基于对称KL距离的用户行为时序聚类方法

李文璟 曾祥健* 李 梦 喻 鹏
(北京邮电大学网络与交换技术国家重点实验室 北京 100876)

摘 要：网络用户随时间变化的行为分析是近年来用户行为分析的热点，通常为了发现用户行为的特征需要对用户做聚类处理。针对用户时序数据的聚类问题，现有研究方法存在计算性能差，距离度量不准确的缺点，无法处理大规模数据。为了解决上述问题，该文提出基于对称KL距离的用户行为时序聚类方法。首先将时序数据转化为概率模型，从划分聚类的角度出发，在距离度量中引入KL距离，用以衡量不同用户间的时间分布差异。针对实网数据中数据规模大的特点，该方法在聚类的各个环节针对KL距离的特点做了优化，并证明了一种高效率的聚类质心求解办法。实验结果证明，该算法相比采用欧式距离和DTW距离度量的聚类算法能提高4%的准确度，与采用medoids聚类质心的聚类算法相比计算时间少了一个量级。采用该算法对实网环境中获取的用户流量数据处理证明了该算法拥有可行的应用价值。

关键词：时序聚类；用户分析；Kullback-Leibler距离
中图分类号：TN915.07 **文献标识码**：A **文章编号**：1009-5896(2018)10-2365-08
DOI: 10.11999/JEIT180016

Time Series Method Clustering in User Behavior Based on Symmetric Kullback-Leibler Distance

LI Wenjing ZENG Xiangjian LI Meng YU Peng
(*State Key Laboratory of Networking and Switching Technology*, *Beijing University of Posts and Telecommunications*, *Beijing* 100876, *China*)

Abstract: Behavioral analysis of Internet users over time is a hot spot in user behavior analysis in recent years, usually clustering users is a way to find the feature of user behavior. Problems like poor computing performance or inaccurate distance metric exist in present research about clustering user time series data, which is unable to deal with large scale data. To solve this problem, a method for clustering time series in user behavior is proposed based on symmetric Kullback-Leibler (KL) distance. First time series data is transformed into probability models, and then a distance metric named KL distance is introduce, using partition clustering method, the different time distribution between different users. For the Large-scale feature of physical network data, each process of clustering is optimized based on the characteristics of KL distance. It also proves an efficient solution for finding the clustering centroids. The experimental results show that this method can improve the accuracy of 4% compared with clustering algorithm using the Euclidean distance metric or DTW metric, and the calculation time of this method is less a quantity degree than clustering algorithm using medoids centroids. This method is used to deal with user traffic data obtained in physical network which proves its application value.

Key words: Time series clustering; User analysis; Kullback-Leibler distance

1 引言

通信技术与移动互联网的高速发展为基于流量监测的用户行为分析提供了前提。网络用户行为分析是指对用户上网数据进行统计、分析，从中发现网络用户行为的规律性[1]。通过用户行为分析，运营商可改善网络服务质量，防范网络攻击[2]；内容提供商可针对用户偏好提供个性化服务并提供社会舆情探测与分析等[3]。网络用户行为分析主要包括会话行为分析，上网喜好分析，社交网络分析和web访问行为分析等，而用户时间行为的模式挖掘

收稿日期：2018-01-04；改回日期：2018-06-27；网络出版：2018-07-30
*通信作者： 曾祥健 zeng_fsh@163.com
基金项目：国家电网公司科技项目(52010116000W)
Foundation Item: The Project of Science and Technology of State Grid Corporation of China (52010116000W)

是各种行为分析方法的基础，主要用于发掘用户网络行为的时间规律。已有学者针对用户时间行为开展了研究，文献[4]通过对微博用户做时序分析，发现微博用户访问微博的时间分布会影响用户在微博上传播信息的能力；文献[5]将用户行为的时间序列聚类应用于情感分析，通过对微博用户时序数据的聚类分析发现了不同群体的情感特征等，以上研究表明网络用户的时间行为分析与用户其他行为关联紧密，了解用户时间分布构成，是分析其他用户行为的基础。

然而对所有用户做时序行为分析十分困难，为了发现存在于不同用户间的时间分布规律，提取用户特征，通常需要先将用户做聚类处理，然后做进一步的分析。准确的聚类结果依赖于合理有效的距离度量方法，因此采用合适的距离度量十分重要。文献[6]在对用户的上网行为做分段处理的基础上，以欧氏距离作为聚类的距离度量方法研究宽带网络用户的上网行为。欧式距离度量因为其简单实用而被广泛应用，但是由于时间序列数据具有波动特性，如纵向缩放、线性漂移、时间漂移等，直接用欧氏距离度量相似性容易受局部波动影响，不能很好地从整体趋势的角度评价相似性。文献[7, 8]采用动态时间规整(Dynamic Time Warping, DTW)方法作为距离度量，DTW也是时序聚类常用的距离度量方法，适用于不定长序列，可以求出两时间序列的最大相似性。DTW方法属于动态规划算法，虽然准确率高，但是计算代价高。文献[9]采用KL距离度量，但其采用的围绕中心点划分(Partitioning Around Medoids, PAM)聚类方法存在质心计算复杂度高的问题，不适合数据量庞大的计算。同时，将采用上述度量方法的聚类算法应用于实网环境时，由于数据量巨大难以满足计算效率要求。

针对上述挑战，本文提出一种基于对称KL(Kullback-Leibler)距离的用户行为时序聚类方法，研究面向实网环境的网络用户时序数据的聚类问题。相比欧式距离和DTW，对称KL距离描述的是对象在概率分布上的差异，能适应数据的平移缩放等变形，摆脱传统距离定义在几何空间上的局限性，提高时间分布差异性描述的准确性。同时针对对称KL距离公式做简化，在传统算法中引入预计算索引的方法，达到快速聚类的目的。另外，本文在证明了概率约束下的聚类子集均值作为质心能与聚类子集内其他向量距离总和最小。最后通过两个实验验证了改进后的聚类算法在聚类准确度上提升了4%，并在计算速度上有了大幅度提升，与采用欧式距离度量的聚类算法相比同样也有很大优势，适合于大规模计算。

本文后续内容安排如下：第2节对用户时序概率模型进行建模；第3节提出基于KL距离的用户行为时序聚类算法；第4节分别基于仿真数据和实网流量数据对算法进行仿真验证，并对仿真结果进行分析讨论；最后总结了全文。

2 用户时序概率模型

2.1 时序概率模型表示

设用户流量行为数据集$S=\{T_i,\ i\in U\}$，U为用户数量，T为某一流量时间序列，序列中各点满足非负性，$T=\{t_0,\ \cdots,\ t_i,\ \cdots,\ t_N|\forall t_i>0,\ t_i\in T\}$，$N$表示序列长度。由于时间序列是高维数据，数据量往往十分庞大[10]，合理的表示将影响最终聚类的效果[11]。当N很大时虽然序列的信息被完整保留，但是计算量会十分庞大。假设将序列由N长度变换为M长度，定义N/M为时间粒度参数，在保证序列物理意义存在的前提下可以根据数据量大小选取该参数，转换方法参照PAA(Piecewise Aggregate Approximation)[12]。转换后的时间序列表示为

$$\begin{aligned}&\overline{T}=\{\bar{t}_1,\cdots,\bar{t}_i,\cdots,\bar{t}_M|\forall\bar{t}_i>0\},\\&\bar{t}_i=\frac{M}{N}\sum_{j=(N/M)(i-1)+1}^{(N/M)i}t_j\end{aligned}\tag{1}$$

$\overline{T}$的概率模型可以描述为

$$p(x)=\frac{\bar{t}_x}{\sum\bar{t}},\ \bar{t}_x\in\overline{T},\ x\in[0,M]\tag{2}$$

因此$\overline{T}$对应的离散概率集合为$P=\{p_1,\ \cdots,\ p_i,\ \cdots,\ p_M|\forall p_i>0,\ \sum_i^M p_i=1\}$。概率分布$p(x)$表示单个用户的流量概率分布模型，$P(x=i)$表示$i$时刻的概率，概率越高代表用户该时刻使用的流量越多。虽然离散概率分布允许概率为零的情况，但对称KL距离的运算不允许出现概率为零的情况，因此要对零概率进行填充。

综上对用户流量行为数据集S重新描述为$S=\{p_i,\ i\in U\}$，i为第i个用户。p_i为第i个用户的概率分布。

2.2 对称KL距离度量

KL距离，又称相对熵，是由Kullback和Leibler[13]提出的用于衡量两个概率分布之间差别的一种方法，由于满足非负性，KL距离可用于描述两组变量之间的距离，即KL距离可以描述近似分布Q接近真实分布P的程度[13]。KL距离关注的是两个概率分布的差异性，它支持连续概率分布和离散

概率分布的计算。

对于离散随机变量，概率分布P和Q的KL距离定义为

$$D_{\mathrm{KL}}(P||Q)=\sum_{i}^{N}p(i)\ln\frac{p(i)}{q(i)},\quad \forall p(i)>0,\ \forall q(i)>0 \tag{3}$$

由于KL距离遵循吉布斯不等式，因此$D_{\mathrm{KL}}(P||Q)\geq 0$，当且仅当$P$=$Q$时有$D_{\mathrm{KL}}(P||Q)=0$。因此对于距离度量来说，KL距离从物理上保持了“距离”含义的成立。但是由于KL距离是对数运算，所以存在非对称性，换言之，$D_{\mathrm{KL}}(P||Q)\neq D_{\mathrm{KL}}(Q||P)$。所以定义对称KL距离为$D(P,Q)=(D_{\mathrm{KL}}(P||Q)+D_{\mathrm{KL}}(Q||P))/2$。

综上对于两个用户的概率分布P, Q，重新定义它们的距离为

$$D(P,Q)=\sum_{i}^{M}\left(p_i\ln\frac{p_i}{q_i}+q_i\ln\frac{q_i}{p_i}\right)\Big/2 \tag{4}$$

其中，M为时序长度，P, Q分别对应不同用户的概率分布，D为两个用户之间的对称KL距离。

3　基于对称KL距离的用户行为时序聚类方法

3.1　确定聚类质心

聚类质心是指在聚类集合中找到一个向量$\bar{\boldsymbol{C}}$，使得$\bar{\boldsymbol{C}}$到该聚类集中其他向量的距离最小，聚类质心的确定是聚类方法的核心，聚类的效果高度依赖于$\bar{\boldsymbol{C}}$选取的质量[10]。

定义向量$\bar{\boldsymbol{C}}$, $\bar{\boldsymbol{C}}$是一个概率分布，又有

$$\bar{\boldsymbol{C}}=\arg\min_{\bar{\boldsymbol{C}}}\sum_{\boldsymbol{x}\in S_i}d(\boldsymbol{x},\bar{\boldsymbol{C}}) \tag{5}$$

其中，S_i表示第i个聚类子集，d是用户向量$\boldsymbol{x}$与向量$\bar{\boldsymbol{C}}$的距离。

目前一般采用medoids的思想用某个实际/向量代表聚类质心，但这样的方法计算量庞大，也有一些文献提出均值聚类质心表示法，但缺少相应场景下的证明[10]。本文证明了当聚类子集为概率分布的集合，并且质心也必须是一种概率分布时，集合的均值作为聚类子集质心将与子集内其他向量距离最小。

为了简化证明，用$d(\boldsymbol{P},\bar{\boldsymbol{C}})=\sum_{t}^{M}p_i\ln\frac{p_i}{\bar{c}_i}$代替式(5)来表示$\bar{\boldsymbol{C}}$与子集内其他向量间的距离。这样做以损失对称性为代价，但可以简化证明和减小计算量，另一方面采用这个距离计算，在质心与子集距离最小化的要求下，得到的质心将覆盖子集中的大概率部分，换言之，质心将体现出子集明显的特征。这是因为在采用$d(\boldsymbol{P},\bar{\boldsymbol{C}})=\sum_{t}^{M}p_i\ln\frac{p_i}{\bar{c}_i}$比较时，$p_i$大的部分，$\bar{c}_i$也必须大，否则差异会很大；而在$p_i$小的部分，无论$\bar{c}_i$取何值都不会与$\boldsymbol{P}$差异太大。假定$i$点$p_i$分别取值为0.9和0.1时在图1中展示该点上的距离d随$\bar{c}_i$的值变化的情况。

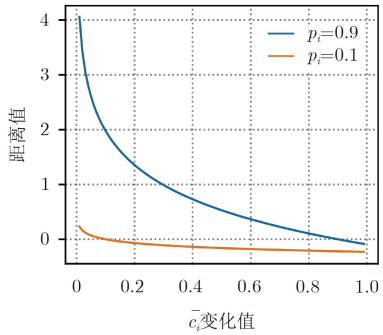


图 1 不同概率值上$\bar{c}_i$取不同值的距离差异

从图1中可以看出，当p_i为0.9时$\bar{c}_i$在远小于0.9的取值上距离最大，在接近0.9时距离最小，而在p_i=0.1下则几乎无变化。按照这个性质，用变形后的距离公式表示$\bar{\boldsymbol{C}}$与子集内其他点的距离，由于所求的$\bar{\boldsymbol{C}}$目标是与子集内所有向量距离最小，所以在$\boldsymbol{P}$中的大概率部分，相应的$\bar{c}_i$若取大概率相比取小概率距离会很多，因此所求出的$\bar{c}_i$能覆盖这部分子集共同的大概率部分，也就是时间段占用较为明显的部分，而不关心局部的小概率事件，因为这样会在总体上距离最小，能体现出子集的共同特征。所以当采用这种距离公式来表示$\bar{\boldsymbol{C}}$与子集内其他向量的距离时是合理且具有优势的。

对于$\bar{\boldsymbol{C}}$可以证明

$$\begin{aligned}\bar{\boldsymbol{C}}&=\arg\min_{\bar{\boldsymbol{C}}}\sum_{\boldsymbol{x}\in S_i}d(\boldsymbol{x},\bar{\boldsymbol{C}})\\&=\left\{p_1,\cdots,p_j,\cdots,p_M\middle|p_j=\frac{\sum_{\boldsymbol{x}\in S_i}x_j}{N_{S_i}},\ \forall p_j>0,\sum_{j}^{M}p_j=1\right\}\end{aligned} \tag{6}$$

其中，$\boldsymbol{x}$为聚类子集S_i的某个概率分布，$\boldsymbol{x}_j$是$\boldsymbol{x}$在j位置的概率，p_j是质心$\bar{\boldsymbol{C}}$在j位置的概率，$\{p_1,\cdots,p_j,\cdots,p_M\}$是构成$\bar{\boldsymbol{C}}$的集合，$p_j=\sum_{\boldsymbol{x}\in S_i}x_j\Big/N_{S_i}$表示$\bar{\boldsymbol{C}}$的概率分布，在3.3节中将采用此方法计算聚类质心。

3.2　算法加速

基于划分的聚类方法需要把数据集中的数据对

象按照距离最近的准则分配到最相似的类中。考虑某样本p和K个聚类中心，第k个聚类中心描述为$\boldsymbol{C}_k$，按照3.1节所描述，$\boldsymbol{C}_k$同样为概率分布，所以p和$\boldsymbol{C}_k$即两个概率分布，它们之间的距离由式(4)表示，式(4)展开为

$$D(\boldsymbol{P},\boldsymbol{C}_k)=\left(\sum_i^M p_i\cdot\ln p_i-p_i\cdot\ln c_i+c_i\cdot\ln c_i-c_i\cdot\ln p_i\right)\Big/2 \tag{7}$$

在K个中心的比较中从式(7)可以看出$p_i\cdot\ln p_i$与p有关，是一个常量，在距离比较中不起作用，可以省去。在计算过程中$\ln p_i$是不变的，在一次迭代中，$\ln c_i$是不变的，因此这两个运算可以通过预计算的方式建立索引。提高计算速度。具体到距离计算中，则用式(8)代替：

$$D(\boldsymbol{P},\boldsymbol{C}_k)=\sum_i^M(-p_i\cdot\ln c_i+c_i\cdot\ln c_i-c_i\cdot\ln p_i) \tag{8}$$

与原式(4)相比，乘除运算较少5次，而且由于不用进行对数运算，将减少一半计算时间。

3.3 聚类算法

综上，采用基于划分的聚类思想，算法包括以下步骤：(1)开始聚类算法前对数据集预先做对数运算并存储；(2)通过随机方式初始选出质心点；(3)开始一轮迭代；(4)按照3.2节的公式计算各个样本与质心的距离，分配到最近的聚类子集里；(5)所有样本遍历结束后，按照3.1节的方法求解每个聚类子集的质心；(6)由于KL距离不像欧氏距离定义在几何空间中，所以很难定义一个合适的阈值来停止迭代，因此采用判断集合数量的办法，当一个聚类子集的数量趋于平稳时可以认为聚类结果稳定，此时可以停止迭代；(7)输出结果，给出聚类子集。

算法过程如表1所示。

表 1 基于KL距离的用户行为时序聚类算法

输入：数据集data，类数目k，停止迭代参数threshold
输出：聚类结果cluster

```
cluster
old_cluster
    //初始化cluster, old_cluster
data_log = pre_computed(data)
    //预计算data的对数运算
k_centers = centers_init(data, k)
    //选择初始质心
for i in iteration:
  clear cluster;
  centers_log = pre_computed(k_centers)
      //预计算质心的对数运算
  for u in data:
    分配u到距离最近的cluster
  k_centers:= computed(cluster, data)
      //重新计算质心
  if i > 3 && overlap(cluster, old_cluster)>threshold:
        //子类重合度大于threshold时停止迭代
    break;
  else:
    old_cluster <- cluster
```

4 仿真结果与讨论

本文仿真分为两部分，第1部分采用人工合成数据，将本文方法与采用欧氏距离Kmeans的时序聚类方法、采用DTW的Kmeans时序聚类方法和采用KL距离的PAM时序聚类算法进行对比。通过这部分实验可以看出本文方法在聚类的准确性上比其他方法平均提高4%左右，在计算耗时上优于其他方法；第2部分采用校园网环境下获取的真实用户流量数据，将用户流量时序数据进行聚类，总共分为7类，每类都有自己的形态特征，聚类的结果可以为实网环境下的用户行为分析提供帮助。

4.1 人工数据仿真结果与讨论

人工数据采用基于cos函数和sin函数的变形数据，在[0, 2π]范围内取200个点，从5个方面对数据进行变形处理：(1)添加随机噪声；(2)平移；(3)幅值按比例缩放；(4)自变量按比例缩放；(5)添加线性误差，相应的变形公式为

$$y=\lambda(f(\alpha t+\beta)+nm(0,\sigma))+lr(t,\phi)+\gamma \tag{9}$$

其中，$f(t)$为cos或sin函数，$nm(0,\ \sigma)$为均值为0，标准差为σ的随机噪声，$lr(x,\ \phi)$为线性函数，形式为$y=\phi x$，γ为偏置系数，由于本文面向的数据是大于零的，这样转化的概率才有意义，所以拟合数据也要大于零。各系数的描述和取值如表2。每个变形参数有3个值，对cos函数和sin函数穷举所有参数的情况可以得到486个数据。人工合成数据总体情况如图2。

评价方法采用FM指数(Fowlkes–Mallows Index, FMI)[14]评价聚类准确性，FM指数定义为

$$\mathrm{FMI}=\frac{\mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})}} \tag{10}$$

假设有两类数据A1, A2, TP(True Positive)指

表 2　变形描述

变形形式	参数
随机噪声	高斯噪声，零均值，标准差σ=0.10, 0.12, 0.15
平移	横轴平移，β=0.10, 0.30, 0.50
幅值缩放	λ=0.50, 1.50, 2.00
自变量按比例缩放	时间轴缩放，α=1.05, 1.08, 1.10
线性误差	斜率ϕ=0.20, 0.50, 1.00
偏置系数	视情况而定，可以取为序列数据中最小值的绝对值，也可以取为常数，本文取γ=3.00

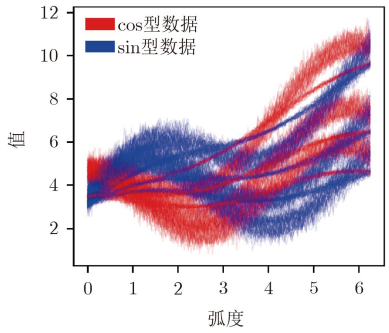


图 2 人工生成数据

真实类别与预测类别一致的数目，FP(False Positive)指实际为A2却预测为A1的数目，FN(False Negatives)指实际为A1却预测为A2的数目。FMI数值越高表示聚类效果与实际类别越接近，效果越好。

为了保证仿真结果的可比性，仿真中代入对比实验的数据都是经过2.1节时序概率模型转换后的数据，从数据的意义上来讲，对于本文的聚类算法是概率模型数据，对于其他对比算法则是流量占比数据，是对原数据的缩放，但这部分转换的计算时间不包括在仿真结果中。

聚类效果仿真结果如表3所示。

表 3　各算法聚类准确率和计算时间

聚类算法和度量方法	Kmeans+欧氏距离	Kmeans+DTW	PAM+KL	本文算法
FMI	0.863	0.856	0.852	0.892
时间(s)	0.099	187.589	0.336	0.046

由于表中所有聚类算法的初始聚类中心都是随机指定，因此聚类结果具有一定程度的随机性。上述实验结果是在多次实验的情况下取平均值。

可以看到相对欧式距离和DTW度量，KL距离在聚类的准确性上有4%左右的提升，且本文算法能够大大减少计算时间，相较欧式距离仍有很大优势，说明本文算法可以用在大规模运算中。DTW由于是关于序列长度平方级别的运算，所以耗费时间长。另一方面DTW和欧式距离在比较序列不同点距离时采用几何空间距离定义，且适合不定长数据，所以在准确性上表现不出优势，对于空间中距离较为接近但分布明显不同的两类时序数据无法较好地区分，而KL距离描述的是概率分布的差异，仿真数据来源于正弦数据和余弦数据，在一个周期内它们具有不同的分布，因此KL距离能在聚类结果上占有优势。而与同样采用KL距离但聚类算法采用PAM的算法相比较，本文算法准确性较高，PAM在计算聚类质心时的时间规模为$O(N^2)$，本文算法计算聚类质心的时间规模则是$O(N)$，因此计算时间上有非常大的优势。人工数据的聚类结果表明本文算法在合理性、准确性和计算效率方面，与现有方法相比都有很大优势，接下来将分析本文算法在真实数据上的应用。

4.2　实际数据仿真结果与讨论

本文采用的实际数据是部署在校园网上服务器所获取的数据，服务器采用旁路部署方式接入校园网某汇聚层路由器，实时获取并保存所有流经该汇聚层路由器的流量信息，并以流记录的形式保留下来，一条流记录是源点与终点之间交互的一系列五元组的统计信息，包括连接时长、传输流量大小等信息。

本文采用2017年7月20号至2017年7月21号之间的流记录数据，本份数据的统计指标如表4所示。

为了从数据中恢复出用户流量时序数据，对上述数据做以下操作。

(1) 首先对数据做预处理，主要包括：(a)去除非内网IP数据；(b)依照目的地址数值化用户ID；(c)时间精度转换为分钟级别；(d)传输速率缩放到分钟级别；(e)缺失值填充。

(2) 按照算法处理数据，得到用户流量数据，

表 4 2017年7月20号至2017年7月21号之间的流记录统计数据

统计指标	数值
流记录规模(10^4条)	270
去重源IP数目	203356
去重目的IP数目	113728
单条最长的流记录时间(ms)	50315672
单条最短的流记录时间(ms)	<1
平均流记录时间(ms)	11586.79
单条流记录最高流量总字节(GByte)	2.126
单条流记录最低流量总字节(Byte)	60
流记录平均流量总字节(Byte)	55945.96
单条流记录最高流量速率(MByte/s)	69.5
单条流记录最低流量速率(Byte/s)	<1
单条流记录平均流量速率(Byte/s)	50866.89

算法时间规模在$O(N+MT)$，其中N是原始数据量，M是用户数，T是时序长度，且T=1440/L，L是时间粒度精度参数，取L=60。

(3) 每个用户流量时序数据按照2.1节的时序概率模型将时序数据转化为概率分布，转换算法如表5所示。

表 5 原始流记录转换为概率模型算法

输入：流记录集D，时间粒度参数L，用户数量M
输出：用户流量概率模型 users
初始化数组 users[M][T], users_tmp[M][T]
for r in D:
  userid:=r对应的用户id
  beginTime:=r开始的时间
  endTime:=r结束的时间
  speed:=r的平均传输速率
  users[userid][beginTime] += speed
  users_tmp[userid][endTime] += speed
for i=0, i<M; i++:
  for j=1; j<L; j++:
    users[i][j]=users[i][j−1]-users_tmp[i][j−1]+users[i][j]
    if users[i][j]==0: //KL距离是对数计算，需要对序列进行平滑
      users[i][j]=1

为了达到更好的聚类效果，首先要选出合适的聚类数K，本文采用轮廓系数(silhouette coefficient)[15]评价聚类效果，对于某个样本i的轮廓系数定义为

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \tag{11}$$

其中，$a(i)$定义为该样本到子集内其他样本的平均距离，$b(i)$定义为样本到其他子集内所有样本平均距离的最小值。所有样本轮廓系数的均值用于评价聚类整体效果的好坏，轮廓系数越接近1聚类效果越好。由于本文采用随机的方法选择初始聚类中心，聚类结果存在不稳定性，因此本文对每一个K值做10次运算，去掉最高和最低值后取平均值作为结果。本文面向的是网络用户群体，在聚类数目上通常会根据实际场景选择需要的数目，这里考虑以尽可能小的聚类数目体现较大的群体特征，文献[6]同样是关于网络用户时间偏好的研究，它采用层次聚类的方法将聚类数目由24最终缩小至6，所以本文将K的范围设为2～10，得到的K值轮廓系数变化图如图3。

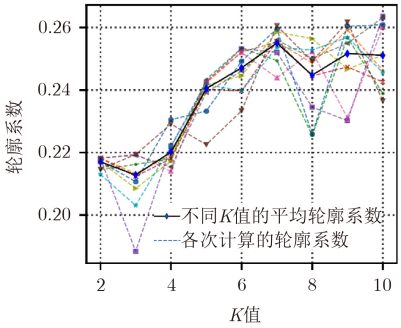


图 3 不同K值的轮廓系数情况

黑实线是取平均的结果，其他虚线是各次计算的结果，可以看到当K=7时效果最好，因此选择K=7作为聚类数目。图4是按照3.3节的聚类算法做的聚类结果，图中每条线代表一类的质心，反映的是这个类别在时间分布上呈现的趋势和时间分布情况。

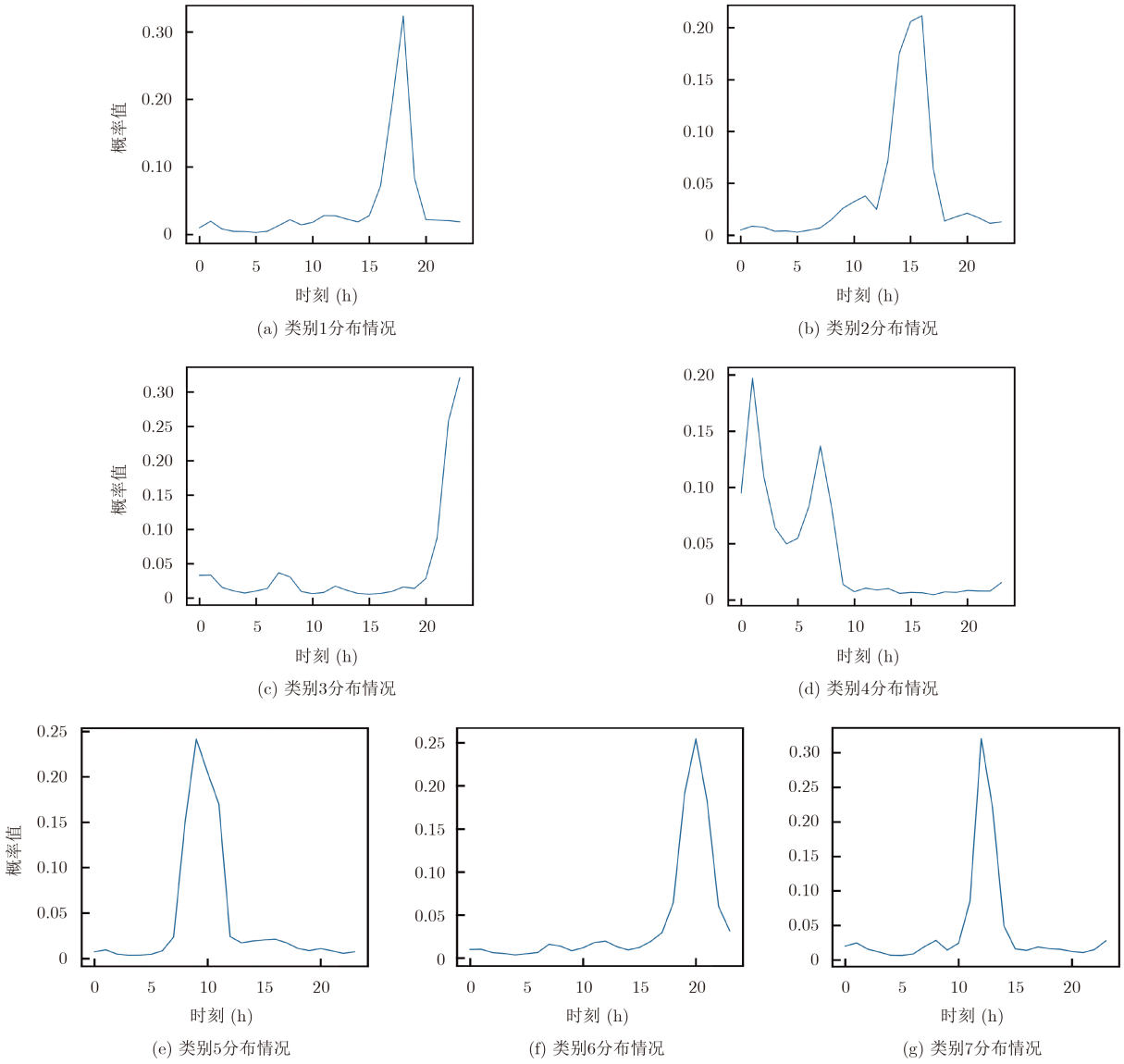
针对上述聚类结果，表6中对每一类人群的时间偏好和分布做了分析，分析的结果可以为群体针对性营销或其他用户行为分析提供帮助。

为了描述不同类别整体的分布情况，我们把各个类别人数的占比与类别中各个时段的概率相乘后，把所有类别合并一起展示成堆叠柱状图，如图5所示，各个时段总和为1，每一种类别用一种颜色表示，高度代表该时刻该类别在所有类别中的比重，从图中可以反映出在各个时段，不同类别用户的分布情况，基于该分布特性，可以针对不同类别的人群做针对性的策略处理。

5 结束语

网络用户的时间行为挖掘对分析网络用户行为有重大意义，准确的时间行为分析是其他行为分析的基础。本文提出的基于对称KL距离的用户行为

时序聚类方法，用概率模型的分布差异描述用户时间行为的差别，摆脱几何空间定义下的距离度量方法的局限性，能适应时序数据的多种变形，同时针对对称KL距离的聚类流程优化，具有计算量小，快速聚类，准确度高的特点，能应用于实际数据的处理中，有较好的应用前景。



(a) 类别1分布情况　(b) 类别2分布情况

(c) 类别3分布情况　(d) 类别4分布情况

(e) 类别5分布情况　(f) 类别6分布情况　(g) 类别7分布情况

图 4 聚类结果

表 6　聚类结果分析

类别	人数(人)	分析
1	2464	主要集中在下午时段使用，占比达到30%，并且持续时间从下午到晚上。
2	2713	这部分用户在15点附近使用幅度最高，总体对网络依赖性较高。
3	4862	所有类别中人数最多的类，主要使用时段是20点之后，并且逐渐升高，使用时段集中。
4	2444	主要在深夜和早上使用网络，其他时段使用较为平均。
5	3250	主要集中在早上使用，使用时间跨度较大。
6	3778	从晚上开始逐渐升高使用量，在20点左右达到最高，这一类人可能是在下班后使用网络。
7	2725	跨度较小，主要集中在中午时段使用。

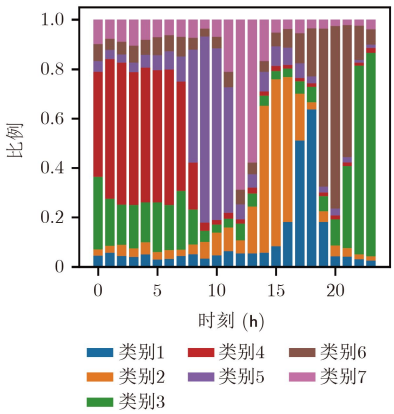


图 5 用户的时间行为整体分布情况

参 考 文 献

[1] 延皓. 基于流量监测的网络用户行为分析[D]. [博士论文], 北京邮电大学, 2011.
YAN H. Network user behavior analysis base on traffic monitoring and measurement[D]. [Ph.D. dissertation], Beijing University of Post and Telecommunications, 2011.

[2] NAJAFABADI M M, KHOSHGOFTAAR T M, CALVERT C, *et al*. User behavior anomaly detection for application layer DDoS attacks[C]. 2017 IEEE International Conference on Information Reuse and Integration (IRI), San Diego, USA, 2017: 154–161. doi: 10.1109/IRI.2017.44.

[3] 方志祥, 于冲, 张韬, 等. 手机用户上网时段的混合Markov预测方法[J]. 地球信息科学学报, 2017, 19(8): 1019–1025. doi: 10.3724/SP.J.1047.2017.01019.
FANG Zhixiang, YU Chong, ZHANG Tao, *et al*. A mixed arkov method to predict the surfing time period of mobile phone users[J]. *Journal of Geo-Information Science*, 2017, 19(8): 1019–1025. doi: 10.3724/SP.J.1047.2017.01019.

[4] 毛佳昕, 刘奕群, 张敏, 等. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014, 37(4): 791–800. doi: 10.3724/SP.J.1016.2014.00791.
MAO Jiaxin, LIU Yiqun, ZHANG Min, *et al*. Social influence anal sis for micro-blog user based on user behavior[J]. *Chinese Journal of Computers*, 2014, 37(4): 791–800. doi: 10.3724/SP.J.1016.2014.00791.

[5] ZHU Jiang, WANG Baixuan, and WU Bin. Social network users clustering based on multivariate time series of emotional behavior[J]. *Journal of China Universities of Posts and Telecommunications*, 2014, 21(2): 21–31. doi: 10.1016/S1005-8885(14)60282-X.

[6] YAN Hao, DOU Yinan, LIU Fang, *et al*. Time division based on analyses of network user time span preference[C]. 2009 IEEE International Conference on Network Infrastructure and Digital Content, Beijing, China, 2009: 177–181. doi: 10.1109/ICNIDC.2009.5360868.

[7] SALGADO C M, FERREIRA M C, and VIEIRA S M. Mixed fuzzy clustering for misaligned time series[J]. *IEEE Transactions on Fuzzy Systems*, 2017, 25(6): 1777–1794. doi: 10.1109/TFUZZ.2016.2633375.

[8] TEERARATKUL T, NEILL D O, and LALL S. Shape-based approach to household electric load curve clustering and prediction[J]. *IEEE Transactions on Smart Grid*, 2017. doi: 10.1109/TSG.2017.2683461.

[9] GHASSEMPOUR S, GIROSI F, and MAEDER A. Clustering multivariate time series using hidden markov models[J]. *International Journal of Environmental Research and Public Health*, 2014, 11(3): 2741–2763. doi: 10.3390/ijerph110302741.

[10] AGHABOZORGI S, SHIRKHORSHIDI A S, and WAH T Y. Time-series clustering — A decade review[J]. *Information Systems*, 2015, 53: 16–38. doi: 10.1016/j.is.2015.04.007.

[11] RATANAMAHATANA C, KEOGH E, BAGNALL A J, *et al*. A Novel Bit level time series representation with implication of similarity search and clustering[C]. 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Hanoi, 2005: 771–777. doi: 10.1007/11430919_90.

[12] KEOGH E J and PAZZANI M J. A simple dimensionality reduction technique for fast similarity search in large time series databases[C]. 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Kyoto, Japan, 2000: 122–133. doi: 10.1007/3-540-45571-X_14.

[13] KULLBACK S and LEIBLER R A. On information and sufficiency[J]. *The Annals of Mathematical Statistics*, 1951, 22(1): 79–86. doi: 10.1214/aoms/1177729694.

[14] FOWLKES E B and MALLOWS C L. A method for comparing two hierarchical clusterings[J]. *Journal of the American Statistical Association*, 1983, 78(383): 553–569. doi: 10.1080/01621459.1983.10478008.

[15] ROUSSEEUW P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. *Journal of Computational and Applied Mathematics*, 1986, 20(1): 53–65. doi: 10.1016/0377-0427(87)90125.

李文璟：女，1973年生，教授，研究方向为网络管理与通信软件、未来网络智能管理.

曾祥健：男，1993年生，硕士生，研究方向为网络管理与智能信息处理.

李 梦：女，1993年生，硕士生，研究方向为网络管理与智能信息处理.

喻 鹏：男，1986年生，副教授，研究方向为基于人工智能的网络管理.