文章编号: 0427-7104(2012)04-0450-08

带正则化项的时间序列聚类算法及其应用

陈 南,钟 敏,许伯熹
(复旦大学 数学科学学院,上海 200433)

摘 要: 为了更加准确地对时间序列数据进行聚类分析,运用了带正则化项的时间序列聚类方法,并实现了该聚类方法的算法.将该方法应用于云南地区水准形变数据的实际研究中,以寻找类之间状态转移与地震的关系.数值结果同时表明了带正则化项的时间序列聚类方法比标准的 K-means 方法更有效,更有优势.

关键词: 正则化; 时间序列聚类方法; 标准 K-means 方法; 水准形变
中图分类号: O 29 **文献标志码**: A

在许多实际问题的研究过程中,研究者常常要面对大量的时间序列数据,为了得到期望的结果,对这些数据的分析往往显得至关重要.通常情况下,研究者都希望能够利用聚类分析的方法将这些看似杂乱的数据进行分类,从而获取蕴含在大量数据内部的一些共同特征.从 20 世纪 80 年代以来对数据聚类的研究中,文献[1-3]中介绍的标准 K-means 方法和模糊 C-means 方法是最传统的基于稳态数据的聚类方法.但是对于处理时间序列数据,这类传统的方法往往无法获取与时间相关的有效信息.文献[4]中介绍的模糊 C-回归模型是随后提出的基于回归模型的一种聚类方法.作为对前两类传统聚类方法的改进,这种聚类方法主要运用于带有时间趋势的数据序列,能够体现出一些数据随时间推移所呈现出来的特征.但是对于大量带有噪声的数据或者数据本身的重叠比较严重时,这种聚类方法将会导致数据在不同分类之间过于频繁的跳跃,从而影响分析结果的准确性和合理性.因此,为了克服模糊 C-回归模型中的不足之处,文献[5,6]中提出了带有正则化项的时间序列聚类方法.

带有正则化项的时间序列聚类方法在对时间序列数据聚类后,每一个类所对应的聚类中心不再是一个常数,而是一个依赖于时间的函数.与模糊 C-回归模型不同,带有正则化项的时间序列聚类方法根据文献[7,8]中提出的 Tikhonov 正则化的思想引入了正则化参数,在模糊 C-回归模型聚类方法的权重(判别数据属于相应类别的概率)上增加了一些必要的限制,从而使各个类之间的转移呈现出一定的光滑性.也就是说,保证了数据随着时间的推移能够平稳地从一个状态转移到另一个状态,而不会出现频繁的跳动,并且减小了数据噪声对结果造成的影响.

在中国地质史上,云南地区处于 3 个一级大地构造单元的交汇区,断层众多,自古地壳运动就十分剧烈.从文献[9]中可以看到,云南省作为我国地震最高发的省份之一,5.0 级以上(包括 5.0 级)地震的发生率非常高,6.0 级以上的地震也经常可见.在对地震研究的过程中,地震的孕育和发生与断层的活动密切相关是普遍的共识.文献[10]指出,跨断层而设的观测站得到的形变数据与地震之间存在着一些呼应关系.断层形变一般分为水准(垂直方向)形变和基准(水平方向)形变两种类型.其中水准形变直接测量断层间垂直方向上的位移,仪器漂移和外界干扰的因素较少,测量结果较为真实可靠,更有利于分析形变数据与地震孕育和发生的关系.目前大部分关于地震的研究都是建立在对断层形变速率研究的基础上,并通过观察结果来寻找形变与地震之间的联系,得到了一些初步的分析结果,相关内容可以参考文献[11].

本文主要运用带有正则化项的时间序列聚类方法定量地对云南地区红河断裂带及楚雄-通海断裂带附近 7 个跨断层而设的观测站采集到的水准形变数据进行聚类及分析,致力于分析水准形变与地震孕育和发生的关系.

收稿日期: 2010-10-16
作者简介: 陈 南(1985—),男,硕士研究生,E-mail: chennan@fudan.edu.cn.

1　聚类方法介绍

记 $x(t)=(x_1(t),x_2(t),\cdots,x_n(t))\in X\subset\mathbf{R}^n$ 为样本数据库 X 的一个观测样本，这个样本 $x(t)$包含了与观测对象相关的 n 个数据，而 t 表示这个样本的编号. 特别地，对于时间序列数据而言，t 表示观测时刻，即 $t\in[0,T]$. 一般来说观测时刻是离散的时间点，因此假设集合 X 有 N 个元素，即 $t\in\{t_j\}_{j=1}^N\subset[0,T]$. 聚类方法就是基于这些观测数据的一个分类方法，目标是将样本数据库 X 划分为 K 个不同的类别，而每个类别分别具有各自的聚类中心. 通常可以利用权重 $w(t)=(w_1(t),w_2(t),\cdots,w_K(t))\in W$ 来体现每个时刻的观测样本隶属于各个类别的可能性，即分量 $w_i(t)$定义为在 t 时刻观测样本 $x(t)$隶属于第 i 类的可能性. 分量 $w_i(t)$常用概率来表示，自然地应该满足如下两个约束条件：

$$\begin{aligned}&\sum_{i=1}^{K}w_i(t)=1,\qquad &\forall\, t\in[0,T],\\&w_i(t)\geqslant 0,\qquad &\forall\, t\in[0,T],i=1,2,\cdots,K.\end{aligned}\tag{1}$$

另外，还利用了距离函数 $d(x(t),\phi_i)$：$X\times\Phi\longrightarrow[0,+\infty)$来描述观测样本与第 i 类的相关程度，其中$\Phi=\{\phi_1,\phi_2,\cdots,\phi_K\}\subset\mathbf{R}^n$ 是 K 个聚类中心的集合.

在聚类方法的理论分析中，距离函数可以有各种各样的定义. 普遍采用的是 Mahalanobis 距离函数：

$$d_M(x(t),\phi_i)=\sqrt{(x(t)-\phi_i)^{\mathrm{T}}\Sigma_i^{-1}(x(t)-\phi_i)}\,.\tag{2}$$

Σ_i 表示观测样本关于聚类中心 ϕ_i 的协方差矩阵. 虽然以 Mahalanobis 距离作为衡量标准在理论分析中有很大的优势，但是在实际计算过程中，计算协方差矩阵及其逆矩阵的计算量是很大的. 因此在实际计算时，较为常用的距离函数是 Canberra 距离函数：

$$d_C(x(t),\phi_i)=\frac{1}{n}\sum_{k=1}^{n}\frac{|x_k(t)-\phi_{i,k}|}{|x_k(t)|+|\phi_{i,k}|},\tag{3}$$

和 Minkowski 距离函数：

$$d_q(x(t),\phi_i)=\Big(\sum_{k=1}^{n}|x_k(t)-\phi_{i,k}|^q\Big)^{1/q}.\tag{4}$$

特别地，当 $q=2$ 时，Minkowski 距离函数即成为最常见的 Euclidean 距离函数.

标准 K-means 方法是最一般也是最常用的聚类方法. 这个算法接受预先给定的分类数目 K，并试图将 N 个观测样本构成的样本数据库 X 划分成 K 个类别，使其满足：同一类别中的样本相关性较高，而不同类别中的样本相关性较低. 其中，类别的相关程度正是利用 Euclidean 距离函数计算得到的. 而聚类中心的集合 Φ 与时间无关，即 ϕ_i 不随观测时刻 t 的变化而变动.

标准 K-means 方法的数学表达是：寻找权重集合 W 及聚类中心集合 Φ 使得目标函数：

$$F(W,\Phi)=\sum_{i=1}^{K}\sum_{j=1}^{N}w_i(t_j)d_2(x(t_j),\phi_i)\tag{5}$$

达到最小值. 该方法由迭代算法来实现，每一次迭代由两个部分组成：求解 W 的部分和求解 Φ 的部分. 在算法的第 m 次迭代中，利用第 $m-1$ 次迭代得到的聚类中心 $\phi_i^{\{m-1\}}$ 先对 W 进行更新，得到每一个 t_j 时刻的权重 $w^{\{m\}}(t_j)$. 更新规则如下：

$$w_i^{\{m\}}(t_j)=\begin{cases}1 & i=\arg\min d_2(x(t_j),\phi_i^{\{m-1\}}),\\0 & \text{其他的 } i.\end{cases}\tag{6}$$

接下来，利用新的权重 $w^{\{m\}}(t_j)$，再对 Φ 进行更新：

$$\phi_i^{\{m\}}=\frac{\sum_{j=1}^{N}w_i^{\{m\}}(t_j)x(t_j)}{\sum_{j=1}^{N}w_i^{\{m\}}(t_j)},\tag{7}$$

得到第 m 次迭代的聚类中心 $\phi_i^{\{m\}}$. 最后,计算目标函数 $F(W^{\{m\}},\Phi^{\{m\}})$. 按照上述的过程进行迭代,逐步极小化目标函数 $F(W,\Phi)$,直到满足停机准则.

如果观测是个连续的过程,即在时间段$[0,T]$的每个时刻上都有观测样本,那么目标函数(5)可以写成积分的形式:

$$F(W,\Phi)=\sum_{j=1}^{K}\int_0^T w_i(t)d_2(x(t),\phi_i)\mathrm{d}t. \tag{8}$$

2 带正则化项的时间序列聚类方法

带正则化项的时间序列聚类方法对传统的 K-means 方法进行了两方面的改进. 首先,聚类中心 Φ 与时间有关. 聚类中心不再只是 $\mathbf{R}^n$ 中的一些固定点,而是随着时间的变化而变动,是一个与时刻 t 有关的量 $\Phi=\{\phi_1(t),\phi_2(t),\cdots,\phi_K(t)\}$. 一般来说,可以定义为多项式函数:

$$\phi_i(t)=\phi_i^{(0)}+\phi_i^{(1)}t\ +\phi_i^{(2)}t^2+\cdots+\phi_i^{(\alpha)}t^\alpha, \tag{9}$$

而 $\alpha=1$ 的情况是最常用的.

另一方面,由于数据本身的原因(带有噪音或者数据重叠),非常容易造成一般聚类方法结果的不准确. 为了克服这样的弊端,基于 Tikhonov 正则化的思想,在权重 W 上增加了一定的正则化条件. 数学上可以表述为

$$\|\mathcal{L}(W)\|^2\leqslant C, \tag{10}$$

其中,$\|\cdot\|$ 表示某种范数. 并且将目标函数进一步演变为

$$F_\mu(W,\Phi)=F(W,\Phi)\ +\mu^2\ \|\mathcal{L}(W)\|^2, \tag{11}$$

其中,μ 表示正则化参数,可以预先给定.

引入正则化条件的目的在于:减小由于个别权重的小扰动对整体造成的影响,并且使权重随时间的变化趋于平稳,避免权重产生剧烈的变动. 因而在实际问题的计算中,通常可以取 $\mathcal{L}$ 为关于时间的微分算子,而取 $\|\cdot\|$ 为 L^2 范数,则有

$$\|\mathcal{L}(W)\|^2=\sum_{i=1}^{K}\int_0^T\left(\frac{\partial}{\partial t}w_i(t)\right)^2\mathrm{d}t\leqslant C_\mu, \tag{12}$$

其中 $C_\mu>0$ 是一个与正则化参数 μ 有关的常数. 因此,目标函数(11)可以具体表示为

$$F_\mu(W,\Phi)=\sum_{i=1}^{K}\int_0^T w_i(t)d_2(x(t),\phi_i(t))\mathrm{d}t+\mu^2\sum_{i=1}^{K}\int_0^T\left(\frac{\partial}{\partial t}w_i(t)\right)^2\mathrm{d}t. \tag{13}$$

接下来,将致力于极小化目标函数(13). 由于目标函数中包含了权重 W 和聚类中心 Φ 这两个量,所以同样地,每一次迭代需要分为两个部分分别进行求解. 具体的算法如下:

1) 给定分类数目 K,正则化参数 μ,并设置迭代容差 TOL,迭代次数 $m=1$;

2) 任意选取初始的权重 $\mathrm{W}^{\{1\}}$,满足约束条件(1),求解

$$\Phi^{\{1\}}=\arg\min_{\Phi}F_\mu(W^{\{1\}},\Phi);$$

3) while $|F_\mu(W^{\{m\}},\Phi^{\{m\}})-F_\mu(W^{\{m-1\}},\Phi^{\{m-1\}})|>TOL$

固定 $\Phi^{\{m\}}$ 的值不变,在 $W^{\{m+1\}}$ 满足约束条件(1)的情况下,计算

$$W^{\{m+1\}}=\arg\min_{W}F_\mu(W,\Phi^{\{m\}});$$

固定 $W^{\{m+1\}}$ 的值不变,计算

$$\Phi^{\{m+1\}}=\arg\min_{\Phi}F_\mu(W^{\{m+1\}},\Phi);$$

$m:=m+1$;

end.

注意到(13)式等号右端第二项中是不包含 Φ 的，因此在每一次迭代中，求解 Φ 的部分只需要考虑等号右端第一项即可. 实际上，可以将 $w_i(t)$ 写成一组有限元基函数的线性组合，并运用有限元的方法将问题离散化. 对于离散化的极小值问题，它关于 Φ 是一个最小二乘问题，而关于 W 是一个二次规划问题. 在每一次迭代中，便可以分别对上述两个问题进行了求解.

3　云南水准形变数据的分析

应用第 2 节中介绍的带正则化项的时间序列聚类方法，对云南地区跨断层而设的 7 个观测站(楚雄 Chuxiong，峨山 Eshan，剑川 Jianchuan，建水 Jianshui，石屏 Shiping，通海 Tonghai，下关 Xiaguan)的水准形变观测数据进行了分析. 每个观测站的数据从 1982 年开始到 2009 年结束，以月为单位一共 336 组. 在图 1 中，我们标示出了 22°N～28°N，98°E～105°E 范围内的 7 个观测站，并标示出了距离观测站 200 km 范围以内的 6.0～6.9 级地震，以及距离观测站 350 km 范围以内的 7.0 级以上地震. 图中括号里的数字分别代表了地震发生的年份及震级.

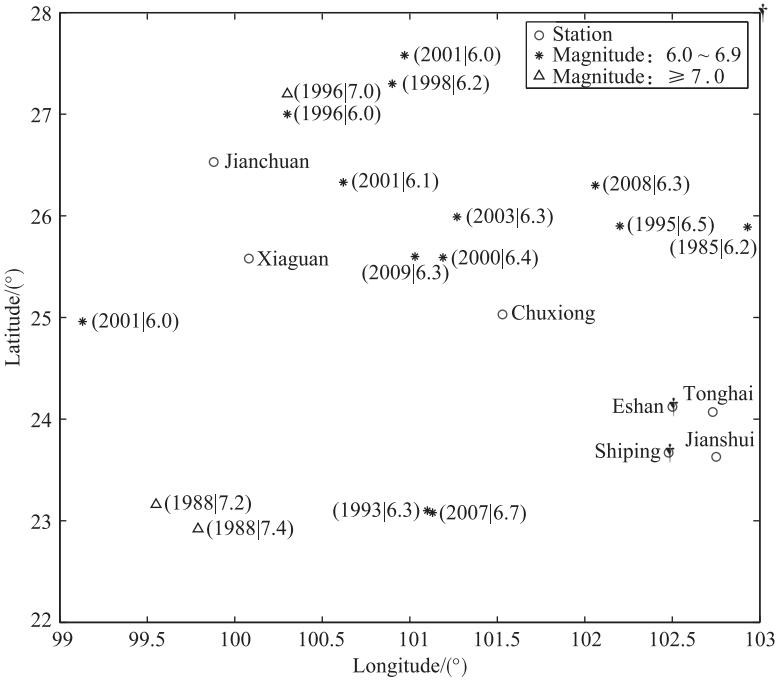


图 1　聚类区域

Fig. 1　The clustering region

首先，将得到的观测数据减去各自的均值处理成相对形变数据，再利用该相对形变数据对红河断裂带以及楚雄-通海断裂带附近的区域进行 Kriging 插值，相关知识可以参阅文献[12]. 运用插值的方法主要是为了解决由于观测站分布不均造成的观测数据在整个空间上分布不均的问题. 由于考虑的是云南地区的整体情况而不是观测站附近的局部情况，因此 Kriging 插值的运用是必要的. 插值的区域是 23.63°N～26.53°N，99.88°E～102.75°E 的地区. 假设插值点为 $\{x_k\}$，并采用 Gauss 型的自相关函数 $\sigma(x_k-x_l)=c\exp(-a_1\|x_k-x_l\|^2)$ 来生成协方差矩阵. 而形变数据在插值区域上的分布函数一般可以由漂移空间 $\{\varphi_l(x)\}$ 中的函数来逼近. 这里取 $\varphi_l(x)=\exp(-a_2\|x-x_l\|^2)$. 在进行 Kriging 插值时，先用观测站的准确数据对参数 c，a_1 和 a_2 进行估计，使插值结果满足一定的光滑性，再利用这组参数来对整个区域进行插值. 这里，取 $c=1$，$a_1=1$ 和 $a_2=10$.

接下来，由于分类数目 K 必须预先给定，则运用文献[5]中提到 Explained Cluster Variance(ECV)准

则来确定聚类方法中所需要的分类数目. 从图 2 中可以看出,在 $K=2$ 时 ECV 准则的增量最为明显,由此可以说明分类数目 $K=2$ 最能够展现聚类方法的分类效果. 因此,选取了 $K=2$ 作为分类数目.

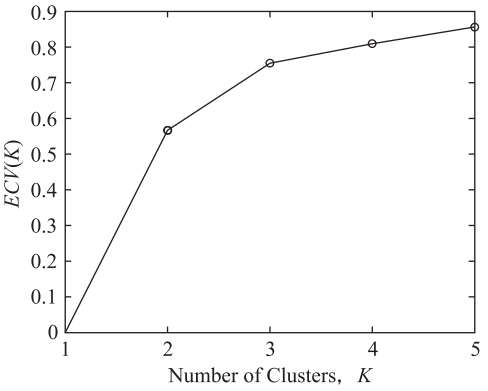


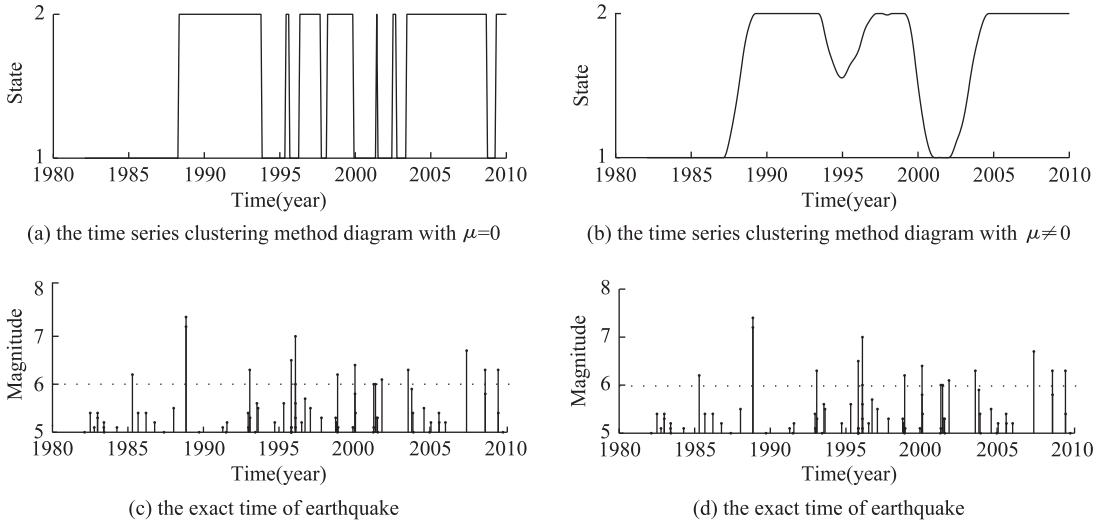
图 2 用 ECV 方法选择分类数目

Fig. 2 Choose clustering number by ECV method

然后,运用带正则化项的时间序列聚类方法对 Kriging 插值后的形变数据进行聚类. 再将权重 W 按时间顺序排列,画出类转移图(权重变化图),结果如图 3 所示.

图 3(a)的聚类结果所选取的正则化参数 $\mu=0$. 图 3(c)是将图 1 中列出的 6.0 级以上地震以及距离观测站 200 km 范围以内的 5.0 级以上地震(包括 5.0 级),依照发生的时刻与震级进行排列. 对比图 3(a)与图 3(c)这两张图,可以发现:两个类之间发生转移的时刻几乎对应了 6.0 级以上地震发生的时刻. 实际上,要特别注意 1988 年 11 月 6 日发生的 7.4 级和 7.2 级澜沧-耿马双震以及 1996 年 2 月 3 日发生的 7.0 级丽江地震,它们在聚类结果中都能充分地体现出来. 两次地震释放了巨大的能量,将云南地区带入了大震过后一段相对平稳的时期,而在类转移图中也对应了一段相对平稳的状态. 而且几乎所有的 6.0 级地震前后都对应着一次类之间的转移,惟一的一次例外是 1985 年 4 月 18 日发生在云南禄劝东北的一次6.2级地震,从类转移图中看不出此次地震带来的转移过程. 由于此次地震隶属的小江断裂带并不在所考察的红河断裂带中,或许正是由于这个原因,聚类结果无法准确地体现出此次地震.

对于正则化参数 $\mu\neq0$ 的情况,结果显示在图 3(b)中. 与 $\mu=0$ 的结果图 3(a)进行对比,可以很清晰地看到,类的转移过程(权重的变化)变得较为平滑,并且没有出现频繁的转移和跳动. 这样,随着正则化参数的引入,考察的将不再是某次地震引起的类转移的时刻,而是一段时间内若干个地震引起的区域能量变化所导致的类转移的时间段. 对比图 3(d)可以看出,类转移的时间段都发生在 7.0 级以上地震或者短时间内 5.0 级或 6.0 级地震频发的时候,并且从总体上反映了云南地区地震发生的周期性规律. 从图中可以看出,在一段时间的地震频发或者某次大震之后,一般会有一段相对平稳的时期,而近年来云南地区处于平静稳定期.



(a) the time series clustering method diagram with $\mu=0$

(b) the time series clustering method diagram with $\mu\neq0$

(c) the exact time of earthquake

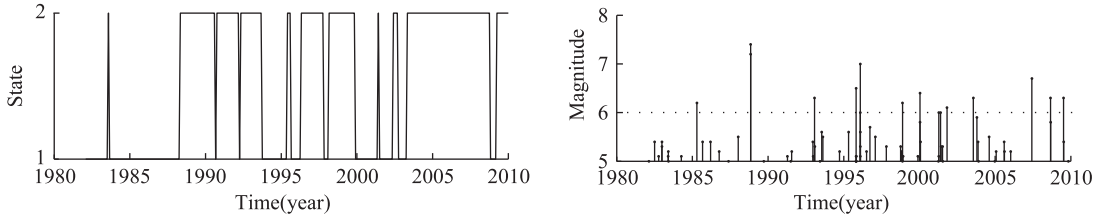
(d) the exact time of earthquake

图 3 时间序列聚类方法得到的类转移图(a)$\mu=0$,(b)$\mu\neq0$,及地震发生时刻图(c)和(d)

Fig. 3 The time series clustering method diagram (a) with $\mu=0$, (b) with $\mu\neq0$, the exact time of earthquake (c) and (d)

作为比较,我们也将标准 K-means 方法运用于云南的水准形变数据中,图 4 展示了标准 K-means 方法得到的结果. 以对比结果来说,本文着重运用的带有正则化项的时间序列聚类方法具有显著的优势. K-means 方法得到的聚类结果在 1983 年附近有一次类的转移,而在 1990 到 1995 年这段时间内的跳动又十

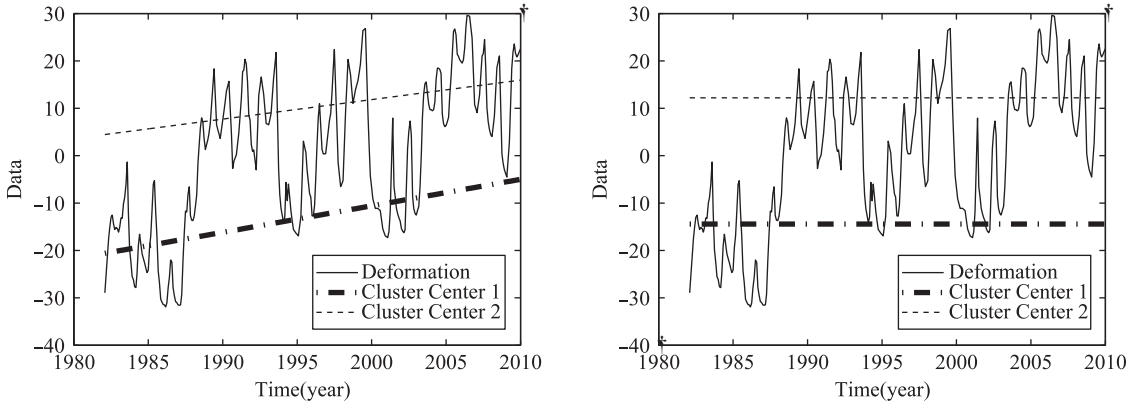
分频繁，从而很难分辨出类转移时刻与地震孕育及发生时刻的内在关系. 而带有正则化项的时间序列聚类方法中引入的正则化参数就是为了避免发生这种类之间剧烈跳动的情况. 对于聚类中心，以峨山观测站为例，如图 5 所示，显然峨山观测站的数据有明显的时间趋势. 若用标准 K-means 方法很难获取数据的时间趋势(图 5(b)). 而带有正则化项的时间序列聚类方法的聚类中心与时间有关，更能体现数据的时间信息(图 5(a)). 因此，对于云南地区在 1982 年到 2009 年这段时间内形变数据与地震数据的综合分析结果来说，带正则化项的时间序列聚类方法比标准 K-means 方法更有优势.



(a) K-means clustering method diagram　　(b) the exact time of earthquake

图 4　标准 K-means 方法得到的类转移图(a)及地震发生时刻图(b)

Fig. 4　K-means clustering method diagram (a) and the exact time of earthquake (b)



(a) the clustering method with regularization　　(b) K-means clustering method

图 5　带正则化项的时间序列聚类方法得到的两个带有时间趋势的聚类中心(a)，及标准 K-meas 方法得到的一般聚类中心(b)

Fig. 5　Clustering centers of clustering method with regularization (a), and K-means clustering method (b)

最后，对聚类结果进行了更深层次的分析，以进一步验证类之间的转移与地震孕育及发生时刻的内在关系. 这里考察了聚类结果的 Markov 状态转移过程，这种方法主要用于分析类之间相互转移的概率. 在运用 Markov 状态转移过程的标准局部高斯核光滑算法[13]之前，首先给出滑动高斯窗口函数的定义：

$$\gamma(t,t_0)=\frac{1}{c}\mathrm{e}^{-\frac{(t-t_0)^2}{\sigma^2}},$$

其中，c 是归一化参数，而 σ 称为噪音强度，是一个用来描述平滑程度的参数. 记 $P_{ij}(t_0)$ 表示包含了 t_0 时刻的一段时间内，从第 i 类向第 j 类转移的概率，那么有

$$P_{ij}(t_0)=\frac{\sum\limits_{t\in A_{ij}}\gamma(t,t_0)}{\sum\limits_{t\in A_i}\gamma(t,t_0)},$$

其中，集合 A_i 包含了这段时间内观测样本属于第 i 个类的所有时刻，而集合 A_{ij} 包含了这段时间内观测样本从第 i 类向第 j 类转移的所有时刻.

依据权重 W 的变化，选取窗口的宽度为 48 个月，沿着时间轴的方向计算类转移的概率，画出 Markov

状态转移图(图 6). 观察图中的 $P_{11}(t)$和 $P_{12}(t)$可以发现,数据停留在第一类的概率在整个时间轴上可以分为相似的 3 个阶段. 其中 1990 年之前为第一阶段,这一阶段中后期,形变数据停留在第一类中的概率迅速下降,而从 1988 年左右起,数据则完全属于第二类. 比较地震发生时刻可以发现,在 1988 年发生了 7.0 级以上的澜沧-耿马双震,由于这次地震产生的巨大能量引起水准形变发生了大幅度的改变,形变数据迅速从第一类转移到第二类. 从 1990 年到 2000 年我们认为是第二阶段,这一阶段中水准形变数据属于第一类的概率也呈现出逐渐减小的趋势,但 1997 年之后又逐步回升. 通过与地震发生时刻以及形变数据进行比对,分析了造成云南地区水准形变数据从第一类向第二类转移的原因是 1996 年在丽江发生的另一次 7.0 级地震. 在这次大地震发生之前,形变已经开始有了转移的趋势,而地震过后转移概率也迅速恢复到原来的状态. 我们可以把 2000 年之后认为是第三阶段,在这一阶段中,转移的趋势与之前两个阶段类似,虽然这个阶段没有 7.0 级以上的大地震,但是 6.0 级以上地震明显增多,这同样引起了水准形变数据的变化,从而导致了类之间的转移. 类似地,考察 $P_{21}(t)$和 $P_{22}(t)$可以发现,云南地区水准形变数据在 2001 年左右向第一类转移的概率最大,而这个时候,恰好连续发生了 3 次 6.0 级以上地震. 而另一个转移的峰值出现在 1996 年,也和丽江的 7.0 级地震吻合,这再一次说明了前文提到的结论:类的转移几乎伴随着地震的发生.

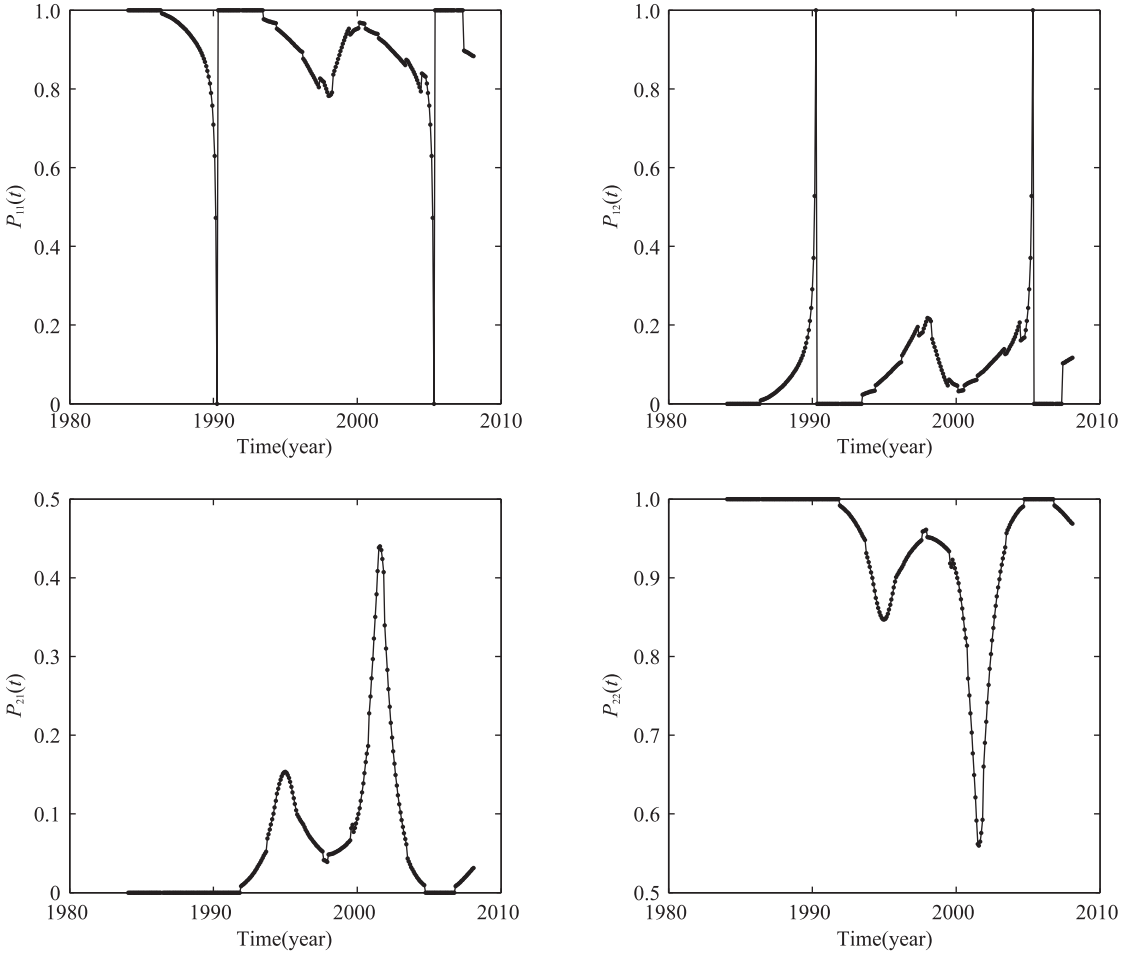


图 6 Markov 状态转移图

Fig. 6 Markov transition diagram

感谢中国地震局地球物理研究所的陈棋福老师提供的水准形变数据. 地震发生时刻与震级的数据来源于:http://www.csndmc.ac.cn/newweb/.

参考文献：

[1] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. New York: Plenum Press, 1981.

[2] Bezdek J C, Hathaway R H, Sabin M J, *et al*. Convergence theory for fuzzy *C*-means: Counterexamples and repairs[J]. *IEEE Trans*, 1987, **17**: 873-877.

[3] Höpner F, Klawonn F, Kruse R, *et al*. Fuzzy cluster analysis[M]. New York: Wiley Press, 1999.

[4] Hathaway R H, Bezdek J C. Switching regression models and fuzzy clustering[J]. *IEEE Transactions on Fuzzy Systems*, 1993, **1**(3): 195-204.

[5] Horenko I. On clustering of non-stationary meteorological time series[J]. *Dynamics of Atmosphere and Oceans*, 2010, **49**(2-3): 164-187.

[6] Horenko I. On simultaneous data-based dimension reduction and hidden phase identification[J]. *Journal of the Atmospheric Sciences*, 2008, **6**: 1941-1954.

[7] Engl H W, Hanke M, Neubauer A. Regularization of inverse problems[M]. Dordrecht: Kluwer, 1996.

[8] Tikhonov A. On the stability of inverse problems[J]. *Dokl Akad Nauk SSSR*, 1943, **39**(5): 195-198.

[9] 苏有锦，李忠华，刘祖荫，等. 20 世纪云南地区 Ms≥5.0 级地震活动的基本特征[J]. 地震研究，2001，**21**(1)：1-9.

[10] 张四新，江在森，王双绪. 南北地震带及青藏块体东部垂直形变与地震活动研究[J]. 西北地震学报，2003，**25**(2)：143-148.

[11] 郭良迁，马　青，杜雪松，等. 华北地区断层形变与地震的关系[J]. 大地测量与地球动力学，2008，**28**(3)：14-20.

[12] 吴宗敏. 散乱数据拟合的模型、方法和理论[M]. 北京：科学出版社，2007.

[13] Loader C. Local regressions and likelihood[M]. New York: Springer, 1999.

Clustering Method with Regularization for Time Series Data and Its Application

CHEN Nan, ZHONG Min, XU Bo-xi

(*School of Mathematical Sciences, Fudan University, Shanghai* 200433, *China*)

Abstract: In order to have a better clustering result of the time series data, a new time series clustering method with regularization is introduced. Applying this method into studying the horizontal deformation data of Yunnan province, the relationship between the transition of the states and the eruption of the earthquakes could be figured out. The numerical results show that, comparing with the standard *K*-means method, this approach manifests its efficiency.

Keywords: regularization; time series clustering method; standard *K*-means method; horizontal deformation