据时代下产生海量无标签的数据模式使得无监督机器学习方法突显其重要性[4]. 以聚类为代表的无监督机器学习方法,使多元时间序列数据领域中的挖掘与分析更具有挑战性,也是实际应用领域中有

收稿日期: 2017-02-18；**修回日期:** 2017-06-13.

基金项目: 国家自然科学基金项目(71771094, 61300139)；福建省社会科学规划项目(FJ2017B065)；福建省科技计划引导性项目(2017H01010065)；福建省高等学校新世纪优秀人才支持计划项目(Z1625112).

责任编委: 刘民.

作者简介: 李海林(1982−), 男, 副教授, 博士, 从事数据挖掘与决策支持等研究；王成(1984−), 男, 副教授, 博士, 从事数据挖掘与机器学习等研究.

†通讯作者. E-mail: hailin@mail.dlut.edu.cn

待解决的科学问题之一.

时间序列的聚类研究主要集中于单变量时间序列数据对象[5],由于多元时间序列数据的高维性和复杂性使得其研究成果相对较少,一般采用特征表示和距离度量来提升传统聚类算法对多元时间序列数据的聚类[6-7]. Brandmaier[8]根据时间序列片段集合的排列分布情况,构建了适合于排列分布特征的距离矩阵,并使用层次聚类方法进行时间序列聚类(PDC). D'Urso等[9]利用小波变换对多元时间序列数据进行特征表示,并结合传统模糊C均值和K中心等聚类方法,提高了多元时间序列聚类性能. Barragan等[10]也利用小波分析对多元时间序列数据进行特征处理,并使用主成分分析距离测度和传统模糊聚类方法实现了多元时间序列的聚类分析. 于重重等[11]提出了基于主成分分析和传统Gath-Geva聚类算法,对多元时间序列进行分割,从而更有效地实现了多元时间序列的分段问题.

目前,相关的主要研究也存在一些问题: 1) 多元时间序列聚类算法大多建立在划分聚类的基础之上,对于非球形的数据分布形态无法获得较好的结果; 2) 鉴于多元时间序列形态的重要性,聚类过程中的不同时间点数据的异步相关性也是需要考虑的问题; 3) 各个分量属性序列对聚类结果具有不同程度的影响,在聚类分析过程中需要进一步分析不同属性的作用.

针对上述问题,本文运用动态时间弯曲方法对多元时间序列数据进行相似性度量,进而反映序列之间的异步相关性和全局形态特征的近似性,还可以抽取分量属性之间的距离,并结合处理非球形数据分布形态的近邻传播方法实现不同属性对聚类结果的影响. 通过综合考虑全局形态特征和分量属性聚类结果,描述原始多元时间序列之间的相关关系,最后使用近邻传播方法对反映原始时间序列数据之间的综合关系矩阵进行聚类分析. 数值实验结果表明,与传统方法相比,基于分量属性近邻传播的多元时间序列数据聚类方法能够获得较好效果.

1 理论基础

动态时间弯曲(DTW)是一种有效的距离度量方法,它能较好地反映多元时间序列数据的形态相似性[12]. 另外,近邻传播(AP)是一种基于距离矩阵的聚类方法,能够处理非球形分布的数据聚类问题[13].

1.1 动态时间弯曲

在时间序列数据挖掘领域中,大部分挖掘任务的前期过程都需要度量数据之间的相似性或距离,例如时间序列相似性检索、聚类、分类、兴趣模式发现和异常点检测等[14-15]. 在众多距离度量函数中,欧氏距离(Euc)和动态时间弯曲(DTW)是最为常用的两种方法. 欧氏距离虽然能够快速度量两条时间序列之间的相似性,但通常被度量的时间序列要求具有相等的长度,而且其度量效果不能很好地反映形态变化效果. 动态时间弯曲是利用动态规划方法在两条时间序列之间找到最优弯曲路径,进而获得最小的距离以反映序列之间的相似性.

对于多元时间序列$X = [X_1, X_2, \cdots, X_P]$和$Y = [Y_1, Y_2, \cdots, Y_P]$,$x_i(t)$、$y_j(t)$表示多元时间序列$X$、$Y$分别在第$i$个属性和第$j$个属性在$t$时刻的数据值,即每个多元时间序列可以用矩阵$X_{n\times P}$表示.

动态时间弯曲是一种从序列X和Y中寻找一条最优弯曲路径$W = \{w_1, w_2, \cdots, w_K\}$,使得弯曲路径对应的距离累加之和最小,即

$$\mathrm{DTW}(X,Y) = \min_p \sum_{k=1}^{K} \mathrm{dist}(w(k)). \tag{1}$$

其中: 弯曲路径元素$w(k) \equiv (x(t_a), y(t_b))$,表示多元时间序列$X$第$t_a$个时间点数据和$Y$第$t_b$个时间点数据相匹配,且$\mathrm{dist}(w(k)) = ||x(t_a) - y(t_b)||^2 = \sum_{i=1}^{P}(x_i(t_a) - y_j(t_b))^2$. 该弯曲路径通常需要满足边界性、单调性和连续性,使得弯曲路径中的始末元素与多元时间序列的始末相互对应,而且下一个弯曲元素$w(k+1)$仅可能出现在累积代价矩阵R中$w(k)$对应单元格的左上角相邻的3个元素. 同时,累积代价矩阵R可由动态规划计算得到,即

$$R(t_a,t_b) = D(x(t_a),y(t_b)) + \min\begin{cases} R(t_a,t_b-1), \\ R(t_a-1,t_b-1), \\ R(t_a-1,t_b). \end{cases} \tag{2}$$

其中: $t_a = 1,2,\cdots,n$, $t_b = 1,2,\cdots,m$; D为X和Y不同时间数据点距离矩阵,有$D(x(t_a),y(t_b)) = \sum_{i=1}^{p}(x_i(t_a) - y_j(t_b))^2$,且初始化$R(0,1:m) = \infty$, $R(1:n,0) = \infty$. 由动态规划计算公式(2)可知,最终得到DTW度量两条多元时间序列X与Y的最小距离为$R(n,m)$,即有$\mathrm{DTW}(X,Y) = R(n,m)$.

由式(2)易知,动态时间弯曲的时间复杂度为$O(mn)$,与时间序列长度成平方阶关系,这不利于大量较长时间序列的相似性比较. 为了解决此类问题,不少学者对提高DTW的计算效率提出了相应的策略和方法. Salvador等[16]使用多分辨率方法,将从粗

到细重复投影过程用来近似计算DTW,使得时间复杂度接近于线性时间序列的长度. Li[17] 针对较长时间序列数据提出利用DTW对较短分段序列进行最优弯曲路径,并通过组合重叠区域的弯曲路径获得最终的近似弯曲路径,不仅能够降维DTW对原始较长时间序列数据的计算复杂度,还能应用于在线时间序列数据的距离计算.

1.2　近邻传播

近邻传播(AP)是由Frey等[13]提出的一种效果较好的聚类方法. 其以数据对象之间的相似性矩阵为基础,初始时刻把每个对象当作代表点,对每个数据的两种信息进行交互与传播,即代表度(Representative)和有效性(Availability),并且逐步迭代至收敛,使得具有簇代表能力的数据有着较高的代表度和有效性.

数据对象之间的相似性矩阵构建通常可利用欧氏距离的相反数来表示,即 $S(i,j) = -(O_i,-O_j)^2$,其中 O_i 和 O_j 表示两个数据对象. 另外,AP算法的聚类簇数目容易受偏向参数Pr(Preference)的影响,该参数的大小决定了最终聚类簇数的多少. 通常情况下,偏向参数取相似性矩阵的中位数,即 $\mathrm{Pr} = \mathrm{median}(S)$,使得最终的聚类结果具有较合适的簇数目.

AP聚类算法中,两种竞争信息通过逐步迭代来传播相关信息,使得数据间能够产生较高代表度和有效性的代表中心点. 代表度 $r(i,j)$ 表示数据对象 O_i 传到候选数据对象 O_j 的信息,反映了 O_j 对 O_i 的代表程度. 有效性 $a(i,j)$ 是指代表对象 O_j 传递数据对象 O_i 的信息,反映了代表对象 O_j 作为数据对象 O_i 的有效程度. AP聚类算法迭代过程中,每次计算每个数据对象 O_i 选择具有较高代表度和有效性的数据对象 O_j 作为其代表对象,即

$$O_j^{(t)} = \arg\max_j\{a^{(t)}(i,j) + r^{(t)}(i,j)\}. \tag{3}$$

$O_j^{(t)}$ 表示第 t 次迭代过程中,O_i 选择数据对象 O_j 作为代表对象. 每次迭代过程中,同一数据对象会选择不同的其他对象作为代表点,直到AP聚类算法迭代收敛为止. AP算法能够自适应地对数据进行聚类,同时兼顾了传统划分聚类和基于密度的聚类方法的优点,能够较好地处理非球形分布的数据聚类问题[18-20].

2　多元时间序列聚类

本文提出一种基于分量属性近邻传播的多元时间序列数据聚类方法(Component attributes based affinity propagation clustering for multivariate time series data, cACM),使用动态时间弯曲对多元时间序列进行距离度量,并从整体序列信息和局部分量属性的角度分析多元时间序列的聚类情况,同时利用两者的聚类信息反映原始多元时间序列之间的相关关系,再通过近邻传播实现多元时间序列数据的聚类分析.

2.1　相关性矩阵

相关性矩阵用于反映不同多元时间序列数据对象隶属于同一个聚类簇的情况,其构建的主要思路是利用动态时间弯曲方法对多元时间序列数据对象构建相似性矩阵 A,通过最优弯曲路径来解释每个分量属性 p 对应的相似性矩阵 B_p,再使用AP聚类对每个相似性矩阵进行聚类分析,进而获得在不同属性视角下的相关性矩阵 R_A 和 R_{B_p}.

考虑任意两条多元时间序列 $X_{n\times P}$ 和 $Y_{m\times P}$,利用DTW对其进行相似性度量,即 $\mathrm{D}(X,Y) = -\mathrm{DTW}(X,Y)$,并获得最优弯曲路径 $W=\{w(1),w(2),\cdots,w(K)\}$. 通过最优弯曲路径 W 可以解析不同多元时间序列数据在同一分量属性描述下的相似性,即第 i 个分量属性序列之间的相似性为

$$D_{\mathrm{DTW}}(x_i,y_i) = -\sum_{k=1}^{K}\mathrm{dist}(w_i(k)). \tag{4}$$

其中: $w_i(k)$ 代表最优弯曲路径 W 中第 k 个元素对应第 i 个分量的值,即 $w_i(k) = (x_i(t_a), y_i(t_b))$ 且 $\mathrm{dist}(w_i(k)) = (x_i(t_a) - y_i(t_b))^2$.

对于含有 P 个分量属性的多元时间序列数据集 $O = \{O_1, O_2, \cdots, O_N\}$,通过DTW计算每对多元时间序列的相似性,不仅可以获得反映多元时间序列数据整体信息的相似性矩阵 A,而且还能解析出反映分量属性序列局部关系的相似性矩阵 $B = \{B_1, B_2, \cdots, B_P\}$,即 $A = \{a_{ij}\}_{N\times N}$ 和 $B_p = \{b_{ij}^p\}_{N\times N}$. 其中: $a_{ij} = \mathrm{DTW}(O_i, O_j), b_{ij}^{(p)} = D_{\mathrm{DTW}}(O_i^{(p)}, O_j^{(p)})$.

根据多元时间序列数据相似性矩阵 A 和分量属性序列相似性矩阵集合 B,利用近邻传播AP方法对这些相似性矩阵进行聚类分类,即 $C_A = AP(A)$ 和 $C_{B_p} = AP(B_p)$,进而获得各种维度视角下的多元时间序列数据聚类标签 C_A 和 C_{B_p}. 将聚类结果中同簇内数据对象关系视为1,不同簇中数据对象视为0,将 C_A 和 C_{B_p} 分别转化为反映多元时间序列数据对象在不同视角下的相关性矩阵,即 $R_A = \mathrm{C2R}(C_A)$ 和 $R_{B_p} = \mathrm{C2R}(C_{B_p})$,其中C2R是将聚类结果转化为关系矩阵的函数.

2.2　相关性近邻传播聚类

关系矩阵用来反映数据对象之间隶属于同簇的程度,其元素值可以看成是两个数据对象共同出现在

同一簇的次数,即共现程度. 特别地,综合关系矩阵是在分析各个分量属性序列聚类结果的基础上构建的,通过对基于DTW距离度量的数据聚类分析,从整体信息和分量属性的角度综合研究对应数据对象在不同聚类结果中的共现程度,其反映了原始数据对象之间的近似关系.

利用相同分量属性序列数据集的聚类结果可以得到相应的关系矩阵,例如第p个分量属性序列构成的数据集所对应的关系矩阵为B_p. 通常情况下,两条相似且在同一簇中的多元时间序列数据,它们会有近似的分量属性序列子集. 从形式上看,在数据集O中,如果任意O_i与O_j为同一个簇中的成员,则存在重要分量属性序列子集$F_{P'} \in F_P$且P'为聚类结果中最大成员数,使得$\mathrm{DTW}(O_i^{P'}, O_j^{P'}) < \varepsilon$,其中$\varepsilon$为较小的距离. 换言之,在$F_{P'}$重要分量属性序列集合中,同一簇中的多元时间序列具有较高的相似度,因此,需要进一步解析哪些分量属性子集能够构造较好的聚类结果.

为了同时考虑整体和分量的信息,将关系矩阵R_A与关系矩阵集合R_B合并成新的关系矩阵集合Q,即$Q = \{R_A, R_{B_1}, R_{B_2}, \cdots, R_{B_P}\}$. 将关系矩阵集合$Q$中的每个关系矩阵视为一个数据对象,使用AP聚类方法对这些关系矩阵对象进行聚类分析,找出具有最大成员对象的簇作为重要分量属性序列子集$F_{P'}$,即

$$F_{P'} = \arg\max_{P'} \mathrm{Tabulate}(\mathrm{AP}(S)), \tag{5}$$

其中Tabulate用来统计AP聚类分类结果中各簇中成员的数目.

由于关系矩阵为稀疏性矩阵,采用夹角公式$\mathrm{Cosim}(\cdot)$计算每对关系矩阵(Q_i和Q_j)的相似性,即$s_{ij} = \mathrm{Cosim}(Q_i, Q_j)$,其中$i, j = 1, 2, \cdots, P+1$. 关系矩阵集合$Q$中不同关系矩阵之间所构建的相似性矩阵为$S$,其能使AP聚类更好地体现数据集中分量属性之间的关系.

式(5)找到具有最大簇数的成员,这些成员反映了关系矩阵集合Q中$F_{P'}$分量矩阵具有较强的相似性,从而得到体现整体和分量信息的关系矩阵G',即

$$G' = \sum_{p'}^{P'} Q_{p'},\ F_{p'} \in F_{P'}. \tag{6}$$

由于对Q进行聚类后,具有最大成员数目的簇可能不包括反映整体信息的关系矩阵R_A,使得G'不能反映多元时间序列数据的整体相关性. 为此,关系矩阵G'不仅需要考虑多元时间序列数据的整体相关性问题,而且需要着重强调整体相关性的作用,即在整体相关性的前提下,考虑分量相关性的影响. 故有修正后的综合关系矩阵

$$G = (G' + R_A) \cdot G' = \left(\sum_{p'}^{P'} Q_{p'} + R_A\right) \cdot \sum_{p'}^{P'} Q_{p'},\ F_{p'} \in F_{P'}, \tag{7}$$

其中$\cdot$表示点乘运算.

关系矩阵G综合了多元时间序列数据之间整体和分量信息的相关性问题,也反映了原始多元时间序列之间的相似性问题,可作为多元时间序列数据的相似性矩阵,故利用基于相关性的近似传播AP方法实现原始多元时间序列数据的聚类,即$C = \mathrm{AP}(G)$. 根据多元时间序列数据中分量属性分析结果的重要性,结合DTW和AP聚类方法的优点,将cACM算法流程归纳为图1.

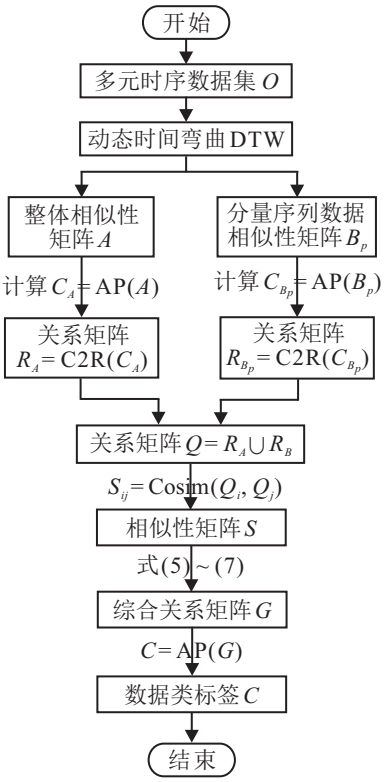


图1 cACM算法流程

首先,针对多元时间序列数据集$O = \{O_1, O_2, \cdots, O_N\}$,利用DTW对每对多元时间序列进行距离度量,不仅可以从整体形态角度获得反映多元时间序列数据之间的相似性矩阵A,还可通过最优弯曲路径解析出每个分量属性p对应数据之间的分量属性序列相似性矩阵$B_p, p = 1, 2, \cdots, P$,进而构建整体信息R_A和分量信息的关系矩阵集合$R_B = \{R_{B_1}, R_{B_2}, \cdots, R_{B_P}\}$. 然后,将整体信息的关系矩阵$R_A$和分量信息的关系矩阵集合$R_B$合并为新数据特征对象集合$Q$,使用夹角公式Cosim度量这些关系矩

阵对象的相似性,即 $S = \mathrm{Cosim}(Q_i, Q_j)$,再利用AP聚类方法对相似性矩阵$S$进行聚类分析,获得最大簇成员对象$F_{P'}$. 最后,根据式(5)～(7)分析最大簇成员对象和整体信息的关系矩阵以构建综合关系矩阵G,通过聚类分析$C = \mathrm{AP}(G)$得到多元时间序列数据集每个对象的类标签.

在cACM聚类算法中,近邻传播聚类AP能够使多元时间序列数据集根据相似性矩阵或关系矩阵得到较好的聚类结果,但其过程执行了3组近邻传播AP算法,使得cACM需要较高的时间代价. 利用DTW计算整体多元时间序列相似性矩阵的时间复杂度为$O(NPm^2)$. 其中: N为数据集中多元时间序列数据的个数,P为多元时间序列分量属性个数(属性维度),为了便于分析,假设每条时间序列的长度为m. 第1组针对相似性矩阵A和每个分量相似性矩阵B_p进行AP聚类,需要时间复杂度为$O(t_1N^3(P+1))$;第2组对维度为$(P+1)\times(P+1)$的相似性矩阵S进行AP聚类,其时间复杂度为$O(t_2P^4)$;第3组对综合关系矩阵G进行AP聚类分析,其时间复杂度为$O(t_3N^3)$. 因此,完成cACM所需要的时间复杂度为$O(NPm^2+(t_1(P+1)+t_3)N^3+t_2P^4)$,其中$t_1$、$t_2$和$t_3$分别为3组AP聚类运算过程的迭代次数. 传统AP对整体多元时间序列数据进行聚类分析记为OPA,其时间复杂度$O(NPm^2+t_0N^3)$,t_0为AP聚类迭代次数. 可见cACM需要较多的计算时间.

需要说明的是,cACM方法主要使用近邻传播AP来实现对相关性矩阵数据信息的聚类,使得该算法性能依赖于AP聚类方法的收敛性. 然而,AP聚类的收敛性可以通过调整阻尼因子λ得到提升,通常情况下也具有较好的收敛性.

3　数值实验与分析

本实验分成两个部分来验证所提出方法的有效性和聚类效果: 一方面,通过仿真实验进一步详述本文方法的聚类过程; 另一方面,通过使用具体的多元时间序列数据集,对比传统方法与本文方法的聚类效果.

3.1　仿真实验

在金融与工业工程等领域中,金融股市的交易价格、成交量和各种相关交易指数以及工业工程行业中机床参数和发动机运行指数等都是常见的时间序列数据. 特别地,在工业工程领域中,Robot execution failure是对Robot进行故障监控的典型代表数据集,该数据集中存在6个属性维度,即3个变量为受力,另外3个变量表示扭矩. 随机选取其中10条多元时间序列数据进行实例演算与分析,每条多元时间序列包含6个分量属性$\{P_1, P_2, \cdots, P_6\}$,时间序列长度为15. 另外,这10条多元时间序列包含4类数据信息,其真实类标签表示为$C_T = [2,1,1,2,2,3,3,4,1,4]$,即10条时间序列数据中第$\{1,4,5\}$、$\{2,3,9\}$、$\{6,7\}$和$\{8,10\}$各为一类.

根据算法流程信息可知: 使用DTW对10条多元时间序列数据进行相似性度量,可以获得整个相似性矩阵A以及分量相似性矩阵集合$B = \{B_1, B_2, \cdots, B_6\}$;利用AP方法对$A$和$B_p$进行聚类分析,将每个相似性矩阵转化为关系矩阵$R_A$和$R_{B_p}$. 每个关系矩阵形如$R_A$所示,即

$$R_A = \begin{bmatrix} 1&0&0&1&1&0&0&0&0&0\\ 0&1&1&0&0&1&1&0&1&0\\ 0&1&1&0&0&1&1&0&1&0\\ 1&0&0&1&1&0&0&0&0&0\\ 1&0&0&1&1&0&0&0&0&0\\ 0&1&1&0&0&1&1&0&1&0\\ 0&1&1&0&0&1&1&0&1&0\\ 0&0&0&0&0&0&0&1&0&1\\ 0&1&1&0&0&1&1&0&1&0\\ 0&0&0&0&0&0&0&1&0&1 \end{bmatrix}.$$

其中: 元素值1表示对应行与列所代表的多元时间序列对象数据被AP方法聚合成同一类,反之说明它们没在同一类中.

通过合并R_A与R_B得到关系矩阵信息集合Q,使用Cosim函数对Q中所有关系矩阵进行相似性度量,获得7个关系矩阵之间的相似性矩阵S,即

$$S = \begin{bmatrix} 1.00&0.90&0.92&0.73&0.84&0.97&0.92\\ 0.91&1.00&0.90&0.83&0.86&0.89&0.90\\ 0.92&0.90&1.00&0.91&0.87&0.96&1.00\\ 0.73&0.82&0.91&1.00&0.84&0.82&0.91\\ 0.84&0.86&0.87&0.84&1.00&0.85&0.87\\ 0.97&0.89&0.96&0.82&0.85&1.00&0.96\\ 0.92&0.90&1.00&0.91&0.87&0.96&1.00 \end{bmatrix}.$$

使用AP方法对相似性矩阵S进行聚类,即$C_S = \mathrm{AP}(S)$,得到$C_S = [6,2,3,3,3,6,3]$,表示在关系矩阵集合Q中,第1个和第6个关系矩阵聚成一类,该类的中心代表点为第6个关系矩阵;第3、4、5和7个关系矩阵聚成一类,其中心代表对象为第3个关系矩阵;第2个关系矩阵自身为一类. 因此,先取最大簇的成员作为重要分量属性序列子集$F_{P'}$,即$P' = [3,4,5,7]$,进一步可以得到体现整体和分量信息的关系矩阵$G = Q_3+Q_4+Q_5+Q_7$,其为重要属性序列子

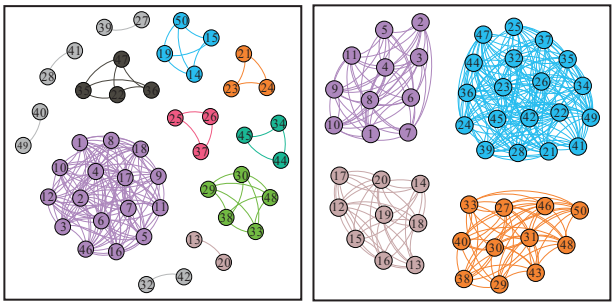
集 $F_{P'}$ 中对应 P' 个关系矩阵的累加之和. 根据式(7)并结合整体关系矩阵 R_A 可得到修正后的综合关系矩阵 G,即 $G' \Rightarrow G$ 为

$$\begin{bmatrix} 4&0&0&2&3&0&0&1&0&0\\ 0&4&3&0&0&2&2&0&3&0\\ 0&3&4&0&0&3&3&1&2&0\\ 2&0&0&4&2&0&0&0&0&2\\ 3&0&0&2&4&1&1&0&1&0\\ 0&2&3&0&1&4&4&1&3&0\\ 0&2&3&0&1&4&4&1&3&0\\ 1&0&1&0&0&1&1&4&0&2\\ 0&3&2&0&1&3&3&0&4&0\\ 0&0&0&2&0&0&0&2&0&4 \end{bmatrix} \Rightarrow \begin{bmatrix} 5&0&0&3&4&0&0&0&0&0\\ 0&5&4&0&0&3&3&0&4&0\\ 0&4&5&0&0&4&4&0&3&0\\ 3&0&0&5&3&0&0&0&0&0\\ 4&0&0&3&5&0&0&0&0&0\\ 0&3&4&0&0&5&5&0&4&0\\ 0&3&4&0&0&5&5&0&4&0\\ 0&0&0&0&0&0&0&5&0&3\\ 0&4&3&0&0&4&4&0&5&0\\ 0&0&0&0&0&0&0&3&0&5 \end{bmatrix}.$$

最后,通过AP方法对综合相关矩阵 G 进行聚类,其结果能够较好地反映原始多元时间序列之间的聚类情况,即 $C_G = [5,6,6,5,5,6,6,8,6,8]$. 聚类结果表明: $\{1,4,5\}$ 为一类,代表对象为第5个多元时间序列; $\{8,10\}$ 为一类,代表对象为第8个多元时间序列;剩下的为一类,代表对象为第6个多元时间序列数据. 10个数据被分为3类,并且各簇代表对象分别为第5、第8和第6个多元时间序列数据.

通过分析易知,由于cACM方法聚类得到的第3个簇中心代表点为第6个多元时间序列,使得该方法的聚类正确率为70%. 若不考虑中心代表点的类标签问题,则可认为cACM方法聚类正确率为80%,仅有第6和第7个数据对象分类错误.

同样,使用cACM方法对Robot execution failure中第1个包含50个多元时间序列数据对象的数据集进行聚类,可以得到若干个簇. 如图2(a)所示,近邻传播AP方法将其分成11个簇. 与真实分类相比,如图2(b)所示,本文方法也能够自适应地将多元时间序列数据划分成关系较近的小簇.



(a) cACM方法聚类情况 (b) 数据真实分类情况

图 2 50条多元时间序列数据的聚类情况比较

3.2 聚类实验

实验数据采用包含工业工程、医疗卫生、手势行为和语言发音等各行业的多元时间序列数据集,包括Australian language(ASL)、Arabic digits(AD)、CMU subject 16(CMUS16)和Japanese vowels(JV)等10个数据集,具体数据信息见表1.

表 1 多元时间序列数据集

序号	名称	变量数	长度	类别数	数据量
1	ASL	22	[45～136]	95	1 425
2	AD	13	[4～93]	10	2 200
3	CMUS16	62	[127～580]	2	29
4	ECG	2	[39～152]	2	100
5	JV	12	[7～29]	9	370
6	LP1	6	15	4	50
7	LP2	6	15	5	30
8	LP3	6	15	4	30
9	LP4	6	15	3	75
10	LP5	6	15	5	100

在表1中,第1～第5个数据集中多元时间序列数据的长度不相等,例如ASL数据集中,多元时间序列数据的时间维度从45到136不等. 然而,第6～第10个数据集中的多元时间序列长度相等,并且这5个数据集属于同一行业的数据,但它们之间具有不同的类别数和数据量.

为了说明本文方法cACM的聚类效果,使用基于整体原始数据信息的近邻传播聚类(OAP)、基于主成分分析的近邻传播聚类(PCA_AP)[13]、基于相似性矩阵的 K-means 聚类[10] 和经典多元时间序列聚类(PDC)[8] 等方法进行实验比较. OAP是通过对某个多元时间序列数据集中的对象进行距离计算,再通过传统AP聚类方法对获得的距离矩阵进聚类分析;PCA_AP是事先对每个多元时间序列数据进行主成分分析及特征变换,分别提取前 $p = \{1,2,\cdots,P/2\}$ 个主成分,并使用近邻传播AP进行聚类分析,其中 P 表示数据集中多元时间序列数据的属性维度; K-means 聚类方法则是建立在整体多元时间序列之间距离的基础之上,使用传统的 K 均值方法对数据进行聚类.

另外,基于近邻传播AP的聚类效果通常受两个参数的影响,即偏向参数Pr和阻尼因子 λ 的影响. 前者主要决定聚类簇的数目,后者主要起算法收敛作用. 鉴于文献[13]已给出了对偏向参数和阻尼因子的设置方法,取相似性矩阵的中位数作为Pr参数的取值和文献[13]算法程序对 λ 设置的默认值0.9作为该参数的值,因此,聚类实验中所有关于AP聚类程序的阻尼因子 λ 取值均为0.9.

使用DTW度量10个数据集中多元时间序列数据或特征序列之间的相似性(或距离),使得各聚类方法具有较好的聚类效果. 另外一组聚类比较实验,对于等长时间序列数据,采用欧氏距离(Euc)进行相似

性或距离度量, 说明本文方法使用 DTW 作为相似性度量的必要性. 另外, 由于 cACM、OAP 和 PCA_AP 都是基于近邻传播的方法, 聚类结果 $C = \{c_1, c_2, \cdots, c_N\}$ 是把同一个簇中成员的类标签标记为该簇的代表对象, 聚类实验的正确率 Pre 定义为

$$\mathrm{Pre} = \sum_{i=1}^{N} \frac{\Phi(\mathrm{Label}(c_i) - \mathrm{Label}(i))}{N}. \tag{8}$$

其中

$$\Phi(x) = \begin{cases} 1, & x = 0; \\ 0, & x \neq 0. \end{cases}$$

$\mathrm{Label}(i)$ 表示第 i 个多元时间序列数据的真实类标签, 而 $\mathrm{Label}(c_i)$ 表示第 i 个多元时间序列的预测标签, c_i 表示第 i 个多元时间对序列数据指向代表对象所对应的序号. 通过实验比较, 各种方法在不同数据集下的聚类结果如表 2 所示, 其中 PCA_AP 以 $P/2$ 次实验中最优的结果作为该方法的聚类结果.

表 2　不同方法在不同数据集中的聚类结果

方法	数据集					
	1	2	3	4	5	6
cACM	0.76	0.99	0.79	0.73	0.95	0.84
OAP	0.51	0.98	0.93	0.81	0.92	0.74
PCA_AP	0.10	0.86	0.72	0.83	0.16	0.64
K-means	0.31	0.48	0.59	0.67	0.67	0.52
PDC	0.08	0.10	0.59	0.68	0.31	0.50

方法	数据集					
	7	8	9	10	mean	sd
cACM	0.73	0.83	0.88	0.64	**0.81**	**0.11**
OAP	0.73	0.67	0.88	0.57	0.77	0.16
PCA_AP	0.70	0.67	0.76	0.66	0.61	0.26
K-means	0.67	0.57	0.77	0.41	0.56	0.14
PDC	0.60	0.67	0.68	0.38	0.46	0.23

通过表 2 聚类结果以及图 3 的比较可知, 本文方法 cACM 获得了较好的聚类结果, 而且从适应各种数据聚类的稳定性来看, 本文方法也具有最小的标准差.

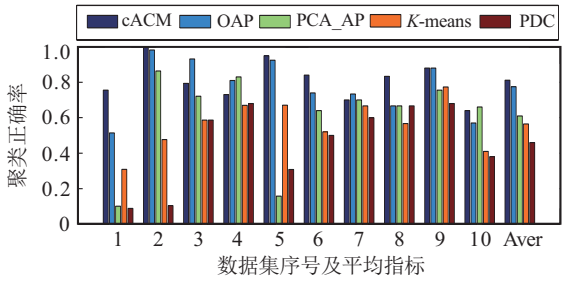


图 3　本文方法 cACM 与其他聚类方法的可视化结果比较

从图 3 的聚类结果比较易知: 本文方法 cACM 对于大部分数据集而言都能表现出较高的聚类正确率; 与其他方法相比, 从聚类效果来看, cACM 在不同数据集中具有较高的胜出率和较强的适用性.

DTW 和 Euc 是常用于度量时间序列相似性的方法, 通常 DTW 具有较好的度量质量, 但需要时间序列长度平方阶的时间复杂度. 为了进一步说明本文方法 cACM, 选择 DTW 作为多元时间序列相似性度量方法的依据, 比较基于这两种方法的等长时间序列数据聚类结果, 如图 4 和图 5 所示.

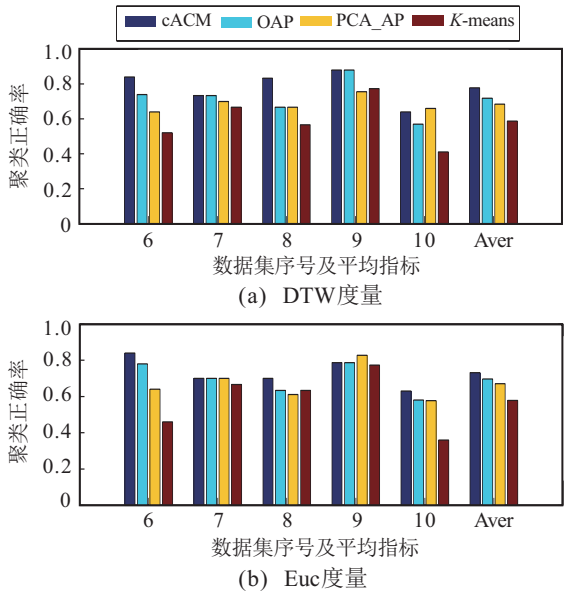


图 4　DTW 和 Euc 度量方法对等长多元时间序列的聚类结果比较

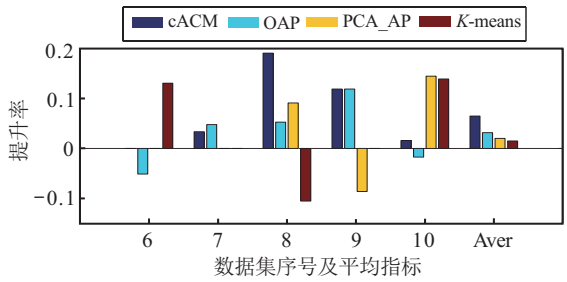


图 5　4 种方法使用 DTW 度量方法的聚类结果提升率

由于 PDC 是需要构建适合于排列分布特征的距离矩阵后进行聚类的方法, 其不参与不同距离度量函数的比较实验. 从基于 DTW 的聚类结果图 4(a) 可知, 与其他方法相比, cACM 方法能够获得稳定且较好的聚类效果. 然而, 从基于 Euc 的聚类结果图 4(b) 可知, cACM 方法无法始终获得最好的聚类结果. 但从平均指标来看, 基于 DTW 和 Euc 的 cACM 方法具有较好的聚类结果. 从聚类结果提升率来看 (见图 5), 与基于 Euc 的多元时间序列聚类方法相比, 基于 DTW 的聚类方法通常可以提升聚类算法的聚类质量.

4　结　论

鉴于传统方法不能很好地直接对多元时间序列数据进行聚类分析, 本文提出了一种基于分量属性近邻传播聚类的多元时间序列数据聚类方法 (cACM). 该方法分别从整体信息和分量属性的角度

出发,使用DTW计算整体信息的相似性且解析出分量属性视角下各数据对象之间的关系矩阵,以近邻传播AP方法为主要聚类手段,通过建立综合关系矩阵来反映两种视角下多元时间序列数据之间的关系,再通过AP方法进行数据聚类分析. 从数据实验结果来看,本文方法能够较好地应用于医疗、语音、文字和故障监控等多元时间序列数据模式识别与知识发现.

本文方法的主要优点是: 1) 考虑整体多元时间序列数据信息的同时,兼顾了分量属性反映多元时间序列之间相关关系对聚类结果的影响; 2) 利用DTW方法一次性计算整体多元时间序列数据之间的相似性,根据最优弯曲路径解析出分量属性序列之间的相似性,提升了算法的运行效率; 3) 使用近邻传播AP方法可以自适应地对反映数据对象之间关系的邻近性矩阵进行聚类分析,每个簇中都有代表性较高的代表对象.

然而,从算法时间复杂度分析可知,本文方法使用了多次AP聚类过程,其时间性能取决于AP聚类的收敛性,通常需要较高的计算时间代价. 为此,研究如何提升AP聚类算法效率或通过分量属性之间的相关性直接导出各分量属性视角下的聚类结果是需要进一步研究的主要工作.

参考文献(References)

[1] 韩萌, 王志海, 原继东. 一种基于时间衰减模型的数据流闭合模式挖掘方法[J]. 计算机学报, 2015, 7: 1473-1483.
(Han M, Wang Z H, Yuan J D. Efficient method for mining closed frequent patterns from data streams based on time decay model[J]. Chinese J of Computers, 2015, 7: 1473-1483.)

[2] Moshtaghi M, Leckie C, Bezdek J C. Online clustering of multivariate time-series[C]. Proc of the 2016 SIAM Int Conf on Data Mining. Miami, 2016: 360-368.

[3] 陈海燕,刘晨晖,孙博. 时间序列数据挖掘的相似性度量综述[J]. 控制与决策, 2017, 32(1): 1-11.
(Chen H Y, Liu C H, Sun B. Survery on similarity measurement of time series data mining[J]. Control and Decision, 2017, 32(1): 1-11.)

[4] 周开乐,杨善林,丁帅, 等. 聚类有效性研究综述[J]. 系统工程理论与实践, 2014, 34(9): 2417-2431.
(Zhou K L, Yang S L, Ding S, et al. On cluster validation[J]. Systems Engineering — Theory & Practice, 2014, 34(9): 2417-2431.)

[5] Aghabozorgi S, Shirkhorshidi A S, Wah T Y. Time-series clustering — A decade review[J]. Information Systems, 2015, 53: 16-38.

[6] Marti G, Nielsen F, Donnat P. Optimal copula transport for clustering multivariate time series[C]. IEEE Int Conf on Acoustics, Speech and Signal Processing. Shanghai, 2016: 2379-2383.

[7] Sun J. Clustering multivariate time series based on Riemannian manifold[J]. Electronics Letters, 2016, 52(19): 1607-1609.

[8] Brandmaier A M. Pdc: Permutation distribution clustering[J]. Psychological Methods, 2015, 18(1): 71-86.

[9] D'Urso P, Maharaj E A. Wavelets-based clustering of multivariate time series[J]. Fuzzy Sets and Systems, 2012, 193(4): 33-61.

[10] Barragan J, Fontes C H, Embiruçu M. A wavelet-based clustering of multivariate time series using a multiscale SPCA approach[J]. Computers & Industrial Engineering, 2016, 95(5): 144-155.

[11] 于重重,吴子珺,谭励,等. 多元时序模糊聚类分段挖掘算法[J]. 北京科技大学学报, 2014, 36(2): 260-265.
(Yu C C, Wu Z J, Tan L, et al. Multivariate time series fuzzy clustering segmentation mining algorithm[J]. J of University of Science and Technology Beijing, 2014, 36(2): 260-265.)

[12] Li H. Distance measure with improved lower bound for multivariate time series[J]. Physica A: Statistical Mechanics and its Applications, 2017, 468: 622-637.

[13] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[14] 李正欣, 张凤鸣, 张晓丰, 等. 多元时间序列相似性搜索研究综述[J]. 控制与决策, 2017, 32(4): 577-583.
(Li Z X, Zhang F M, Zhang X F, et al. Survey of similarity search for multivariate time series[J]. Control and Decision, 2017, 32(4): 577-583.)

[15] 李海林. 基于变量相关性的多元时间序列特征表示[J]. 控制与决策, 2015, 30(3): 441-447.
(Li H L. Feature representation of multivariate time series based on correlation among variables[J]. Control and Decision, 2015, 30(3): 441-447.)

[16] Salvador S, Chan P. Toward accurate dynamic time warping in linear time and space[J]. Intelligent Data Analysis, 2007, 11(5): 561-580.

[17] Li H. On-line and dynamic time warping for time series data mining[J]. Int J of Machine Learning and Cybernetics, 2015, 6(1): 145-153.

[18] Sun L, Guo C. Incremental affinity propagation clustering based on message passing[J]. IEEE Trans on Knowledge and Data Engineering, 2014, 26(11): 2731-2744.

[19] 郭昆,郭文忠,邱启荣, 等. 基于局部近邻传播及用户特征的社区识别算法[J]. 通信学报, 2017, 36(2): 68-79.
(Guo K, Guo W Z, Qiu Q R, et al. Community detection algorithm based on local affinity propagation and user profile[J]. J on Communication, 2017, 36(2): 68-79.)

[20] 张建朋,陈福才,李邵梅, 等. 基于密度与近邻传播的数据流聚类算法[J]. 自动化学报, 2014, 40(2): 277-288.
(Zhang J P, Chen F C, Li S M, et al. Data stream clustering algorithm based on density and affinity propagation techniques[J]. Acta Automatica Sinica, 2014, 40(2): 277-288.)

(责任编辑: 李君玲)