HMM-Cluster:面向交通量过载发现的轨迹聚类方法

潘 立[1,2],邓 佳[1],王永利[1]
PAN Li[1,2], DENG Jia[1], WANG Yongli[1]

1.南京理工大学 计算机科学与工程学院,南京 210094
2.中国人民解放军火箭军参谋部,北京 100085
1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2. Staff of PLA Rocket Force, Beijing 100085, China

PAN Li, DENG Jia, WANG Yongli. HMM-Cluster: Trajectory clustering for discovering traffic volume overload. Computer Engineering and Applications, 2018, 54(1):77-85.

Abstract: With the development of economy, the urban traffic congestion has become an urgent problem in China. The traffic volume overload discovering is an effective method for solving the problem of traffic congestion. A kind of trajectory clustering method based on the HMM model, named HMM-Cluster, is put forward, which can find out the traffic volume overload conditions. HMM-Cluster extracts the feature points of spatio-temporal trajectory data firstly, and it uses dimension reduction technique to decrease the trajectory data volume, as well as save the cost of storage. Secondly, it trains a HMM model for each reference trajectory based on density function to get a trajectory affinity similarity matrix. Finally, the HMM-Cluster algorithm aggregates similarity trajectory effectively and forms the clustering results of trajectory data. The contrast experiments on actual data prove that the HMM-Cluster method has a good effect, which can obtain moving objects' pattern and discover traffic volume overload effectively and conveniently. The proposed method has significant values in real application.

Key words: traffic volume overload; spatio-temporal data; trajectory clustering; Hidden Markov Model(HMM)

摘 要:随着经济的发展,城市交通拥堵问题亟待解决,交通量过载发现是解决交通拥堵问题的有效方法之一。提出一种基于HMM模型的轨迹聚类算法HMM-Cluster,可有效地发现交通量过载情况。该算法首先提取时空轨迹特征点,并采用维数约简技术减少轨迹数据量,根据参照轨迹拟合HMM模型,基于密度函数得到轨迹相似度矩阵,最后给出聚合的相似性轨迹。真实轨迹数据集上的对比实验结果表明,提出的HMM-Cluster可有效地挖掘移动对象运动模式,准确发现交通量过载情况,具有一定实用价值。

关键词:交通量过载;时空数据;轨迹聚类;隐马尔科夫模型
文献标志码:A **中图分类号:**TP311 doi:10.3778/j.issn.1002-8331.1612-0528

1 引言

近年来,对于时空轨迹数据的研究持续升温,在国内外获得了广泛的关注。相关定位设备可记录包含时间戳和经度、纬度位置信息的轨迹数据,即一系列带有时间戳信息的位置数据(二维点)的有序集合,也称为时空轨迹数据,时空轨迹描述了移动物体位置随时间产生变化的情况[1]。由于交通量是一个聚集数(aggregate number)并且在持续的变化,不同的路之间的交通量还会互相影响,这些因素使得有效地描述交通量变化趋势,对交通量过载的发现成为颇具挑战的问题。

基金项目:国家自然科学基金(No.61170035);"江苏省六大人才高峰"高层次人才项目(No.WLW-004);中央高校基本科研业务费专项资金(No.30916011328);江苏省科技成果转化专项资金(No.BA2013047)。

作者简介:潘立(1977—),男,硕士研究生,主要研究领域为海量数据分析、物联网数据处理、数据挖掘等;邓佳(1990—),女,硕士研究生,主要研究领域为大数据分析、交通数据流分析、位置服务;王永利(1974—),男,博士,教授,主要研究领域为数据库技术、情境感知、物联网数据处理、模式识别等,E-mail:yongliwang@njust.edu.cn。

收稿日期:2017-01-03 **修回日期:**2017-05-12 **文章编号:**1002-8331(2018)01-0077-09
CNKI**网络优先出版:**2017-09-21, http://kns.cnki.net/kcms/detail/11.2127.TP.20170921.1559.010.html

对某时刻的轨迹位置进行聚类,可得到该时刻交通路网中移动对象的相对密集区域和这些区域的分布情况,即该时刻的交通量过载区域;对某段时间的对象轨迹进行聚类,可发现交通路网中的拥堵路段和这些路段的分布情况,即可获得该时段的交通量过载区域。进行轨迹聚类的目的在于,分析轨迹的特征属性,聚集在某时刻具有相似性的轨迹,获取移动对象的行为模式,通过聚类结果,对交通量过载区域进行发现。

本文提出一种针对移动对象轨迹数据的聚类方法HMM-Cluster,主要工作如下:(1)利用隐马尔科夫模型(HMM)抽取移动对象轨迹的隐含状态以及状态之间的相关信息,并运用AIC(Akaike Information Criterion)信息量准则度量模型的最佳状态;(2)提出一种轨迹相似性度量方法Sim-HMM,基于BP(Bicego-Panuccio)距离计算各轨迹在模型下的极大似然概率,根据相似度度量函数,构建轨迹相似度矩阵;(3)在此基础上,结合相似度矩阵降维及模糊C均值方法完成轨迹聚类,与已有方法相比,HMM-Cluster提升了聚类结果的$F1$值,可准确发现区域交通量过载。

2 相关工作

目前有两种实现移动轨迹数据聚类的算法[2]:一是将轨迹数据聚类转化成传统的点的聚类,使用相似度计算方法,其缺陷是相似度计算的过程难度相当高,一旦选择的方法不合适,可能会导致得到轨迹没有意义的特征而忽略轨迹重要的信息;二是基于模型进行聚类,假设数据集由指定模型生成,估计模型的相关参数,从而实现移动轨迹数据的聚类分析。

实现移动对象轨迹数据聚类的关键问题是计算轨迹之间的相似性,已有较为常见的相似度度量方法有最长公共子序列(Longest Common Sub-Sequence,LCSS)[3]、实序列编辑距离(Edit Distance on Real-sequence,EDR)[4]、约束的离散邻距离(w-constrained Discrete Fréchet distance,wDF)[5]、动态时间规整(Dynamic Time Warping,DTW)[6]、欧式几何距离(Euclidean Distance)[7]、豪斯多夫距离(Hausdorff Distance)[8]等。 Euclidean较为简单,但是对数据噪声很敏感,当前多采用其改进方法。DTW和LCSS方法适用于轨迹长度不等的情况,但是时间复杂度很高,使用受到限制。Hausdroff距离运用到方向相反的轨迹时,得到的效果不是很理想,文献[9]在Hausdorff距离基础之上加入轨迹点速度测量。

Wei等[10]提出了多项式拟合轨迹数据的方法。Lee等[11]针对轨迹数据的聚类提出了一种“划分-聚合”框架——TRACLUS,首先使用最小描述长度(Minimum Descriptive Length,MDL)来对原始轨迹进行划分,这些被划分的片段被称为子轨迹,然后使用DBSCAN算法[12]对子轨迹进行聚类,最后使用具有代表性的轨迹表示聚类簇。Gudmundsson等[13]提出一种融合了数据挖掘、计算几何和字符串处理技术的轨迹数据聚类方法。Lee等在文献[14]中拓展了文献[15]中的聚类方法进行轨迹分类。目前应用比较广泛,聚类效果比较好的一些传统聚类方法还有层次聚类(Hierarchical Clustering)、k-均值聚类(k-means)等[16-17]。现有方法多是基于距离的轨迹距离,无法适应动态的轨迹长度不确定的实时分析场景。

通常,基于模型进行聚类的方法不考虑轨迹本身的长度不同带来的限制,将轨迹转化成模型进行描述,可以适应轨迹的特殊属性。而在这类方法中,隐马尔科夫模型(Hidden Markov Model,HMM)[18]能够统计分析轨迹的隐含和观察状态,以及它们之间的相关信息,还能统计轨迹的位置转换概率,目前在轨迹序列的聚类中应用越来越广泛。本文提出一种基于HMM模型的高效轨迹聚类方法,以弥补现有轨迹聚类方法存在的精度低,适应性差低问题。

3 轨迹特征点提取

由于交通轨迹受到路网的约束,与一般自由运动的移动轨迹有所差别。首先由于轨迹受到路网的空间限制,包括交叉口的汇聚和分流,移动对象只能在公路马路等特定区域运动等,需要将这些因素的影响加以考虑;其次,轨迹在大部分路段运动时,若无障碍物,轨迹近似直线,空间特征是确定的,除非运动速度和方向有较大变化,其他细节对分析结果不会产生太大影响,可以不予考虑。

本文提出一种在采样轨迹基础上提取轨迹特征点的方法,使用较少的离散点保留较多的轨迹时空特征。

以如图1所示的场景为例,本文的特征点提取方法以X-Y二维空间为参照,将一段直行道路标志为A-E,假设途中经过公园P,分三类提取特征点:

(1)“路段交叉口”特征点。按时间顺序记录移动对象分别经过某路段两个交汇点的信息,将该点作为特征点之一,如图中点A、B、C、D、E。

(2)“区域范围”特征点。将移动对象运动特征相似的范围,如某段直行道路、公园类休闲娱乐区域、高速公路等,这些区域范围内没有交叉口,且不会对移动对象轨迹整体运动模式的分析产生巨大作用,可以忽略移动对象在其中的运动细节,而将它浓缩为一个特征点,如上图将公园内轨迹用点P表示。

(3)“运动速度或方向变化较大”特征点。在同一路段,速度变化较大的点或者出现转向等情况,对于轨迹运动模式的发现也有重要作用,需要作为特征点提取考虑。移动对象在C、D之间运动时,属于同一路段,但是S点速度变化较大,故将其作为特征点保留下来。

该方法在保留轨迹时空特征的基础上降低存储空

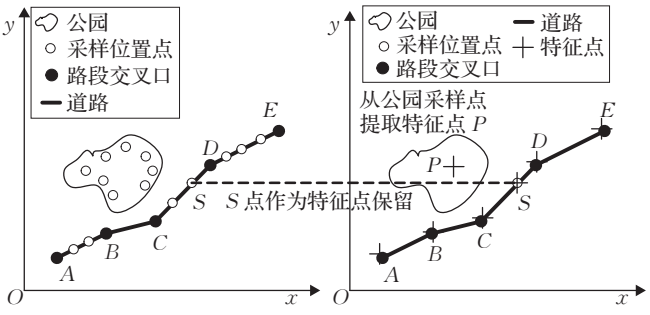


图1　轨迹特征点提取方法图

间的需求，由此得到的轨迹也是下文进行轨迹聚类时相似度计算的基础单位。

4　基于HMM的轨迹聚类方法

4.1　方法框架

本文采用HMM模型描述轨迹数据蕴含的隐藏状态。令 λ 表示HMM模型的参数集合值，轨迹特征HMM模型可表示为五元组参数：$\lambda=(X,Y,\Pi,A,B)$。其中 $X=\{x_1,x_2,\cdots,x_n\}$ 表示隐状态；$Y=\{y_1,y_2,\cdots,y_n\}$ 表示隐状态观测序列（显状态）；$\Pi=\{\pi_j\},j\in X$ 表示初始概率（隐状态）；$A(N\times N)=\{a_{ij}\}$ 表示状态转移矩阵，$i,j\in X$，$a_{ij}=P(x_j|x_i)$，是指序列由隐式状态 x_i 转换到隐式状态 x_j 的概率；$B(N\times M)=\{b_{jk}\}$ 表示观察概率矩阵，$j\in X$，$k\in Y$ $b_{jk}=P(y_k|x_j)$，是指隐式状态 x_j 转换为显式状态 y_k 的概率。本文重点解决HMM评估、解码、学习3个基本问题。

由于对象的轨迹长度、运动速度、时空分辨率差异较大等因素的影响，通过欧式距离、LCSS、Hausdorff距离等简单距离计算方法得到轨迹的相似度难度较大，用于实际分析的可行性不大。本文考虑通过HMM概率模型表示移动对象的轨迹分布模式，将轨迹序列之间的相似度距离计算问题转化成轨迹序列在概率模型下的产生概率的计算问题。基于HMM概率模型的轨迹聚类方法框架如图2所示。

首先，参照Baum-Welch算法为每条轨迹拟合隐马尔可夫模型，并运用AIC信息量准则（Akaike Information Criterion）度量模型的最佳状态；其次使用本文提出的轨迹相似度度量方法——Sim-HMM方法，包含计算各轨迹在模型下的极大似然概率、BP距离，相似度度量函数的提出，得到轨迹相似度矩阵等；然后进行PCA矩阵降维处理；在此基础上使用FCM（Fuzzy C Mean，模糊C均值）进行轨迹特征聚类。该方法简称为HMM-Cluster。

4.2　训练HMM模型

定义1　参考轨迹序列：对给定的轨迹集 $T=\{t_1,t_2,\cdots,t_n\}$，可从其 n 条轨迹中随机地选取 r 条轨迹作为参考序列 $R=\{R_1,R_2,\cdots,R_r\}$，其中 $R\subset T,1\leqslant r\leqslant n$。

对参考轨迹序列 R 中的每一条轨迹 $R_i(i<r)$ 拟合对应的HMM模型 λ_i，得到大小为r的隐马尔可夫模型集 $HMMs=\{\lambda_1,\lambda_2,\cdots,\lambda_t\}$。于是，建模的时间复杂度从 $O(n)$ 降到 $O(r)$。

对轨迹拟合训练HMM模型的过程是一个估计模型参数的过程，不断地计算交通轨迹序列在模型下的后验概率，不断地调整模型的参数，使给出的交通轨迹序列与模型的匹配度最高。

本文使用Baum-Welch算法为观测序列拟合一个最合适的HMM，首先初步估计隐马尔可夫模型的参数，然后通过使用给定轨迹数据对参数的价值进行评估，重新修改这些参数，减少错误；以梯度下降的形式寻找最小错误测度的方式，确定最合适的 $\lambda=(\Pi,A,B)$ 三元组来描述已知序列。

给定观察序列 $Y=(y_1,y_2,\cdots,y_T)$，其中 y_i 表示轨迹序列在 i 时刻的位置，T 表示序列长度。调整HMM的参数 $\lambda_0=(\Pi,A,B)$，使得观测序列在模型下的条件概率 $P(Y|\lambda)$ 最大，即两者匹配度最高。使用式(1)和式(2)进行模型参数重估，得到 $\lambda'=(\Pi',A',B')$：

$$a'_{ij}=\frac{\sum_{t=1}^{T-1}P(q(t)=x_i,q(t+1)=x_j|O,\lambda)}{\sum_{t=1}^{T-1}P(q(t)=x_i|O,\lambda)}=\frac{\sum_{t=1}^{T-1}\xi_{ij}(t)}{\sum_{t=1}^{T-1}\delta_i(t)} \tag{1}$$

$$b'_{ik}=\frac{\sum_{t=1}^{T}{}_{\{y(t)=k\}}P(q(t)=x_i|O,\lambda)}{\sum_{t=1}^{T}P(q(t)=x_i|O,\lambda)}=\frac{\sum_{t=1}^{T-1}{}_{\{y(t)=k\}}\delta_t(i)}{\sum_{t=1}^{T-1}\delta_t(i)} \tag{2}$$

其中，变量 $\delta_i(t)$ 表示在上述给定条件下，t 时刻位于隐藏状态 x_i 的概率；变量 $\xi_{ij}(t)$ 表示在上述给定条件下，t 时刻位于隐藏状态 x_i 且 $t+1$ 时刻位于隐藏状态 x_j 的概率。在重估模型之前，首先通过前向-后向算法计算 $\delta_i(t)$ 和 $\xi_{ij}(t)$ 的值，定义 $\alpha_j(t)$ 和 $\beta_i(t)$，如式(3)、(4)。

$$\alpha_j(t)=P(y_1,y_2,\cdots,y_t,q(t)=x_j|\lambda) \tag{3}$$

$$\beta_i(t)=P(y_{t+1},y_{t+2},\cdots,y_T|q(t)=x_i,\lambda) \tag{4}$$

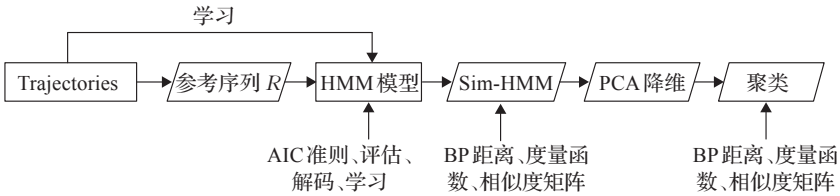


图2　HMM-Cluster方法框架

其中，$\alpha_j(t)$表示，在上述给定条件下，轨迹于t时刻状态x_i，产生序列为$Y=(y_1,y_2,\cdots,y_t)$的前向概率；同理$\beta_i(t)$表示后向概率。使用从穷举到递归的方法反复迭代计算$\alpha_j(t)$、$\beta_i(t)$及$\delta_i(t)$、$\xi_{ij}(t)$，并进行模型重估，直至$P(Y|\lambda)$收敛。

在HMM模型的拟合中，状态数目的选择是一个比较困难的问题。由于只有一条序列参与训练，这里采用AIC准则来确定每个模型的最佳状态数，即：

$$AIC=-2L/n+2K/n \tag{5}$$

式中，L是模型中的对数极大似然值，n是观测值数目，k是被估计的参数个数（模型的状态数），AIC准则要求该式子的值越小越好。AIC的大小取决于L和k，k取值越小，AIC越小；L取值越大，AIC值越小。k小意味着模型简洁，L大意味着模型精确，因此可以使用AIC准则评价模型的简洁性和精确性。

4.3 轨迹相似性度量方法Sim-HMM

对轨迹数据集T的参考序列进行HMM模型训练，得到大小为t的参考HMM模型集$\{\lambda_1,\lambda_2,\cdots,\lambda_t\}$，即由序列$T_i$训练得到的HMM模型参数集合$\lambda_i$，对整个轨迹集$T$中的每条轨迹$T_i(i<n)$进行模型学习，可获得轨迹序列$T_i$在模型$\lambda_j$下的似然概率密度函数$p_{ij}=f(T_i,\lambda_j)$，其时间复杂度为$O(n\times t)$，一般使用前向算法求该概率函数值，本文提出一种度量轨迹相似度方法Sim-HMM，其基本思想如下：

在轨迹相似度分析中，对称距离、POR（Porikli）距离、BP（Bicego-Panuccio）距离等都是基于产生概率度量不同轨迹之间距离的方法，这里采用BP距离来度量轨迹之间的相似性，当轨迹序列的模型有较大差距时，根据它得到的分析结果最好。设有轨迹序列T_i和T_j，T_i和T_j之间的BP距离为：

$$dist(i,j)=\left(\frac{l_{ji}-l_{ii}}{l_{ii}}+\frac{l_{ij}-l_{jj}}{l_{jj}}\right)\Big/2 \tag{6}$$

式中，$l_{ij}=\dfrac{\log p_{ij}}{length(T_i)}$、$length(T_i)$表示序列$T_i$具有的长度。

定义2 相似性度量函数：给出一个概率距离阈值ε和两个轨迹序列T_i和T_j，两个序列的相似性度量函数表示如式（7）所示：

$$f_{\varepsilon ij}=f_\varepsilon(T_i,T_j)=\begin{cases}0, dist(i,j)>\varepsilon\\ 1-\dfrac{dist(i,j)}{\varepsilon}, \text{else}\end{cases} \tag{7}$$

轨迹相似度的域定为[0,1]，度量值在0到1之间变化，两条轨迹之间的相似度越高，函数值越大。当两个轨迹点完全一样的时候，函数值为1；当轨迹点的概率距离大于ε，函数值为0时表示轨迹全无相似性。

计算得到n个轨迹序列在每个模型下的产生概率后，计算相互之间的BP距离，根据相似性度量函数得到轨迹的相似度矩阵Affinity Matrix，如式（8）所示，以上度量相似度的方法即为Sim-HMM方法，以矩阵的第i行作为该对应轨迹T_i的特征表示。

$$\boldsymbol{AM}=\begin{bmatrix}f_{\varepsilon 11},f_{\varepsilon 12},\cdots,f_{\varepsilon 1n}\\ f_{\varepsilon 21},f_{\varepsilon 22},\cdots,f_{\varepsilon 2n}\\ \vdots\quad\vdots\qquad\vdots\\ f_{\varepsilon n1},f_{\varepsilon n2},\cdots,f_{\varepsilon nn}\end{bmatrix} \tag{8}$$

根据以上相关描述给出算法1：

算法1 Sim-HMM(HMM,Trajectory)

Input：HMM，Trajectory T_i，Trajectory T_j，ε

Output：Affinity

1. $p_{ij}\leftarrow f(T_i,\lambda_j)$
2. $l_{ij}\leftarrow\dfrac{\log p_{ij}}{length(T_i)}$
3. $dist(i,j)\leftarrow\left(\dfrac{l_{ji}-l_{ii}}{l_{ii}}+\dfrac{l_{ij}-l_{jj}}{l_{jj}}\right)\Big/2$
4. If $dist(i,j)>\varepsilon$ Then
5. $f_{\varepsilon ij}=f_\varepsilon(T_i,T_j)=0$
6. Else
7. $f_{\varepsilon ij}=f_\varepsilon(T_i,T_j)=1-\dfrac{dist(i,j)}{\varepsilon}$
8. End else
9. $Affinity\leftarrow f_{\varepsilon ij}$
10. Return Affinity

该算法的时间复杂度为$O(1)$，第1行获得轨迹序列T_i在模型λ_j下的似然概率密度函数，第2行得到对数密度函数，第3行计算BP距离，4至9行通过判断语句计算相似度。

一般采样的轨迹数目较多，通过建模和相似度度量后会得到的维数较高的Affinity Matrix，在聚类之前，采用主元分析法（Principal Component Analysis，PCA）对轨迹特征Affinity Matrix降维，PCA的实质是在能尽可能好的代表原特征的情况下，将原特征进行线性变换，映射至低纬度空间中。在确定轨迹的特征表示时，仅保留矩阵中差异性大的列，以降低后续处理的复杂性。

4.4 轨迹聚类结果获取

每个聚类代表了一种运动行为，由于上述过程得到的轨迹特征维数均相同，本文采用FCM对轨迹特征进行聚类，并根据其聚类结果对原始的轨迹进行类别标注。该方法的实现如下：

把n个向量$\boldsymbol{x}_i(i=1,2,\cdots,n)$分为$c$个模糊组，并求每组的聚类中心，使得非相似性指标的价值函数达到最小。与引入模糊划分相适应，隶属矩阵U允许有取值在0、1间的元素，不过归一化规定一个数据集的隶属度的和总等于1，如式（9）：

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, 2, \cdots, n \tag{9}$$

FCM的目标函数如式(10)：

$$J(\boldsymbol{U}, c_1, c_2, \cdots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{10}$$

这里 $0 < u_{ij} < 1$，c_i 为模糊组 i 的聚类中心，$d_{ij} = \|c_i - x_j\|$ 为第 i 个聚类中心与第 j 个数据点的欧式距离，$m \in [1, \infty)$ 是一个加权指数。

聚类中心公式如式(11)、(12)：

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{11}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{12}$$

模糊C均值聚类算法以下列步骤进行迭代：

步骤1　用值在0、1间的随机数初始化隶属矩阵 U，使其满足式(9)中的约束条件。

步骤2　用式(11)计算 c 个聚类中心 c_i，$i = 1, 2, \cdots, c$。

步骤3　根据式(10)计算价值(目标)函数。如果它小于某个确定的阈值，或它相对上次价值函数值的改变量小于某个阈值，则算法停止。

步骤4　用式(12)计算新的 U 矩阵；返回步骤2。

上述算法也可以先初始化聚类中心，然后再执行迭代过程。由于不能确保FCM收敛于一个最优解，算法的性能依赖于初始聚类中心。因此要么采用另外的快速算法确定初始聚类中心，要么每次用不同的初始聚类中心启动该算法，多次运行FCM。

在对轨迹进行聚类的过程中，聚类数目的确定是一个非常重要的问题，如果聚类数目设定的不合适，聚类的效果将会很差。本文算法中使用Xie-Beni指标来确定最佳的轨迹分布模式数目，Xie-Beni指标是针对模糊聚类算法提出的一种聚类有效性函数，在应用中效果较好，其思想是使用最小的类与类中心距离平方来衡量类间分离度，使用类中各点与类中心的距离平方和来衡量类内紧密度，在类内紧密度与类间分离度之间寻找一个平衡点。

基于上述描述以及轨迹特征点提取方法，提出基于HMM的轨迹聚类算法2。

算法2　HMM-Cluster

Input：A Set of Trajectory: Tset, Number of Clusters: k
Output：Clusters
1. $R = \{R_1, R_2, \cdots R_r\} (r < n) \Leftarrow T = \{t_1, t_2, \cdots t_n\}$
2. For R_i in R DO
3.　HMM $\lambda_i \leftarrow$ Train(R_i)
4. End for
5. For T_i in Tset
6.　For λ_i in HMM set
7.　$p_{ij} \leftarrow f(T_i, \lambda_j)$
8.　End for
9. End for
10.For T_i in Tset
11.　For T_j in Tset
12.　get the BP distance $dist(i,j)$ of trajectories
13.　Then Calculate the similarity of them $f_{\varepsilon ij}$
14.　End for
15.End for
16.Do dimensionality reduction using PCA
17.Do clustering using FCM with BP distance
18.Return Clusters

设数据集中原始轨迹序列集大小为 n，参照轨迹集大小为 $r(r < n)$。轨迹聚类的时间主要消耗在针对轨迹建立HMM模型、轨迹相似性度量计算、PCA降维和轨迹聚类4个环节，其中针对轨迹建立HMM模型时间复杂度为 $O(r)$，轨迹相似性度量计算，包括轨迹在模型下学习的似然概率获取，计算相似度组成相似度矩阵两部分，时间复杂度分别为 $O(n \times r)$、$O(n^2)$，PCA降维的时间复杂度为 $O(n)$，轨迹聚类的时间复杂度为 $O(n^2)$。因此，算法的总时间复杂度是 $O(r + n + nr + 2n^2)$。

5　实验与测试

本文实验的硬件环境配置为Intel® Core™ i3-4130 3.40 GHz，4.00 GB(RAM)，500 GB，Windows10，实验数据采用Geo-Life项目的真实数据集T-Drive Trajectory，包含10 357辆北京出租车行驶一周的GPS轨迹，共计1 500万个数据点，每条轨迹包含了序号、时间戳、纬度、经度等信息，是典型的时空数据集。本文实验针对数据集中近两百万条轨迹训练HMM模型，计算轨迹相似度距离，并进行聚类处理，用来发现某时间段交通量过载状况，算法采用Matlab进行实现。

5.1　轨迹特征点提取前后存储空间对比

依次选取周一产生的 n 条轨迹数据，如表1所示，将轨迹特征点提取前后的数据集分别存储在.txt文件中，表中分别列出所占存储空间大小。通过实验证明，提取交通轨迹数据特征点，可以降低轨迹数据量，节约存储空间。

表1　轨迹特征提取前后所占存储空间大小

数量/条	前存储空间/KB	后存储空间/KB
50	35	23
500	365	215
1 000	742	436
3 000	2 316	1 153
6 000	4 682	2 316

5.2　轨迹相似性分析性能评估

将提出的轨迹相似性度量方法Sim-HMM与目前已有的时空轨迹相似性度量方法做对比，如LCSS[18]、EDR[19]、wDF[20]等，评估其准确率和效率，这3种典型方法的性能统计如表2所示。

表2　相似度计算法性能统计

度量方法	时间变化	时间敏感	空间特征	映射	空映射
LCSS	√	√	离散	1—1	√
EDR	√	√	离散	1—1	√
wDF	√	√	无	1—1	—

传统的相似度度量方法通常更适合用于静态的数据计算空间距离，而轨迹作为一种动态数据，本文提出的Sim-Cluster方法具有更好的有效性。在对比实验过程中，设定以上4种时空相似度计算算法的空间阈值为100 m，时间阈值为300 s。为了对比距离函数的性能，定义一个1最近邻(1NN)分类器，在被标记的训练数据中，1NN分类器用来预测训练集中最近邻轨迹的标签。虽然不同的距离函数在相似值方面具有不同的变化范围，1NN分类器能很好地解决这个问题。通过1NN分类器，给出一个距离函数和一条轨迹后，被给出轨迹可以用其最近邻轨迹的标签进行预测，也就是说，对给出的轨迹进行标签预测，会影响距离函数的性能。

使用“失误率”对结果进行评价，“失误率”是指预测错误的轨迹数量在轨迹总数中所占比例。$A_$表示预测错误的轨迹数，R表示轨迹总数，“失误率”*errorRate*的含义如式(13)：

$$errorRate=\frac{A\_}{R} \tag{13}$$

相似度度量方法的对比结果如图3所示，图(a)中Sim-Cluster的概率阈值采用$\varepsilon=0.80$，图(b)中采用$\varepsilon=0.85$，k表示轨迹分类数目的变化。wDF方法的失误率较高，在0.7左右，而LCSS方法稍低，在0.4左右，EDR方法失误率较前两者低，在0.3左右，而Sim-Cluster方法是失误率最低的，在0.2左右，究其原因，前4种方法需要对轨迹进行时间和空间度量，涉及到的计算量很大，失误可能性大，本文方法是在HMM建模的基础上进行概率距离计算的，可靠度高。

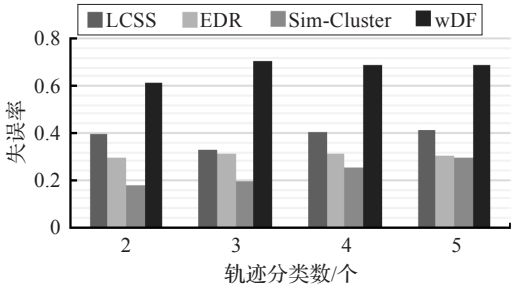


图3(a)　相似度度量方法失误率对比结果($\varepsilon=0.80$)

接下来利用数量不同的轨迹数据集，对轨迹相似度的计算进行多次测试，对比本文提出的Sim-Cluster相似

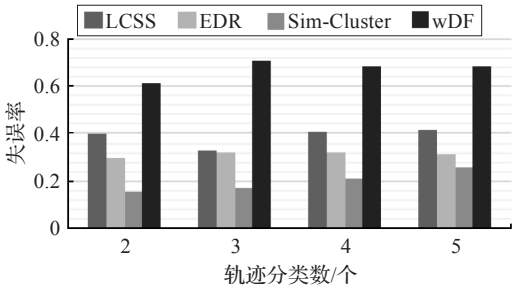


图3(b)　相似度度量方法失误率对比结果($\varepsilon=0.85$)

性度量方法与这4种方法在轨迹相似性分析时的执行效率。如图4所示，当轨迹数量少于1 000时本文所提出的时空相似性度量方法Sim-Cluster执行时间较短，当轨迹数量变多，其执行效率的优势更加明显。传统方法的时间相似性度量基于不同时间间隔，空间和时间度量单位不同，需要进行权重参数选择和计算。本文所提出的时空相似性度量方法不仅通过轨迹特征提取减小数据量，更重要的是基于HMM训练模型，得到概率密度函数后，该方法可以省略对其时空相似性的计算，进一步减少了时间消耗，因此执行效率更高。

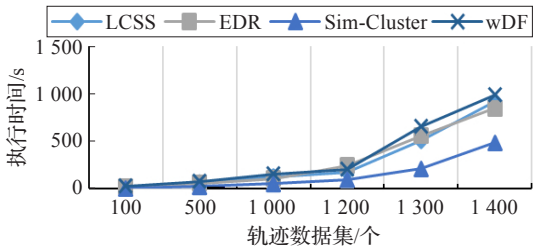


图4　相似性度量方法执行效率对比结果

通过对这4种方法进行实验分析，表明相似性度量方法将直接影响到轨迹聚类结果，执行时间短，正确率高的方法，在轨迹聚类方法具有更大的优势，也能更好地发掘出交通热点，预测交通拥堵情况。

5.3　基于HMM轨迹聚类方法性能评估

轨迹聚类的意义在于能够发现轨迹数据集最有用的分组来对数据进行分析研究，发掘移动对象的运动模式，在此基础上研究交通情况，发现交通热点。通常就是需要得到估计数据簇，使得簇内的轨迹对象具有很高的相似度，而且能够被很好的划分。本文实验对基于HMM轨迹聚类方法HMM-Cluster的划分结果进行评价，并与已有方法做出对比分析。

令$M=\{m_1,m_2,\cdots,m_s\}$是原始数据集的已知的分割结构，$N=\{n_1,n_2,\cdots,n_t\}$是数据集T聚类后的数据结构，此时对于T的任意两个轨迹对象(t_v,t_u)，存在以下4个数据对。

PP：两条轨迹在M和N中都属于同一个簇。

PQ：两条轨迹在M中属于不同的簇，而在N中属于同一个簇。

QP：两条轨迹在M中属于同一个簇，而在N中属于不同的簇。

QQ：两条轨迹在 M 和 N 中都属于不同的簇。

设 PP、PQ、QP、QQ 的统计值分别为 a、b、c、d。$S=a+b+c+d$，这样可以使用4个值，可以计算聚类结果的准确率 P 和召回率 R，分别如式(14)、(15)：

$$\text{Precision:}P=a/(a+c) \tag{14}$$

$$\text{Recall:}R=a/(a+b) \tag{15}$$

F 值是根据 RI 指标衍生出来的评价方法，由 P 和 R 两个值决定的，如式(16)所示：

$$\text{RandIndex:}RI=(a+d)/S,$$

$$F-\text{measure:}F_\beta=(\beta^2+1)PR/(\beta^2P+R) \tag{16}$$

β 的值决定了准确率 P 在 F 值得计算中所占权重值，一般情况下使用 $\beta=1$，也就是 $F1$ 指标来进行聚类结果的评价。$F1$ 值越高，表明聚类效果越好。

如图5显示的是HMM-Cluster方法使用不同参考轨迹数目 r（3种不同颜色深度的柱形分别表示 $(r=100, r=300, r=500)$ 所取得的聚类结果的 $F1$ 值。其中横坐标表示轨迹类数目的变化，不同的矩形高度显示了不同参考轨迹数目导致的 $F1$ 值不同。由该图可知，参考轨迹数目越大，HMM-Cluster方法获得的聚类结果的 $F1$ 值越大；轨迹分类个数越多，评价HMM-Cluster方法聚类效果的 $F1$ 值具有下降的趋势。客观地看，每增加200条参考轨迹，$F1$ 值会增加4%左右，这说明随着参考轨迹数据的数目的增加，能得到更多的轨迹属性信息，提高了轨迹之间的区分度，从而提高 $F1$ 值，也就是轨迹聚类效果能得到明显提升。

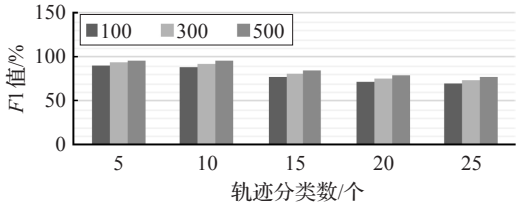


图5　HMM-Cluster聚类结果的$F1$值

下面将基于欧几里德距离的 k-均值聚类方法（k-means）和DBSCAN聚类方法获得的 $F1$ 值显示在图6中，且两者的作用对象是HMM-Cluster方法中获得的基于隐马尔可夫模型的Affinity Matrix。图6的相关设定同图5，图6(a)表示 k-means方法的聚类 $F1$ 值，图6(b)表示DBSCAN方法的聚类 $F1$ 值。当分类数为5时，三者的 $F1$ 值差距不是很明显，k-means和DBSCAN很接近，跟HMM-Cluster相差6%~10%，然后随着分类数的增加，差距有所增长，HMM-Cluster方法可高出30%。

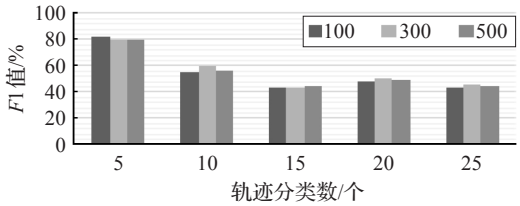


图6(a)　k-means方法聚类结果的$F1$值

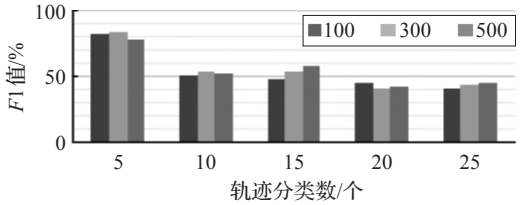
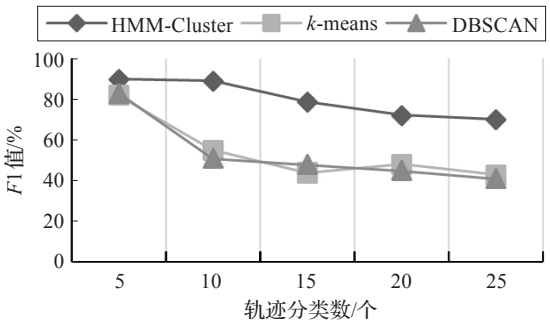
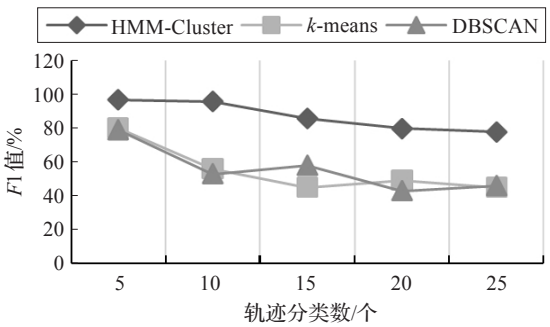


图6(b)　DBSCAN方法聚类结果的$F1$值

为了更直观地对比3种方法聚类的效果差距，将3种方法的 $F1$ 值显示在图7的折线图中，其中，图7(a)为设定参考轨迹数据为100的时候，3种方法获得的 $F1$ 值比较图，图7(b)为设定参考轨迹数目为500的时候，3种方法的 $F1$ 值比较图，横坐标轴表示轨迹类数目的变化。通过分析两张图可以发现，HMM-Cluster的聚类效果优于 k-means和DBSCAN方法，而且在参考轨迹为500时，差距较为明显。



(a)参考轨迹数据为100



(b)参考轨迹数据为500

图7　聚类方法$F1$值对比结果

由此可见，HMM-Cluster方法优点如下：(1)提高了聚类效果。HMM-Cluster方法首先选择参考轨迹序列，并基于参考序列拟合HMM模型，使得到的轨迹相似性矩阵涵盖的信息相对减少。但是通过实验结果发现HMM-Cluster方法集合了BP距离、模糊C均值聚类的优点，并使用相似性度量函数和PCA降维方法进行相似度矩阵的构建和优化，通过对它与两个经典聚类算法的 $F1$ 值分析，发现其聚类结果 $F1$ 值比前两者高。(2)HMM-Cluster计算轨迹的概率距离并建立轨迹的Affinity Matrix，使得这个过程的时间复杂度由原来的 $O(n^2)$ 降到了现在的 $O(nr)(r<<n)$，当 n 很大时，r 的取值远小于 n，因此很大程度地降低了算法的时间复杂度，提高了算法的执行效率。

5.4 交通量过载发现实验结果与分析

利用本文提出的方法对轨迹进行聚类,可以得到反映了移动对象的运动规律和行为模式的时空邻近的轨迹簇,即聚类结果,该聚类结果可用于发现区域交通量过载。本文实验将对聚类发现交通过载的情况进行展示和分析。

实验环境和实验数据集同上,聚类结果如图8所示,以北京市地理轮廓为背景,横坐标为经度变化情况,纵坐标为纬度变化情况。首先将记录了北京市约1 600 000条原始出租车轨迹数据的.txt文件载入Matlab后,进行预处理,然后在矢量地图文件中进行标注,结果如图8(a)蓝色部分所示。下面设定经验值 N 为判断交通量过载区域的阈值,将聚类结果不同的轨迹数据值与 N 作比较,当某区域的聚类结果大于 N 时,判定为过载区域。不同的 N 值对交通量过载区域判定的结果影响不同,经测试当 $N=2$,判定结果比较接近真实情况。

分别针对上班交通早高峰时间段7:30—8:30和下午交通闲暇时间段15:00—16:00进行轨迹聚类分析,用红色标记发现的交通量过载区域。

图8(b)表示在下午交通闲暇时间段15:00—16:00发现的交通量过载区域,可以发现该时间段交通量过载区域较小,交通较为通畅。

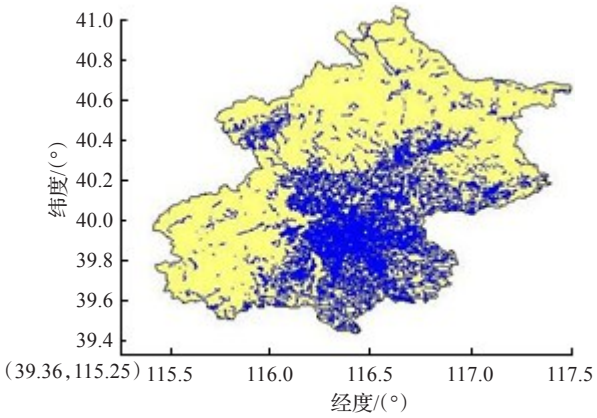


图8(a) 出租车轨迹数据标注结果

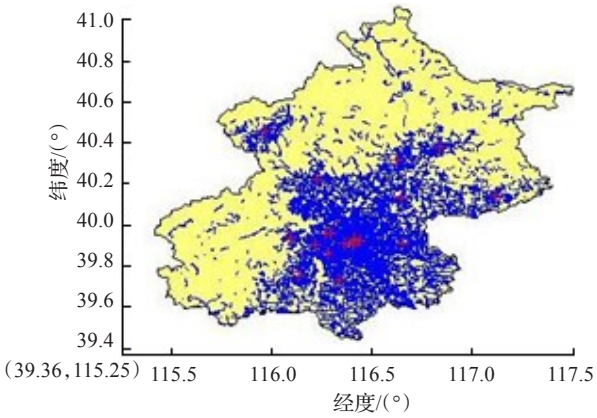


图8(b) 轨迹聚类结果图(闲暇时段)

图8(c)表示表示在上班交通早高峰时间段7:30—8:30发现的交通量过载区域,可以发现该时间段交通量过载区域较大,交通比拥堵情况比较严重。

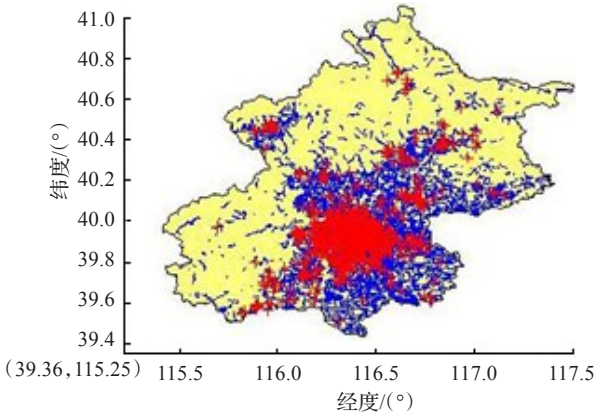


图8(c) 轨迹聚类结果图(早高峰时段)

6 结束语

提出了一种基于HMM模型的轨迹聚类方法,首先对时空轨迹进行特征点提取,减少轨迹数据量,节约存储成本;然后为参照轨迹拟合HMM模型,针对HMM模型对轨迹数据集中的每条轨迹计算似然概率;接下来基于BP距离,提出相似度度量函数,得到估计相似度矩阵;最后使用模糊C类均值算法进行聚类,并提出聚类算法HMM-Cluster,实验验证了算法的运行效率。HMM-Cluster算法可以有效地对交通轨迹进行聚类分析,挖掘移动对象的运行模式,从而发现交通密集区域,进行交通量过载发现。

参考文献:

[1] Niedermayer J, Zufle A, Emrich T, et al.Probabilistic nearest neighbor queries on uncertain moving object trajectories[J].Proceedings of the VLDB Endowment, 2014,7(3):205-216.

[2] Ferreira N, Klosowski J T, Scheidegger C E, et al.Vector field *k*-means: Clustering trajectories by fitting multiple vector fields[J].Computer Graphics Forum, 2013, 32:201-210.

[3] Vlachos M, Hadjieleftheriou M, Gunopulos D, et al.Indexing multidimensional time-series[J].Journal of VLDB, 2006, 15(1):1-20.

[4] Chen L, Özsu M T, Oria V.Robust and fast similarity search for moving object trajectories[C]//Proceedings of SIGMOD, 2005:491-502.

[5] Ding H, Trajcevski G, Scheuermann P.Efficient similarity join of large sets of moving object trajectories[C]// Proceedings of International Symposium on Temporal Representation Reasoning, 2008:79-87.

[6] Keogh E, Ratanamahatana C A.Exact indexing of dynamic

time warping[J].Knowledge and Information Systems，2005，7（3）：358-386.

[7] Agrawal R，Faloutsos C，Swami A N.Efficient similarity search in sequence databases[C]//Proc of Intl Conf on Data Organazation，1993：69-84.

[8] Dubuisson M P，Jain A K.A modified Hausdorff distance for object matching[C]//Proceedings of International Conference on Pattern Recognition，1994：566-568.

[9] 陈锦阳，宋加涛，刘良旭，等.基于改进Hausdorff距离的轨迹聚类算法[J].计算机工程，2012，38（17）：157-161.

[10] Wei J，Yu H，Chen J H，et al.Parallel clustering for visualizing large scientific line data[J].Large Data Analysis and Visualization，2011，2（1）：47-55.

[11] Lee J G，Han J，Whang K Y.Trajectory clustering：A partition-and-group framework[C]//Proceedings of ACM SIGMOD International Conference on Management of Data，2007：593-604.

[12] Ester M，Kriegel H P，Sander J，et al.A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of 2nd International Conference on KDD（KDD-96），1996：226-231.

[13] Gudmundsson J，Thom A，Vahrenhold J.Of motifs and goals：Mining trajectory data[C]//Proceedings of the 20th International Conference on Advances in Geographic Information Systems，2012：129-138.

[14] Lee J G，Han J，Li X，et al.TraClass：Trajectory classification using hierarchical region-based and trajectory-based clustering[J].Proceedings of the VLDB Endowment，2008，1（1）：1081-1094.

[15] Han J，Kamber M.Data mining：Concepts and techniques[J].Data Mining Concepts Models Methods & Algorithms，2000，5（4）：1-18.

[16] Xu D，Tian Y.A comprehensive survey of clustering algorithms[J].Annals of Data Science，2015，2（2）：165-193.

[17] INRIX home page[EB/OL].[2016-12-01].http://www.inrix.com/default.asp.

[18] Celebi M E，Kingravi H A，Vela P A.A comparative study of efficient initialization methods for the *k*-means clustering algorithm[J].Expert Systems with Applications，2013，40（1）：200-210.

[19] Cao H，Mamoulis N，Cheung D W.Mining frequent spatio-temporal sequential patterns[C]//Proceedings of International Conference on Data Engineering，2005：82-89.

[20] Jeung H，Liu Q，Shen H T，et al.A hybrid prediction model for moving objects[C]//Proceedings of International Conference on Data Engineering，2008：70-79.

（上接63页）

[7] Li J，Stoica P，Wang Z.On robust Capon beamforming and diagonal loading[J].IEEE Transactions on Signal Processing，2003，51（7）：1702-1715.

[8] Mohammadzadeh A B，Mahloojifar A.Eigenspace-based minimum variance beamforming applied to medical ultrasound imaging[J].IEEE Transactions on Ultrasonics Ferroelectrics，and Frequency Control，2010，57（11）：2381-2390.

[9] Hollman K W，Rigby K W，O'Donnell M.Coherence factor of speckle from a multi-row probe[C]//Proceedings of Ultrasonics Symposium，1999，2：1257-1260.

[10] Mallart R，Fink M.Adaptive focusing in scattering media through sound speed inhomogeneities：The van Cittert Zernike approach and focusing criterion[J].The Journal of the Acoustical Society of America，1994，96（6）：3721-3732.

[11] Li P C，Li M L.Adaptive imaging using the generalized coherence factor[J].IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control，2003，50（2）：128-141.

[12] Xu M，Yang X，Ding M，et al.Spatio-temporally smoothed coherence factor for ultrasound imaging[J].IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control，2014，61（1）：182-190.

[13] Asl B M，Mahloojifar A.Minimum variance beamforming combined with adaptive coherence weighting applied to medical ultrasound imaging[J].IEEE Transactions on Ultrasonics，Ferroelectrics and Frequency Control，2009，56（9）：1923-1931.

[14] Wang S L，Li P C.MVDR-based coherence weighting for high-frame-rate adaptive imaging[J].IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control，2009，56（10）：2097-2110.

[15] 吴文焘，蒲杰，吕燚.最小方差波束形成与广义相干系数融合的医学超声成像方法[J].声学学报，2011，36（1）：66-72.

[16] Zeng X，Chen C，Wang Y.Eigenspace-based minimum variance beamformer combined with Wiener postfilter for medical ultrasound imaging[J].Ultrasonics，2012，52（8）：996-1004.

[17] Zhao J，Wang Y，Yu J，et al.Subarray coherence based postfilter for eigenspace based minimum variance beamformer in ultrasound plane-wave imaging[J].Ultrasonics，2015，65：23-33.