DOI: 10.3969/j.issn.1009-6868.2018.03.011
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20180522.1113.002.html

基于结构特征的时序聚类方法研究

Time Series Clustering Based on Structural Features

中图分类号：TN929.5 文献标志码：A 文章编号：1009-6868 (2018) 03-0061-006

孟志浩/MENG Zhihao
刘建伟/LIU Jianwei
韩静/HAN Jing
（中兴通讯股份有限公司，广东 深圳 518057）
(ZTE Corporation, Shenzhen 518057, China)

摘要： 数据驱动的智能运维对提高云平台的管理效率有重要意义。提出一种基于结构特征的时序聚类方法以用于云平台大量性能数据的智能分类。该方法采用分级处理的方式用于降低聚类复杂度，首先基于傅里叶变换将时序分为明显周期型和非明显周期型两大类，然后从时序中提取季节性指标、趋势性指标、偏度、相对熵、样本熵、自相似性和李雅普诺夫系数等7个特征，最后在每个大类中基于特征空间进行K均值聚类分析。实验数据仿真表明：所提方法能够有效将不同波形特性的时序分开。

关键词： 特征提取；时序聚类；数据挖掘；云平台

Abstract: Data-driven intelligent Operation & Management (O&M) has significant importance for improving the efficiency of cloud platform maintenance. In this paper, a time series clustering method based on structural features is proposed for classifying large-scale metrics in cloud platform. A hierarchical scheme is adopted to reduce the complexity of clustering. First, the time series are classified into two big categories based on Fourier transformation: significant periodicity and non-significant periodicity. Secondly, seven features are extracted from the data: seasonal degree index, trend degree index, skewness, relative entropy, sample entropy, self-similarity and Lyapunov coefficient. And then, the k-means algorithm is used to cluster the time series in the feature space for each big category. The real data experiment shows that the method proposed is able to distinguish the time series which contain different characteristics.

Key words: feature extraction; time series clustering; data mining; cloud platform

计算机（IT）集群在各行各业均有广泛运用，以电信运营商为例，其核心网、网管中心和数据中心等均以IT集群为依托。一般来讲，IT集群规模庞大，配置的硬件和软件数目和种类繁多。IT集群又是对正常运行时间有严格要求的不间断系统，若出现软件错误和硬件故障不仅使用户体验急剧下降，而且耗费大量维护费用。因此集群的管理和运维一直是重要而又具有挑战性的任务。

随着虚拟化和软件自定义网络（SDN）等技术的引入，传统IT集群向云化转变，集群规模进一步增大，上层软件应用和业务类型日趋增多，所需监查控制的性能指标数量有百万级乃至更多。因此，传统人工设定阈值进行监查控制的方法已经难以满足应用需求，不仅人工成本增加，且运维效率和准度下降。基于机器学习实现智能化运维对解决此问题具有重要意义，已在业界得到普遍关注。采用时序建模和数据挖掘的方法根据性能数据的历史分布设定其阈值范围，实现自动化的动态阈值设定，可降低人工设定阈值的时间成本并提高阈值精准度[1]。另一方面，集群性能数据种类千差万别，时序分布特性各有不同，难以简单只采用一种时序建模算法就可以实现对所有序列的建模。因此，需要对不同特性的时序数据采用各自合适的阈值算法，才能更满足阈值设定精度。这对时序的自动分类提出了要求，在完成对时序的自动分类后，再根据其类别选择合适的阈值模型。

时序数据的分类在数据挖掘领域是一个多年的研究热点，其分析多个输入时间序列存在的共性与差异，将具有相同结构的序列归为一类，而将结构不同的序列尽量区分开来。相比于一般的聚类问题（静态聚类），由于时间序列带有时间维度的动态性，使其聚类问题变得更为复杂。因此，除了在原始时间序列空间做聚类分析，更有效的方法是通过间接的方式，先对时间序列做特征提取或建模，再进行聚类分析。本文中我们采用的方法即是先从时序中提取周期

收稿日期：2018-04-23
网络出版日期：2018-05-22
基金项目：上海市青年科技英才扬帆计划（18YF1423300）

性、趋势性、非高斯性等结构特征，然后在此特征空间对时序进行聚类。此外，为降低聚类复杂度，本文通过基于傅里叶变换的方法先将时序分为两大类，然后在各大类中进行聚类分析。在云平台实际采集的性能数据的仿真表明：本文所提方法具有有效性。

1 时序聚类简介

一般时间序列的聚类方法可分为3种[2]：基于原始信号的聚类、基于建模的聚类、基于特征提取的聚类。

（1）基于原始信号的聚类方法指直接在原始信号空间进行聚类分析。为测量信号之间的相似性，常用的距离有欧式距离[3]、余弦距离[4]和动态时间规整（DTW）[5]等。这种方法简单直接，但易受干扰，例如：对于存在缺失值的时序无法处理，对时序波形过于敏感，只能够表征时序的一些局部特性。

（2）基于建模的聚类方法先对时序进行统计建模，例如：自回归滑动平均（ARMA）[6]、隐马尔科夫模型（HMM）[7]等，然后在模型系数空间采用一定的方法进行聚类。这种方法能对不同长度的时序进行分析，提高聚类分析的鲁棒性；但每种模型背后存在较多严格假设，限制了其运用的范围。

（3）基于特征提取的聚类方法先提取相关特征以表示时序某种特性，然后在此特征空间进行时序聚类[8]。这种方法能够处理不同长度的时序，且无论时序长度都可将其压缩在一定维度的特征空间上，避免高维度聚类难题。所提取特征可表征时序的全局特性，从而避免局部特性的影响。通过提取多种类型的特征，可从不同角度描述时序，具有更广泛的运用范围。

无论在何种空间进行聚类，都需要一定的聚类算法，本领域常用的算法有K均值聚类[9]、层次聚类[10]等。K均值聚类算法简单，计算效率高，但需要事先指定类别个数；层次聚类可实现类别个数的自动选择，但算法复杂度较高，且算法收敛的条件也需要事先指定。

2 特征提取方法

我们将从周期性、趋势性、非高斯性和非线性4个角度描述时序特征，所提取的特征为季节性指标、趋势性指标、偏度、相对熵、样本熵、自相似性和李雅普诺夫系数。下面我们对这7个特征进行说明。

2.1 季节性与趋势性

假定一个序列的数学表示为：$X_t=\{x_1,x_2,\cdots,x_N\}$，为计算序列的季节性程度和趋势性程度，先采用季节-趋势分解（STL）[11]时序分解法将序列分解为3个分量：

$$X_t=T_t+S_t+E_t \tag{1}$$

T_t、S_t和E_t分别表示趋势性成分、季节性成分和随机性成分。完成时序分解后，季节性指标的相应计算公式为：

$$s_{\deg}=1-\frac{\mathrm{var}(E_t)}{\mathrm{var}(X_t-T_t)} \tag{2}$$

趋势性指标的计算公式为：

$$t_{\deg}=1-\frac{\mathrm{var}(E_t)}{\mathrm{var}(X_t-S_t)} \tag{3}$$

其中，var(·)表示求序列的方差。

2.2 偏度

偏度用来表征时序概率分布的拖尾（非对称）现象，对于正态分布，其偏度等于0，因此偏度可作为一种非高斯性的度量。随机变量X的偏度定义为：

$$skew(X)=E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \tag{4}$$

通过推导，可得偏度的简化计算方法为：

$$skew(X)=\frac{E\left[X^3\right]-3\mu\sigma^2-\mu^3}{\sigma^3} \tag{5}$$

其中μ和σ表示序列的均值和标准差，$E[\cdot]$表示求均值计算。

2.3 相对熵

相对熵是描述两个概率分布差异的一种方法[12]。设存在两个分布P和Q，P相对于Q的相对熵定义为：

$$D(P||Q)=\int P(x)\log\frac{P(x)}{Q(x)}dx \tag{6}$$

对于离散变量，有：

$$D(P||Q)=\sum P(i)\log\frac{P(i)}{Q(i)} \tag{7}$$

本文中，我们将序列X的相对熵定义为序列分布$P(X)$相对于正态分布$N(X)$的偏差，其中$N(X)$的均值和标准差等于序列的均值和标准差。我们对序列X做直方图，假设将取值范围分成m段，那么相对熵的计算方法可以为：

$$relative\_entropy=\sum_{i=1}^{m}P(i)\log\frac{P(i)}{N(i)} \tag{8}$$

其中，$P(i)$表示直方图第i段区间的概率，$N(i)$表示对应区间的正态分布累积概率。如此，文中的相对熵可作为序列非高斯性的一种度量。

2.4 样本熵

样本熵表征时间序列的复杂度，是一种非线性度量[13]。对于时间序列X_t，定义长度为m的模板向量为：

$$X_m(i)=\{x_i,x_{i+1},\cdots,x_{i+m-1}\},i=1,2,\cdots,N-m+1 \tag{9}$$

定义两个模板向量的距离为：

$$d[X_m(i),X_m(j)]=\max_{k=0,\cdots,m-1}\left\|x_{i+k}-x_{j+k}\right\|,i\neq j \tag{10}$$

则序列的样本熵计算为：

$$SampEn=-\log\frac{A}{B} \tag{11}$$

其中，B表示长度为m的模板向量对

距离小于某个阈值 r（$d\left[X_m(i),X_m(j)\right]<r$）的个数，$A$ 表示长度为 $m+1$ 的模板向量对距离小于阈值 $r(d[X_{m+1}(i),X_{m+1}(j)]<r)$ 的个数。根据一般经验，我们取 $m=2,r=0.2\times std$，std 表示序列标准差。

2.5 自相似性

自相似性表示序列的长期依赖性，是一种非线性度量。假设序列零均值后表示为：

$$X_t^{'}=X_t-mean(X_t) \tag{12}$$

令 $X_t^{'}$ 累积和序列为 Y_t，其第 i 个元素 y_i 表示为：

$$y_i=\sum_{k=1}^{i}x_i^{'} \tag{13}$$

计算 Y_t 取值范围 $R=\max(Y_t)-\min(Y_t)$，则采用如下 Hurst 指数[14]来表征该自相似性，其定义如下：

$$R/\sigma=(N/2)^K \tag{14}$$

其中，K 是 Hurst 指数，σ 是序列标准差，N 是序列长度，从而可得：

$$K=\frac{2}{N}\log(R/\sigma) \tag{15}$$

2.6 李雅普诺夫系数

李雅普诺夫指数表征序列的混沌性，也是一种非线性测量[15]。假定序列的某个以 i 下标作为起始点的子序列为：

$$X_i=\left\{x_i,x_{i+\tau},x_{i+2\tau},\cdots,x_{i+(m-1)\tau}\right\} \tag{16}$$

即该子序列的长度等于 m（嵌入维度），τ 表示延迟步数。设与该子序列欧拉距离最小的另一子序列为 X_j，则可求得子序列 X_{i+k} 和 X_{j+k} 的距离则为 $d_{i(k)}$。李雅普诺夫指数 λ 的定义为：

$$d_{i(k)}=d_{i(0)}e^{\lambda k} \tag{17}$$

即：

$$\lambda k+\log d_{i(0)}=\log d_{i(k)} \tag{18}$$

为计算 λ，对某个特定 k，可计算求得所有子序列 $d_{i(k)}$ 的平均值 $\bar{d}_{(k)}$。改变 k 可计算得到相应的 $\bar{d}_{(k)}$，然后再对 k 和 $\log\bar{d}_{(k)}$ 做线性拟合求其斜率即是 λ。一般经验取嵌入维度 m=10，延迟步数 τ 为序列自相关系数 R_τ 小于 $1-1/e$ 对应的值。

3 聚类方法

本文中，我们采用基于傅里叶变换的方法先将时序分为两大类，然后在各大类中采用 K 均值的方法进行聚类。

3.1 基于傅里叶变换的周期型分类

由于性能数据的形态特性种类繁多，直接进行聚类分析较为复杂。通过对大量的实际性能数据的观察，作者发现有些时序存在明显的周期形态，有些则不明显。因此我们提出先将时序按傅里叶变换的方法分为具有明显周期形态和不具有明显周期形态两大类，然后在各大类中采用基于结构特征的聚类方法进行更小类的细分，这种分等级的处理方法可以降低聚类复杂度。由于数据是离散序列，采用如下的离散傅里叶变换求取频率幅值谱：

$$|F[k]|=\left|\sum_{n=0}^{N-1}x(n)e^{-j\frac{2\pi}{N}kn}\right|,0\leqslant k\leqslant N-1 \tag{19}$$

然后求取幅值谱的最大值 $|F|_{\max}$，均值 $|F|_{mean}$ 和标准差 $|F|_{std}$。如果满足：

$$|F|_{\max}>|F|_{mean}+c\cdot|F|_{std} \tag{20}$$

其中，c（一般取不小于 3）是一个设定系数，当 $|F|_{\max}$ 对应的周期等于设定值（设定值一般为 1 天），则该时序具有明显周期形态。

3.2 K 均值聚类

由于 K 均值聚类算法简单直接，我们采用此方法对提取的结构特征向量进行聚类分析。设要将 n 个样本划分为 k 个聚类，k 均值聚类即是要确定这 k 类的中心（均值），使每个点离他最近的均值的距离和最小，即：

$$\arg\min_{\mu_i}=\sum_{i=1}^{k}\sum_{x_j\in S_i}\left\|x_j-\mu_i\right\|^2 \tag{21}$$

一般采用如下的算法流程完成上述优化问题：

（1）从数据集中随机取 k 个样本，作为 k 个簇的中心；

（2）分别计算各样本到 k 个簇中心的距离，将这些样本分别划归到与之距离最近的中心的簇；

（3）根据聚类结果，重新计算 k 个簇各自的中心；

（4）重复步骤 2 和 3，直至收敛；

（5）输出各类中心和各样本所属类别标签。

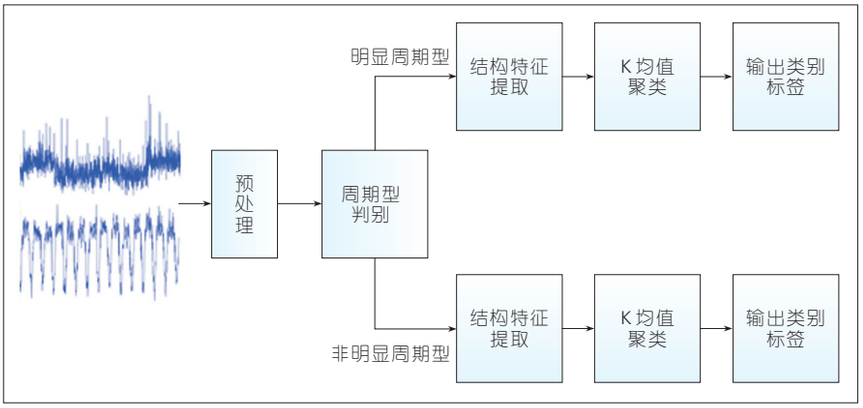
综合第 2 节和第 3 节，我们提出的聚类方案如图 1 所示。

4 仿真验证

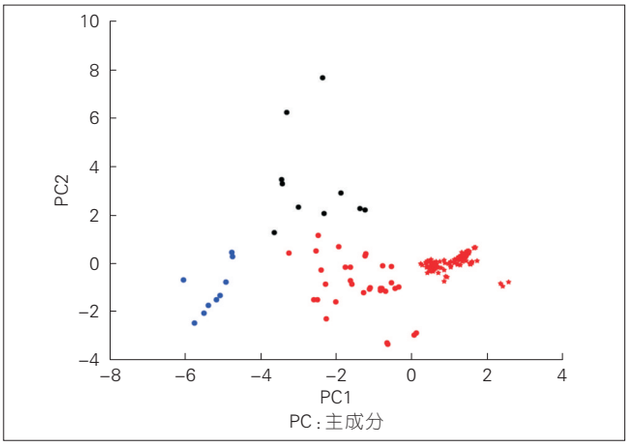
为验证本算法有效性，我们从实际商用云平台采集了 407 个网络端口流量数据，采集时间长度为 2 周，采样粒度为 15 min（即每天采集 96 个点）。在进行聚类分析之前，我们首先对每个时序进行去除极端噪声的预处理；然后采用基于傅里叶变换的方法进行周期型分类。仿真结果显示：407 个序列分成 165 个明显周期型序列和 242 个非明显周期型序列。

对于明显周期型序列，对每个序列提取完第 2 节所述 7 个结构特征，然后在此特征空间进行 K 均值聚类分析，仿真表明可将这些时序分为 4 类。图 2 是采用主成分分析（PCA）将特征样本点降维到 2 维平面的散点图，每种颜色或形状的散点表示其中一类时序。红色圆点的示例时序如图 3a）所示；蓝色圆点的示例时序如图 3b）所示；黑色圆点的示例时序如图 3c）所示；红色星点的示例时序如图 3d）所示。

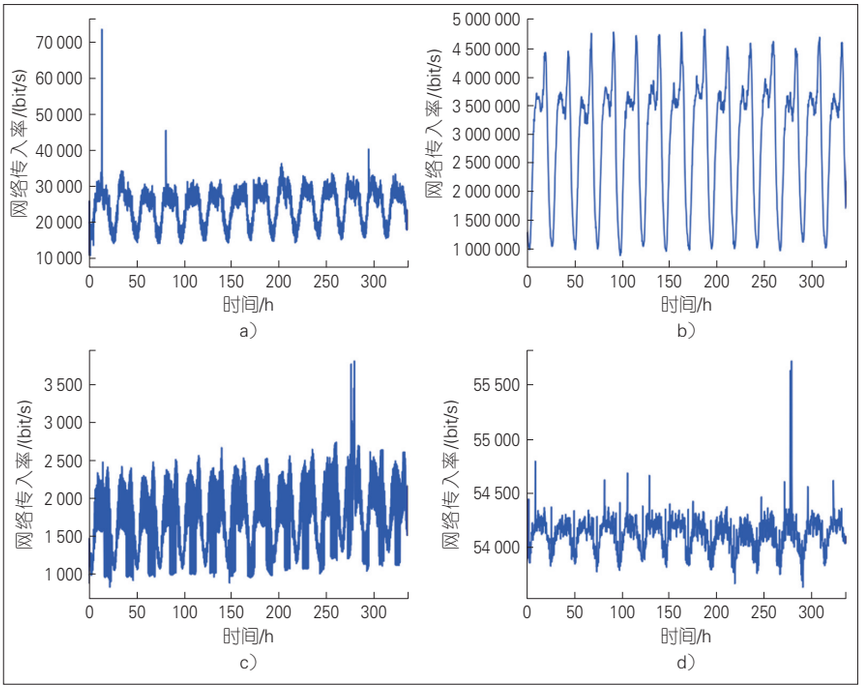
如上介绍，每种时序的代表性时序波形如图 3 所示。可见虽同样是



▲图1 时序聚类流程图



◀图2 周期型数据的特征空间散点示意图



▲图3 周期型数据4种类别时序波形示意图

有明显周期型的时序，其波形特性仍有差异，例如：右上角时序最为平滑规整，而左下角时序则带有显著的波动性。因此，其适用的时序建模方法将有所差别。

对于非明显周期型数据，同样对每个序列提取完第2节所述7个结构特征，然后在此特征空间进行K均值聚类分析，仿真表明可将其分为5类，图4是经过PCA降维后的特征散点示意图，每种颜色或形状的散点表示其中一类时序。红色圆点的示例时序如图5a）所示；蓝色圆点的示例时序如图5b）所示；黑色圆点的示例时序如图5c）所示；红色星点的示例时序如图5d）所示；蓝色星点的示例时序如图5e）所示。

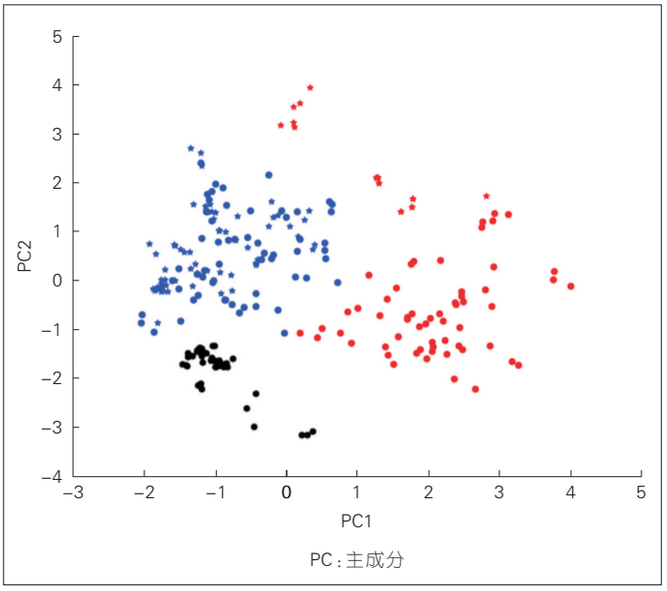
每种时序的代表性时序波形如图5所示，这些时序的特性也是各有变化，例如：图5a）时序带有大量的高脉冲，图5d）时序带有一定的趋势波动性，而图5e）时序较为平稳。

从图3和图5的仿真结果表明：本文所提的聚类方法能够将不同波形特性的时序分别开来，而这些不同特性的时序数据，可以预见有着不同适用的时序建模方法和动态阈值方法。表1、表2分别为非明显周期型数据和明显周期型数据各个子类别的时序数目。

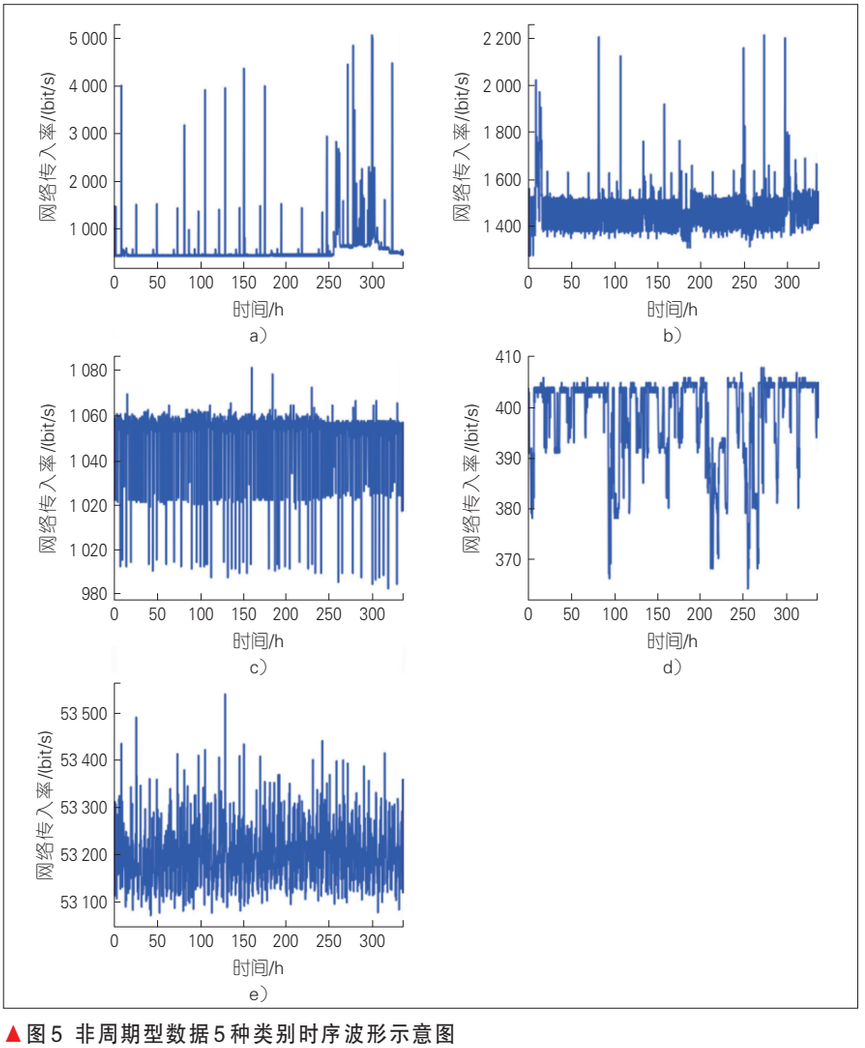
5 结束语

云平台产生大量的性能数据可用于系统运行状态的监测控制。本文提出一种基于结构特征的聚类方法对这些性能数据进行自动化分类。该方法采用分级处理的方式，首先将时序分为明显周期型和非周期型两大类，然后从各类时序提取7个结构特征，最后在此特征空间进行聚类分析。实验结果表明该方法能够将时序分为具有不同波形特性的数据。本文所提工作的主要创新和贡献主要如下：

（1）针对云平台大规模特性不一的数据，创新性地提出了一种基于结



◀图4 非周期型数据的特征空间散点示意图



▲图5 非周期型数据5种类别时序波形示意图

▼表1 非明显周期型数据各子类别的时序数目汇总

非周期型	时序数目
类别1	58
类别2	60
类别3	43
类别4	19
类别5	62

▼表2 明显周期型数据各子类别的时序数目汇总

周期型	时序数目
类别1	31
类别2	114
类别3	11
类别4	9

构特征的自动聚类方法，所提特征用于表征周期性、趋势性、非高斯性、非线性等多种时序结构。

（2）针对所采集的时序数据特点，创新性地提出了一种分级聚类框架，用于降低聚类复杂度。

该自动化时序分类结果为性能数据的进一步分析提供了基础，具有重要的应用价值。在作者所在的智能运维项目，已经着手根据每种类别时序的特点进行相应的建模和动态阈值预测。这方面内容超过本文的讨论范围，不再做细述。

参考文献
[1] MARVASTI M A, POGHOSYAN A V, HARUTYUNYAN A N, et al. An Enterprise Dynamic Thresholding System[C]//ICAC. USA: USENIX Association, 2014: 129–135

[2] LIAO T W. Clustering of Time Series Data—A Survey [J]. Pattern Recognition, 2005, 38 (11): 1857–1874
[3] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient Similarity Search in Sequence Databases [J]. Foundations of Data Organization and Algorithms, 1993: 69–84
[4] GOLAY X, KOLLIAS S, STOLL G, et al. A New Correlation–Based Fuzzy Logic Clustering Algorithm for FMRI [J]. Magnetic Resonance in Medicine, 1998, 40(2): 249–260
[5] RATANAMAHATANA C A, KEOGH E. Three Myths about Dynamic Time Warping Data Mining[C]//Proceedings of the 2005 SIAM International Conference on Data Mining. USA: Society for Industrial and Applied Mathematics, 2005: 506–510. DOI: 10.1137/1.9781611972757.50
[6] PICCOLO D. A Distance Measure for Classifying ARIMA Models [J]. Journal of Time Series Analysis, 1990, 11(2): 153–164
[7] LI C, BISWAS G. Temporal Pattern Generation Using Hidden Markov Model Based Unsupervised Classification [J]. Advances in Intelligent data analysis, 1999: 245–256
[8] WANG X, SMITH K, HYNDMAN R. Characteristic–Based Clustering for Time Series Data [J]. Data mining and knowledge Discovery, 2006, 13(3): 335–364
[9] HALKIDI M, BATISTAKIS Y, VAZIRGIANNIS M. On Clustering Validation Techniques [J]. Journal of intelligent information systems, 2001, 17(2): 107–145
[10] 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
[11] CLEVELAND R B, CLEVELAND W S, TERPENNING I. STL: A Seasonal–Trend Decomposition Procedure Based On Loess [J]. Journal of Official Statistics, 1990, 6(1): 3
[12] MARIAN P, MARIAN T A. Relative Entropy is An Exact Measure of Non–Gaussianity [J]. Physical Review A, 2013, 88(1): 012322
[13] RICHMAN J S, MOORMAN J R. Physiological Time–Series Analysis Using Approximate Entropy and Sample Entropy [J]. American Journal of Physiology–Heart and Circulatory Physiology, 2000, 278(6): H2039–H2049
[14] WERON R. ESTIMATING Long–Range Dependence: Finite Sample Properties and Confidence Intervals [J]. Physica A: Statistical Mechanics and its Applications, 2002, 312(1): 285–299
[15] ECKMANN J P, KAMPHORST S O, RUELLE D, et al. Liapunov Exponents from Time Series [J]. Physical Review A, 1986, 34(6): 4971

作者简介

孟志浩，中兴通讯股份有限公司虚拟化中心控制器研发总工；主要研究方向为移动边缘计算、智能运维。

刘建伟，中兴通讯股份有限公司虚拟化中心高级算法工程师；从事云平台智能运维相关工作，主要研究方向为机器学习、数据挖掘、信号处理和自动化控制；已发表论文14篇。

韩静，中兴通讯股份有限公司虚拟化中心智能运维总工；负责云平台智能运维总体规划、AI算法方向演进。