时间序列数据挖掘中的聚类研究综述

李海林[1,2*]，张丽萍[1]

(1. 华侨大学信息管理与信息系统系 福建 泉州 362021；2. 华侨大学应用统计与大数据研究中心 福建 厦门 361021)

【摘要】鉴于时间序列数据的高维性和复杂性给数据挖掘带来的困扰以及聚类分析在时间序列数据挖掘领域中的重要性，对目前该领域国内外相关时间序列数据聚类研究的状况进行综述。时间序列聚类总体上可分为整体时间序列聚类、子序列聚类和时间点聚类 3 种，分别从特征表示、相似性度量、聚类算法和簇原型等方面来研究，同时也结合了具体的应用分析。根据时间序列数据挖掘中聚类存在的主要问题，提出了部分未来值得关注和研究的内容和方向，以便更好地促进时间序列数据聚类分析的研究与发展。

关 键 词 聚类分析； 数据挖掘； 高维性； 时间序列； 时间序列聚类

中图分类号 TP273 **文献标志码** A doi:10.12178/1001-0548.2022055

Summary of Clustering Research in Time Series Data Mining

LI Hailin[1,2*] and ZHANG Liping[1]

(1. Department of Information Management and Information Systems, Huaqiao University Quanzhou Fujian 362021;
2. Research Center for Applied Statistics and Big Data, Huaqiao University Xiamen Fujian 361021)

Abstract In view of the high dimensionality and complexity of time series data bringing trouble to data mining and the importance of clustering analysis in the field of time series data mining, this paper summarizes the research status of time series data clustering at home and abroad. Time series clustering can be divided into the whole-time-series clustering, the subsequence clustering, and it can be studied from the aspects of feature representation, similarity measurement, clustering algorithm and cluster prototype, as well as the specific applications analysis. According to the main problems existed in the time series clustering, this work proposes some contents and directions that are worthy of being researched in the future. All the work is to better promote the research and development of time series data clustering.

Key words clustering analysis; data mining; high dimensionality; time series; time series clustering

大数据背景下，数据挖掘与分析成为信息处理和知识管理等相关学科领域重点关注的研究对象[1]。在各种复杂数据类型中，广泛存在于金融市场和工业工程等领域的时间序列是一种与时间密切相关的数据，根据变量属性维度的大小其可分为单变量和多变量两种时间序列。相应地，时间序列数据挖掘是从时间序列数据库中发现信息与知识的理论与方法，为帮助政府和企业管理者在相关领域中提供更为可靠的辅助决策与技术支持[2]。时间序列的高维性具有时间维度长、属性变量多、数据体量大等特征，给传统数据挖掘技术的实施带来了极大困扰，在一定程度上阻碍了其在时间序列数据分析领域中的应用与发展。因此，运用数据挖掘技术从高维时间序列数据中发现信息和知识成为了数据分析领域中具有挑战性且最主要的研究方向之一[3]。

传统时间序列数据分析主要基于某种数据分布假设，再选取和制定计量经济模型来对时间序列数据预测分析。在大数据时代，除了需要传统的统计模型对时间序列数据进行预测与分析之外，鉴于时间序列数据具有时间维度长、属性变量多和数据体量大等高维性特征，借助机器学习、模式识别、智能计算和数据挖掘等模型和算法对高维时间序列数据可以进行深入研究与挖掘。聚类是数据挖掘相关研究和应用中非常重要

收稿日期：2022－02－25；修回日期：2022－03－28
基金项目：国家自然科学基金面上项目 (71771094)
作者简介：李海林 (1982－)，男，博士，教授，主要从事数据挖掘与人工智能方面的研究.
*通信作者：李海林，E-mail：blihailin@163.com

的方法，涉及计算机科学、模式识别、人工智能和机器学习等多个研究领域，同时也常被用于教育、营销、医学和生物信息学等学科，在大数据、人工智能和机器人等热点领域有突出贡献[4]。如在大规模群体决策中，聚类分析被用于划分大规模群体、处理非合作行为和社区发现等[5-6]。聚类分析也是一项重要而且基础的工作，其过程包括了时间序列的数据表达、特征提取、相似性度量以及具体聚类模型与算法等。为此，本文对时间数据挖掘中的聚类分析进行综述研究，首先介绍了目前时间序列聚类方法分类，然后分别从特征表示、相似性度量、聚类算法和簇原型等方面进行国内外研究状况分析，最后分析了目前研究存在的不足，同时给出了未来的研究方向。

1　时间序列聚类

时间序列聚类研究大体上可分为 3 种类型[7]，分别为整体时间序列聚类、子序列聚类和时间点聚类。整体时间序列聚类把每条时间序列视为数据对象，对具有共同数据特征的时间序列对象进行聚类。它常以相似性度量为基础，结合数据降维和特征表示来找出两个数据对象之间的共性，进而实现时间序列数据的簇划分。

如图 1 所示，分别使用主成分分析 (principal component analysis, PCA) 和对称性主成分分析 (asynchronism-based principal component analysis, APCA)[8] 对 10 条 Synthetic_Control 时间序列数据进行特征表示，并使用相应的相似性度量方法结合层次聚类实现整体时间序列的聚类分析。

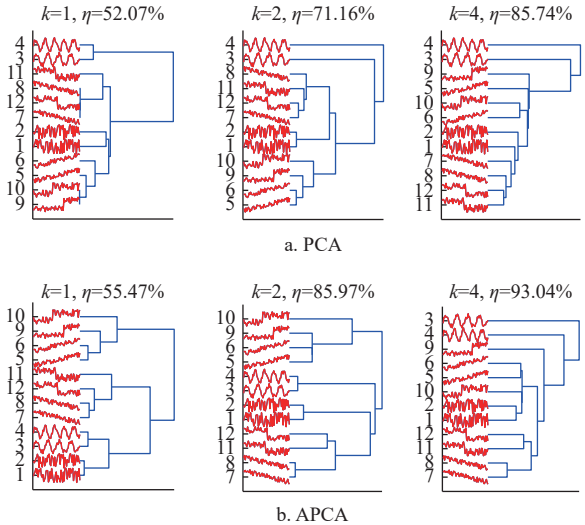


图 1　两种方法对整体时间序列数据层次聚类

子序列聚类通常指对一条时间较长的一元序列利用滑动窗和矢量量化等方法进行子序列划分，并使用相应聚类方法实现分段子序列的聚类。子序列聚类方法可以有效地发现较长时间序列中的频繁模式和异常片段，也能够发现不同时间序列数据之间存在的共同模式和关联关系。

时间点聚类则是从时间点和相应数据点两个角度出发来研究基于时间点的数据对象之间的近似性，把具有较高相似性的时间点聚合成同一簇，进而实现时间序列数据点划分[7-8]。该方法能够用来对一条时间序列进行分段划分，实现数据降维和特征表示，与传统时间序列分割表示方法相比，具有较高的时间效率。

目前国内外学者对于子序列聚类的研究目前尚存一些争议[9]。鉴于整体时间序列聚类的模型与算法可直接或间接应用于子序列聚类和时间点聚类，大部分集中于对整体时间序列聚类的研究。主要研究方法有：1) 传统聚类方法，如 K-Means、模糊聚类和基于密度的等聚类方法，根据时间序列的数据特征定制合适的距离度量函数，实现原始时间序列数据聚类[10]；2) 对时间序列数据通过特征空间转化[11]，将原始时间序列数据转化为另一特征空间的数据对象，再选取合适的传统聚类方法在特征空间中对数据对象进行聚类[12]；3) 通过时间序列数据的多分辨率解析，在不同分辨率视角下结合不同方法进行聚类分析，提升传统方法的聚类效果[13]。

2　国内外研究现状

针对时间序列数据挖掘中的聚类分析主要集中在整体时间序列的聚类研究，通常整体时序数据聚类方法也可用于子序列聚类中，使得整体时间序列聚类显得更为重要。由于时间的连续性，对时间点聚类的研究相对较少。

如图 2 所示，重点从整体时间序列聚类的视角来分析时间序列数据挖掘领域中的聚类研究状况。有关整体时间序列聚类的国内外相关文献主要从 4 个方向对其进行了相关理论和方法的研究，分别为数据降维与特征表示、相似性度量、聚类模型与算法和簇原型，采用不同的技术手段和理论方法从这 4 个方向进行分析与探究。

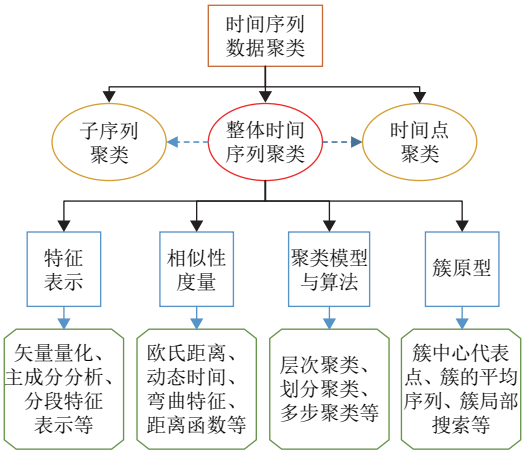


图 2 时间序列数据聚类的主要研究问题

2.1 研究地位

目前已经出现了不少成熟经典的聚类模型与算法，但一些基本问题始终是该领域的研究重点，其中包括不同结构特征数据的相似性度量、高维数据的降维与特征表示、基于噪声数据的聚类鲁棒性、大规模数据集聚类算法的有效性选择等[14]。高维时间序列数据与传统数据不同，随着时间维度的增加，各时间点产生的数据具有不确定性[15]，在聚类分析过程中除了要解决因高维性给有关模型和算法带来精度不高和复杂度过大等问题，还需要考虑动态实时、不确定性和高噪声等其他特征因素给聚类结果带来的影响。另外，时间序列聚类结果所产生的模式通常也被用于其他时间序列挖掘任务和方法中，如时间序列的数据降维与特征表示、模式匹配、关联分析、分类、数据可视化等[16-18]，使得整个时间序列数据挖掘任务具有更为出色的效果。

时间序列数据挖掘包括特征表示、相似性度量、聚类、分类、关联规则、模式发现和可视化等重要任务和关键技术[2]。聚类分析与特征表示和相似性度量方法一样，通常作为其他时间序列挖掘任务的子程序或中间件，以便更好地提升相关挖掘技术的性能和质量[10]。时间序列聚类分析研究的另一个重要动力来自于实际应用领域中超大容量数据的获取，包括经济金融、电子信息、医疗行业、航空航天、天体气象等。这些与时间相关的高维数据隐藏着大量有价值的信息和知识，需要通过聚类分析对时间序列数据进行模式发现，进而有针对性地对相关模式和知识进行处理，以便数据科学家和管理者进行技术分析和决策支持。

由于时间序列数据自身存在一定的特殊性，使得数据降维与特征表示以及相似性度量方法成为其他时间序列数据挖掘方法研究的基础任务，其质量好坏在一定程度上影响其他挖掘任务的效果[19]。文献 [20] 对单变量和多变量两种时间序列数据的特征表示和相似性度量进行了较为系统的研究，研究成果能较好地改善和提高有关挖掘技术和方法的质量和效率。同时，聚类自身也可用来发现时间序列中的频繁模式或时间序列数据库中的奇异模式，甚至作为一种降维手段来实现数据特征表示[21]。另外，在大部分情况下，时间序列聚类通常是建立在特征表示和相似性度量基础之上的一种机器学习方法，实现获得较高质量的聚类分析结果[10]。

2.2 数据降维与特征表示

数据降维和特征表示是高维时间序列数据挖掘中至关重要的过程，其目的是对高维数据进行数据变换，在低维空间下使用相应的特征来表示原始时间序列的关键信息，进而提高时间序列聚类算法的效率和质量。目前，已有一些较为成熟的方法对一元时间序列进行特征表示，包括矢量量化[22]、分段表示[23]、聚合符号化表示[24]、多项式回归参数[25] 和模型参数[26] 等。鉴于多元时间序列数据的广泛性和重要性，主要从序列的时间和属性两个维度进行数据降维，代表性方法有基于主成分分析的[27]、基于独立成成分的[12]、基于奇异值分解的[28] 等。

将时间序列数据转化为复杂网络方法，再使用复杂网络的拓扑结构特征来表示原始时间序列数据也是目前较为常用的一种时间序列数据特征表示方法，通常包括基于可视图、基于相空间重构法、基于递归法和基于符号模式等建网方法[29]。特别地，可视图可以将周期时间序列、随机时间序列和分形时间序列分别转化为规则网络、随机网络和无标度网，其拓扑结构能够较好地反映时间序列的数据特征。若时间序列中两个数据所表示的直方条能够画一条不与任何中间直方条相交的直线，则此直方条组所对应的数据组之间可以形成网络连边，即：

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a} \tag{1}$$

时间序列任意两点$a(t_a, y_a)$ 和$b(t_b, y_b)$ 之间有一点$c(t_c, y_c)$，满足式 (1) 说明数据点 a 和 b 满足可见性。通过复杂网络的节点度中心性、聚集系数、结构洞等拓扑结构特征来描述时间序列的特征，结合常用的聚类分析方法来有效实现时间序列数据的聚类。

基于数据降维和特征表示的时间序列聚类主要从基于形态的、基于特征的和基于模式的等方面来

研究。基于形态的时间序列聚类[30]主要从数据形态变化的角度来匹配序列之间的相似性，包括同步形态和异步形态，进而聚类算法可将具有相似性形态变化特征的时序对象归入同一簇。基于特征的时间序列聚类[31]将时间序列进行数据转化，在低维的特征空间中进行时间序列的聚类分析。基于模式的时间序列聚类[10]则是将原始时间序列转化为模型参数，结合传统聚类算法实现时间序列的模式识别。

2.3　相似性度量

相似性度量也是时间序列聚类算法中必不可少的中间件，基于相似性度量的聚类算法有时间序列数据划分聚类、层次聚类和基于密度的聚类等。文献 [32] 提出了时间序列相似性搜索过程中距离度量的理论基础，要求设计的快速近似度量函数满足真实距离的下界性，以免相似性检索时发生漏报情况。

目前存在各种不同的时间序列距离度量方法[19]，最典型的两种方法为欧氏距离 (Euclidean distance, ED) 和动态时间弯曲方法 (dynamic time warping, DTW)[11, 33-34]。欧氏距离通常要求两条时间序列具有相等的长度，即对于两条时间序列 A 与 B，有：

$$\mathrm{ED}(A,B)=\sqrt{\sum_{i=1}^{m}(a_i-b_i)^2} \tag{2}$$

式中，$|A|=|B|=m$；a_i与b_i 分别为时间序列 A 与 B 中的第 i 个数据点。DTW 则是从两条时间序列中的数据点之间寻找一条最优弯曲路径 P，并且最优弯曲路径 P 的元素$p_k=(i,j)_k$要满足边界性、单调性和连续性等条件，使得最优路径产生的距离值最优，即：

$$\mathrm{DTW}(A,B)=\min_{P}\left\{\frac{1}{K}\sum_{k=1}^{K}\mathrm{dist}(p_k)\right\} \tag{3}$$

式中，$\mathrm{dist}(p_k)=\left\|a_i-b_j\right\|_2$。

如图 3 所示，欧氏距离对时间序列进行了同步硬性度量，动态时间弯曲方法根据最优化匹配路径，实现异步相似形态的度量。前者满足三角不等式，比较适用于时间序列的相似性搜索，但其结果易受异常数据点的影响，且无法度量不等长时间序列之间的相似性；后者利用动态规划方法从两条时间序列中找到一条距离最优的弯曲路径，使具有相似形态的异步数据相互匹配，进而实现不等长时间序列之间的距离度量，但其平方阶的时间复杂度限制了其在高维时间序列数据聚类过程中的应用。

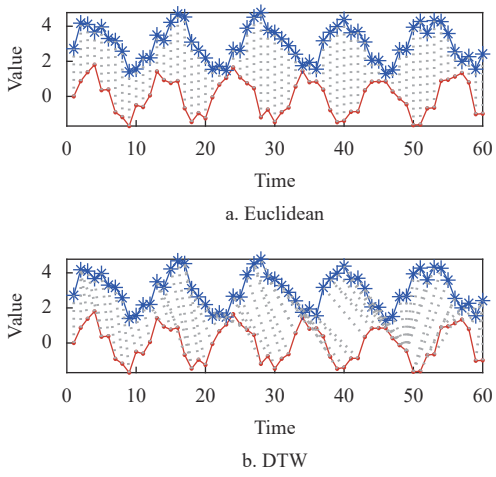


图 3　欧氏距离与动态时间弯曲度量

基于形状的距离度量 (shape-based distance, SBD)[35]是一种通过寻找两条时间序列之间的最优子序列交叉相关性来反映原始时间序列数据的相似性。针对两条时间序列 A 与 B，长度为$2m-1$的交叉相关性序列可定义为：

$$\mathrm{CC}_s(A,B)=R_{s-t}(A,B)\quad s\in\{1,2,\cdots,2m-1\}$$

式中，

$$R_g(A,B)=\begin{cases}\sum_{l=1}^{t-g}a_{l+g}b_l & g\geqslant 0\\ R_{-g}(B,A) & g<0\end{cases} \tag{4}$$

通过查找标准化交叉相关性的值计算两条时间序列的形状距离值，即：

$$\mathrm{SBD}(A,B)=1-\max_{s}\left(\frac{\mathrm{CC}_s(A,B)}{\sqrt{R_0(A,A)R_0(B,B)}}\right) \tag{5}$$

大量实验表明，在时间序列数据聚类中，使用 SBD 可以获得比使用 DTW 更好的聚类性能和效果。

另外，一些基于特征表示的距离度量方法也常用于时间序列的聚类分析，如基于多项式参数的[25]和基于主成分分析的[27]等距离度量方法在时间序列数据挖掘中起到了提升聚类效果的作用。

2.4　聚类算法

时间序列数据聚类主要包括层次聚类、划分聚类、基于模型的聚类、基于密度的聚类、基于格的聚类和多步聚类等[18]。时间序列层次聚类[36]是一种具有直观效果的聚类方法，分为基于凝聚和基于分裂的层次聚类。特别地，为了检索特征表示或相似

性度量方法的有效性，通常被用来直观显示基于形态的或基于特征的时间序列聚类情况。

划分聚类[37] 是时间序列聚类算法研究中最为常用的方法之一，通常借助于相似性度量函数来实现簇划分，具体方法包括 K-Means、K-Medoids 和 FCM。例如，在时间序列 SBD 距离计算中，使用 K-Means 的思想来对时间序列进行快速有效地聚类，通过寻找最优参数来达到目标评价函数最优，即：

$$\boldsymbol{\mu}_k^* = \arg\max_{\boldsymbol{x}_i \in P_k} \mathrm{NCC}_c(\boldsymbol{x}_i, \boldsymbol{\mu}_k)^2 \tag{6}$$

式中，$\mathrm{NCC}_c(\cdot,\cdot)$用于计算两条时间序列之间的标准化交叉相关性；$\boldsymbol{\mu}_k$ 为簇中心代表序列；P_k 表示第 k 个簇中的时间序列数据集。

基于划分的聚类方法需要事先设定聚类个数，但在应用中通常无法确定聚类个数，特别是对海量高维时间序列数据来说，该参数的确定显得更加困难。文献 [38] 研究了适当的初始中心对时间序列 K-Means 聚类的质量和效率有很大影响。文献 [39] 认为 K-Means 和 K-Medoids 与层次聚类相比，其具有较好的时间性能，比较适用于时间序列的聚类分析。与这两种聚类相比，FCM 是一种基于模糊理论的软划分，该方法在一定程度上考虑了时间序列数据对象的不确定性问题[40]。

基于模型的聚类与其他方法不同，它假设同一簇中数据服从某种模型的数据分布，通过数据模型学习来试图调整近似模型，使其接近数据客观存在的真实模型[41]。目前也有一些较为成熟的方法[10]，如自组织映射、多项式回归分析、高斯混合模型、ARIMA 模型、马尔可夫链和隐马尔可夫模型等。然而，基于模型的聚类方法存在一些问题有待研究：一方面，模型需要用户事先设定假设模型和模型参数，若假设模型与真实模型相差甚远，则会导致最终的聚类结果不准确；另一方面，此类模型需要较长的计算时间，不利于高维时间序列数据和动态时间相关数据的聚类分析。

基于密度的和基于格的聚类方法[42-43] 先将时间序列数据转化为另一种数据形态，使其能够适用于传统数据挖掘中的聚类算法，如 DBSCAN、OPTICS、STING 和 Wave Cluster 等方法。多步聚类方法[44] 则是从聚类算法设计和分析的角度出发，通过多种方法对时间序列数据进行分步聚类，其效果通常要优于传统基于特征表示的、基于相似性度量的和基于模型的聚类方法。

数据挖掘中的聚类算法[4] 较为成熟，除了具有较为完善的理论基础，其在许多领域都有很好的应用效果，因此，它们也可以直接或间接应用于时间序列数据的聚类分析。然而，由于时间序列数据具有时间和变量高维性、概念漂移、随机性和混沌现象等特点[29,45]，需要进行数据降维、特征表示和相似性度量，也包括异常点发现等前期处理工作。根据传统聚类算法思路设计适用于时间序列数据聚类的模型和算法，如将传统聚类思路结合复杂网络特征，实现多变量时间序列数据的聚类[46-47]。

2.5 簇原型

簇原型[48] 是指某一特定簇的近似代表对象，其质量好坏直接影响某些聚类算法的分类效果，如 K 均值、模糊聚类和近邻传播聚类 (AP) 等算法都需要定义相应的簇原型。在时间序列数据聚类领域中，簇原型大体可分为 3 种，分别是簇中心代表点[49]、簇的平均序列[50] 和基于局部搜索的簇原型[37]。

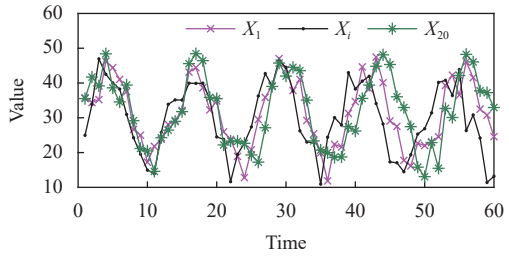
簇中心代表点是从簇中找到某个时间序列数据对象作为簇的代表对象，该对象到簇中其他对象的平均距离最小。簇平均序列是指对簇中的所有数据对象在相同时间点作平均值并随时间顺序获得的序列。然而，该方法通常需要簇中所有数据对象具有相同的长度，其结果不能很好地处理同簇中时间序列数据对象之间具有异步相似形态的情况。为此，文献 [50] 提出了基于动态时间弯曲的平均序列表示方法 (DTW barycenter averaging, DBA)，并将其应用于时间序列聚类算法中的簇原型，使其能够将具有异步相似形态的不等长时间序列数据实现较高质量的聚类。假设存在一个簇中心序列 $S=[s_1,s_2,\cdots,s_w]$，使用 DTW 度量它与其他簇成员时间序列存在的最优弯曲匹配路径，即 s_i 与其他 l 条时间序列的数据点集 $A_j=\{a_j^1,a_j^2,\cdots,a_j^{k_j}\}$ 相匹配，则有：

$$s_i = \frac{\sum_{j=1}^{l}\sum_{k=1}^{k_j} a_j^k}{\sum_{j=1}^{l} k_j} \tag{7}$$

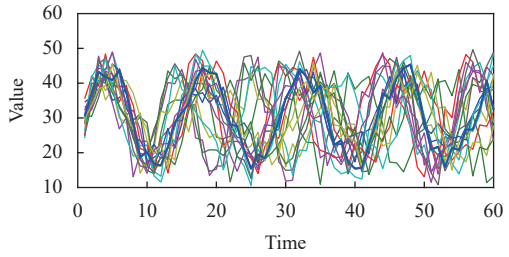
通过基于 DTW 的距离计算来重复交替迭代计算簇中心序列和分配簇成员，实现时间序列数据的聚类。

如图 4 所示，图 4a 显示了同一个簇中 3 条时间序列样本的形态波动示例，图 4b 中较粗曲线表

示了 DBA 方法的簇平均序列，易发现基于 DBA 的簇平均序列的形态波动与簇成员的形态波动相似。基于局部搜索的簇原型[37] 是一种在簇类中进行局部搜索找出簇原型的方法，与基于簇中心代表点的和基于簇平均序列的 K 中心点聚类算法相比，基于局部搜索簇原型的 K 中心聚类具有较好的挖掘效果。



a. Time series samples



b. DBA

图 4　簇平均代表序列

2.6　应用研究

时间序列数据挖掘中的聚类研究成果主要应用于两个方面：1) 将聚类算法作为其他时间序列数据挖掘技术和方法的子程序或中间件，其聚类结果可以辅助其他数据分析任务的顺利进行并提高数据挖掘任务的效果[51-52]；2) 聚类算法可被运用在具体的实际生产和生活领域中，如生物信息、天体气象、经济金融、医疗卫生、语音识别和工业工程等[53-57]，根据具体背景知识来发现相关行业时间序列数据中的兴趣模式、异常模式和频繁模式。

在工业工程领域中，文献 [58] 提出了一种基于灰色关联聚类的特征提取算法，利用灰色关联度作为动态聚类欧氏距离的思想，构建以某型涡扇发动机为例的灰色关联聚类特征提取模型，以便满足故障诊断要求。特别地，在金融数据分析应用领域中，通过时间序列聚类分析方法可以发现股票市场中相似的股票群，结合滑窗法可以实现基于动态衍化聚类的股票识别。金融市场被众多因素共同影响，不仅有宏观的政治、经济等环境因素，还有微观的企业运作方式、人们的心理作用等因素。通过对金融时间序列进行聚类，可以挖掘出金融市场的内在机制，对揭示数据背后的发展变化规律有重要作用。文献 [59] 提出一种基于影响力计算模型的股票网络中心节点层次聚类算法，利用社区发现方法对股票时间序列进行聚类。文献 [60] 提出一种基于支持向量回归和自组织神经网络的聚类方法，提取投资组合选择方法，实现了金融股票价格和波动率的预测分析，对印度国家证券交易所 102 支股票进行最优投资组合，具有低风险高盈利的特征。文献 [61] 针对金融股价时间序列数据的时间属性变量高维性，提出利用三阶段聚类模型，对股票进行增量式聚类，进而发现上市公司之间的联动关系。

3　主要问题与未来研究方向

时间序列数据聚类研究主要集中于整体序列数据对象，在数据降维与特征表示、相似性度量、聚类模型与算法、簇原型和应用等方法上取得了一定的学术进展，拥有应用价值，但仍存在一些问题有待探讨，以便系统性地研究和提高聚类分析在时间序列数据资料中的挖掘质量和应用性能。

1) 传统时间序列聚类模型与算法主要以一元时间序列数据为研究对象，通过数据转换、特征表示或模型参数实现时间维度的降维，利用数据挖掘中经典聚类方法进行分析，缺少兼顾时间序列数据时间维度长、属性变量多和数据体量大等高维性的问题。特别地，数据体量大造成算法在运行期间需要消耗巨大的内存空间，使得以静态处理方式为基础的传统聚类算法在此类高维时间序列数据集中无法得到有效地运行。因此，根据高维时间序列数据自身的特点，需要研究适用于高维时间序列数据的实时、动态或者增量运行的聚类算法。

2) 基于原始数据的时间序列距离度量通常需要较大的时间复杂度和空间复杂度，并且大部分距离度量方法对数据中的扭曲数据较为敏感，使其在计算过程中无法获得直观的度量效果。数据降维后的特征表示在一定程度上能够改善此种境况，降低了聚类算法和模型的复杂性，但特征表示对高维原始时间序列数据的精确定位难以实现，最终影响聚类模型与算法的精度。针对基于特征表示和相似性度量方法的高维时间序列数据聚类模型与算法研究，需要改变传统方法仅对高维数据进行某个特定维度上的数据降维和特征表示，制定适用于特征表示后由于信息丢失而造成距离度量不准确的情况，进而提升相关模型与算法的聚类效果。

3) 动态时间弯曲是时间序列数据挖掘领域中最为常用的相似性度量方法，它能有效地匹配时间序列数据中的近似形态趋势，对时间点异常数据不敏感，能够度量不等长时间序列之间的相似性，具有较好的度量质量和较高的准确性等优势。在高维时间序列数据库中，过长的时间维度、过多的属性变量和过大的数据体量造成动态时间弯曲方法容易平滑时间序列局部形态的特点，不能有效反映高维时间序列数据对象之间的形态变化关系，无法实现真实距离的有效度量，进而影响基于动态时间弯曲的高维时间序列聚类效果。如何提高动态时间弯曲方法的精确度量和计算性能是高维时间序列聚类研究中需要解决的问题之一。

4) 关于时间序列聚类的研究目前大多数集中在提升特征表示、距离度量和簇原型的质量或效率上，对于聚类本身的设计与研究相对不足。虽然已有学者利用多步聚类方法在一定程度上改进了传统聚类算法在时间序列数据中的分析效果，但也存在步骤繁琐、聚类结果易受参数设置影响、计算性能较低等问题。同时，由于时间序列数据高维性和其他特征因素的影响，聚类方法在相关应用中大多数局限于变量属性少、时间维度短和数据量少等对象的分析，较少考虑不确定性和高噪声等因素的影响，使得聚类分析理论和方法在实际应用中具有局限性。为此，通过总结现有的时间序列聚类算法优缺点，结合具体问题中的数据特征，在考虑多种特征因素影响的情况下来构建符合高维时间序列数据的高性能聚类算法值得深入研究。

4 结 束 语

本文梳理了目前常用的时序聚类算法，综述了该领域中的相关研究成果，归纳了已有研究存在的不足，提出了一些值得研究的方向。研究发现，时间序列数据挖掘中聚类模型与算法的研究顺应了大数据时代潮流，解决了高维性给传统时间序列聚类分析带来不能快速有效挖掘的问题，提高和拓展了时间序列数据挖掘领域中的相关理论和方法。时间序列数据聚类研究成果能给政府部门和企业对相关事务决策提供更为完备成熟的理论基础与技术，以便进行更为科学合理的智能管理。

参 考 文 献

[1] 李学龙, 龚海刚. 大数据系统综述[J]. 中国科学: 信息科学, 2015, 45(1): 1-44.
LI X L, GONG H G. A survey on big data systems[J]. Scientia Sinica (Information Sciences), 2015, 45(1): 1-44.
[2] ESLING P, AGON C. Time-series data mining[J]. ACM Computing Surveys, 2012, 45(1): 1-12.
[3] YANG Q, WU X. 10 challenging problems in data mining research[J]. International Journal of Information Technology & Decision Making, 2006, 5(4): 597-604.
[4] EZUGWUA A E, IKOTUNA A M, OYELADE O O, et al. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects[J]. Engineering Applications of Artificial Intelligence, 2022, 110: 104743.
[5] CHAO X, KOU G, PENG Y, et al. Large-scale group decision-making with non-cooperative behaviors and heterogeneous preferences: An application in financial inclusion[J]. European Journal of Operational Research, 2021, 288(1): 271-293.
[6] CHAO X, KOU G, PENG Y, et al. An efficient consensus reaching framework for large-scale social network group decision making and its application in urban resettlement[J]. Information Sciences, 2021, 575: 499-527.
[7] MÖRCHEN F, ULTSCH A, HOOS O. Extracting interpretable muscle activation patterns with time series knowledge mining[J]. International Journal of Knowledge Based Intelligent Engineering Systems, 2005, 9(3): 197-208.
[8] LI H L. Asynchronism-based principal component analysis for time series data mining[J]. Expert Systems with Applications, 2014, 41(C): 2842-2850.
[9] KEOGH E, LIN J. Clustering of time-series subsequences is meaningless: Implications for previous and future research[J]. Knowledge and Information Systems, 2005, 8(2): 154-177.
[10] LIAO T W. Clustering of time series data—a survey[J]. Pattern Recognition, 2005, 38(11): 1857-1874.
[11] 李海林, 梁叶, 王少春. 时间序列数据挖掘中的动态时间弯曲研究综述[J]. 控制与决策, 2018, 33(8): 1345-1353.
LI H L, LIANG Y, WANG S C. Review on dynamic time warping in time series data mining[J]. Control and Decision, 2018, 33(8): 1345-1353.
[12] 郭崇慧, 苏木亚. 基于独立成分分析的时间序列谱聚类方法[J]. 系统工程理论与实践, 2011, 31(10): 1921-1931.
GUO C H, SU M Y. Spectral clustering method based on independent component analysis for time series[J]. Systems Engineering-Theory & Practice, 2011, 31(10): 1921-1931.
[13] LAI C P, CHUNG P C, TSENG V S. A novel two-level clustering method for time series data analysis[J]. Expert Systems with Applications, 2010, 37(9): 6319-6326.
[14] 王骏, 王士同, 邓赵红. 聚类分析研究中的若干问题[J]. 控制与决策, 2012, 27(3): 321-328.
WANG J, WANG S T, DENG Z H. Survey on challenges in clustering analysis research[J]. Control and Decision, 2012, 27(3): 321-328.
[15] 张旭, 张亮, 金博, 等. 基于不确定性的多元时间序列分类算法研究[J]. 自动化学报, 2021, 48(11): 1-15.
ZHANG X, ZHANG L, JIN B, et al. Uncertainty-based

multivariate time series classification[J]. Acta Automatica Sinica, 2021, 48(11): 1-15.
[16] 施沫寒, 王志海. 一种基于时间序列特征的可解释步态识别方法[J]. 中国科学:信息科学, 2020, 50(3): 438-460.
SHI M H, WANG Z H. An interpretable gait recognition method based on time series features[J]. Science in China (Information Sciences), 2020, 50(3): 438-460.
[17] SORIANO-VARGAS A, WERNECK R, MOURA R, et al. A visual analytics approach to anomaly detection in hydrocarbon reservoir time series data[J]. Journal of Petroleum Science and Engineering, 2021, 206: 108988.
[18] LI X, LIN J, ZHAO L. Time series clustering in linear time complexity[J]. Data Mining and Knowledge Discovery, 2021, 35: 2369-2388.
[19] 孙冬璞, 曲丽. 时间序列特征表示与相似性度量研究综述[J]. 计算机科学与探索, 2021, 15(2): 195-205.
SUN D P, QU L. Survey on feature representation and similarity measurement of rime series[J]. Journal of Frontiers of Computer Science and Technology, 2021, 15(2): 195-205.
[20] 李海林. 时间序列数据挖掘中的特征表示与相似性度量方法研究[D]. 大连: 大连理工大学, 2012.
LI H L. Research on feature representation and similarity measure methods in time series data mining[D]. Dalian: Dalian University of Technology, 2012.
[21] LI H L. Accurate and efficient classification based on common principal components analysis for multivariate time series[J]. Neurocomputing, 2016, 171(1): 744-753.
[22] 范晓诗, 雷英杰, 路艳丽, 等. 基于 DTW 的长期直觉模糊时间序列预测模型[J]. 通信学报, 2016, 37(8): 95-104.
FAN X S, LEI Y J, LU Y L, et al. Long-term intuitionistic fuzzy time series forecasting model based on DTW[J]. Journal on Communications, 2016, 37(8): 95-104.
[23] 江艺美. 基于 GM(1, 1) 模型的时间序列分段表示方法[J]. 系统工程, 2014(7): 137-142.
JIANG Y X. Method of time series piecewise representation based on model GM(1, 1)[J]. Systems Engineering, 2014(7): 137-142.
[24] SHIEH J, KEOGH E. ISAX: Indexing and mining terabyte sized time series[C]//Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S.l.]: ACM, 2008: 623-631.
[25] FUCHS E, GRUBER T, NITSCHKE J, et al. On-line segmentation of time series based on polynomial least-squares approximation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(12): 2232-2245.
[26] D'URSO P, GIOVANNI L D, MASSARI R. GARCH-based robust clustering of time series[J]. Fuzzy Sets and Systems, 2016, 305(15): 1-28.
[27] KARAMITOPOULOS L, EVANGELIDIS G. PCA-based time series similarity search[J]. Data Mining Annals of Information Systems, 2010, 8: 255-276.
[28] WENG X, SHEN J. Classification of multivariate time series using two dimensional singular value decomposition[J]. Knowledge-Based Systems, 2008, 21(7): 535-539.
[29] ZOU Y, DONNER R V, MARWAN N, et al. Complex network approaches to nonlinear time series analysis[J]. Physics Reports, 2019, 787: 1-97.
[30] ZAKARIA J, MUEEN A, KEOGH E. Clustering time series using unsupervised-shapelets[C]//Proceedings of IEEE 12th International Conference on Data Mining (ICDM). [S.l.]: IEEE, 2012: 785-794.
[31] 谢聪慧, 吴世新, 张晨, 等. 基于谱系聚类的全球各国新冠疫情时间序列特征分析[J]. 地球信息科学学报, 2021, 23(2): 236-245.
XIE C H, WU S X, ZHANG C, et al. Analysis of time series features of COVID-19 in various countries based on pedigree clustering[J]. Journal of Geo-Information Science, 2021, 23(2): 236-245.
[32] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence databases[J]. Foundation of Data Organization Algorithms, 1993, 46: 69-84.
[33] LI H L. Time works well: Dynamic time warping based on time weighting for time series data mining[J]. Information Sciences, 2021, 547: 592-608.
[34] PETITJEAN F, FORESTIER G, WEBB G, et al. Dynamic time warping averaging of time series allows faster and more accurate classification[C]//Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM). [S.l.]: IEEE, 2014: 470-479.
[35] PAPARRIZOS J, GRAVANO L. Fast and accurate time-series clustering[J]. ACM Transactions on Database Systems, 2017, 42(2): 1-49.
[36] KEOGH E, PAZZANI M. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback[C]//Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining. New York: ACM, 1998: 239-241.
[37] HAUTAMÄKI V, NYKÄNEN P, FRÄNTI P. Time-series clustering by approximate prototypes[C]//Proceedings of the 19th International Conference on Pattern Recognition. Piscataway: IEEE, 2008: 1-4.
[38] FAYYAD U M, REINA C, BRADLEY P S. Initialization of iterative refinement clustering algorithms[C]// Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: ACM, 1998: 194-198.
[39] BERINGER J, HÜLLERMEIER E. Online clustering of parallel data streams[J]. Data & Knowledge Engineering, 2006, 58(2): 180-204.
[40] LI H L, WEI M. Fuzzy clustering based on feature weights for multivariate time series[J]. Knowledge-Based Systems, 2020, 197: 1-11.
[41] CORDUAS M, PICCOLO D. Time series clustering and classification by the autoregressive metric[J]. Computational Statistics & Data Analysis, 2008, 52(4): 1860-1872.
[42] CHANDRAKALA S, SEKHAR C C. A density based method for multivariate time series clustering in kernel feature space[C]//Proceedings of IEEE International Joint Conference on Neural Networks. Piscataway: IEEE, 2008: 1885-1890.

[43] WANG W, YANG J, MUNTZ R. STING: A statistical information grid approach to spatial data mining[C]// Proceedings of the International Conference on Very Large Data Bases. New York: ACM, 1997: 186-195.

[44] ZHANG X, LIU J, DU Y, et al. A novel clustering method on time series data[J]. Expert Systems with Applications, 2011, 38(9): 11891-11900.

[45] JASTRZEBSKAA A, NÁPOLESB G, SALGUEIRO Y, et al. Evaluating time series similarity using concept-based models[J]. Knowledge-based Systems, 2022, 238(28): 107811.

[46] LI H, LIU Z. Multivariate time series clustering based on complex network[J]. Pattern Recognition, 2021, 115: 107919.

[47] LI H, DU T. Multivariate time-series clustering based on component relationship networks[J]. Expert Systems with Applications, 2021, 173: 114649.

[48] BAGNALL A, JANACEK G. Clustering time series with clipped data[J]. Machine Learning, 2005, 58(2-3): 151-178.

[49] COPPI R, D'URSO P, GIORDANI P. Fuzzy c-medoids clustering models for time-varying data[M]//BERNADETTE B M, GIULIANELLA C, RONALD R Y. Modern Information Processing: From Theory Applications. [S.l.]: Elsevier, 2006: 195-206.

[50] PETITJEAN F, KETTERLIN A, GANÇARSKI P. A global averaging method for dynamic time warping with applications to clustering[J]. Pattern Recognition, 2011, 44(3): 678-693.

[51] 闫相斌, 李一军, 邹鹏, 等. 动静态属性数据相结合的客户分类方法研究[J]. 中国管理科学, 2005, 13(2): 95-100.
YAN X B, LI Y J, ZOU P, et al. Customer segmentation based on integration of dynamic and static attributes[J]. Chinese Journal of Management Science, 2005, 13(2): 95-100.

[52] CAI L, QU S, YUAN Y, et al. A clustering-ranking method for many-objective optimization[J]. Applied Soft Computing, 2015, 35: 681-694.

[53] 王德青, 何凌云, 朱建平. 基于函数型自适应聚类的股票收益波动模式比较[J]. 统计研究, 2018, 35(9): 79-91.
WANG D Q, HE L Y, ZHU J P. Comparing returns volatility patterns of SSE50 stock shares via functional adaptive clustering[J]. Statistical Research, 2018, 35(9): 79-91.

[54] LI H, WU Y J, CHEN Y. Time is money: Dynamic-model-based time series data-mining for correlation analysis of commodity sales[J]. Journal of Computational and Applied Mathematics, 2020, 370: 1-20.

[55] 刘乃龙, 周晓东, 刘钊铭, 等. 基于多变量时间序列的接触状态聚类分析[J]. 电子科技大学学报, 2020, 49(5): 660-665.
LIU N L, ZHOU X D, LIU Z M, et al. Contact state clustering analysis based on multivariate time series[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(5): 660-665.

[56] 李乃和, 汤世强. 特大型连续服务供应商的顾客细分与新顾客归类判别研究[J]. 中国管理科学, 2015(S1): 602-609.
LI N H, TANG S Q. Research on customers' segmentation and new comers' classification by super-large scale continuous service provider[J]. Chinese Journal of Management Science, 2015(S1): 602-609.

[57] VÍCTOR R, HÉCTOR D, DAVID C. Analysing temporal performance profiles of UAV operators using time series clustering[J]. Expert Systems with Applications, 2017, 70(15): 103-118.

[58] 鲁峰, 黄金泉. 基于灰色关联聚类的特征提取算法[J]. 系统工程理论与实践, 2012, 32(4): 872-876.
LU F, HUANG J Q. Feature extraction algorithm of clustering based on grey relational theory[J]. Systems Engineering-Theory & Practice, 2012, 32(4): 872-876.

[59] 王浩, 李国欢, 姚宏亮, 等. 基于影响力计算模型的股票网络社团划分方法[J]. 计算机研究与发展, 2014, 51(10): 2137-2147.
WANG H, LI G H, YAO H L, et al. Stock network community detection method based on influence calculating model[J]. Journal of Computer Research and Development, 2014, 51(10): 2137-2147.

[60] CHOUDHURY S, GHOSH S, BHATTACHARYA A, et al. A real time clustering and SVM based price-volatility prediction for optimal trading strategy[J]. Neurocomputing, 2014, 131: 419-426.

[61] AGHABOZORGI S, YING W T. Stock market co-movement assessment using a three-phase clustering method[J]. Expert Systems with Applications, 2014, 41(4): 1301-1314.

编 辑 叶 芳