# 有序秩聚类及对地震活跃期的分析*

朱建平　方匡南

**内容提要**:本文在对 Fisher 最优求解有序聚类方法和有序近邻聚类方法剖析的基础上,提出了有序秩聚类分析方法,并对 Fisher 最优求解、有序近邻聚类和有序秩聚类在计算效率上进行了比较分析,研究表明有序秩聚类在处理海量数据具有明显的优势。最后利用该方法对我国南北地震带活跃期进行分析,取得了良好的效果。

**关键词**:有序秩聚类;海量数据;计算速度;地震活跃期
**中图分类号**:O212　　**文献标识码**:A　　**文章编号**:1002 - 4565(2008)12 - 0083 - 04

## Ordinal rank cluster and analysis of active period of earthquakes

Zhu Jianping & Fang Kuangnan

**Abstract**: This paper gives a new method of cluster for ordered samples-ordinal rank cluster based on the Fisher and near-neighbour cluster methods, and compares these three methods on the efficiency of computation. The results show that the ordinal rank cluster is superior to other methods on analysis of massive data. At last, this method is applied to analyze the active period of earthquakes of north-south earthquake belt in china, and it have good effect.

**Key words**: Ordinal rank cluster; Massive data; Computation speed; Active period of earthquakes

## 一、引言

系统聚类和 $K$ 均值等聚类方法,样品的地位是彼此独立的,没有考虑样品的次序。但在实际应用中,有时样品(或变量)的次序是不能变动的,这就需要进行有序聚类分析。假设 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 表示 $n$ 个有序的样品,该样品是根据客观或实际问题需要排序后的数据,在聚类分析中,顺序不能打乱。例如在地质勘探中,需要通过岩心了解地层结构,此时按深度顺序取样,样品的次序不能被打乱;又比如根据儿童不同年龄的体重和身高等指标研究儿童生长发育情况,此时样品是根据不同年龄排序的,顺序也不能被打乱;再比如对国民经济的不同发展阶段进行聚类划分,此时经济数据是按时间排序的,顺序也不能被打乱。

Fisher 最优求解有序聚类法(又称最优分割法)是 Fisher 于 1958 年最先提出,是目前国内外最常用的有序聚类方法。但是 Fisher 最优求解方法的计算量很大,尤其对于海量数据处理时需要耗用很大的计算资源,甚至根本无法得到聚类结果。有序近邻聚类方法克服了 Fisher 最优求解计算速度慢的缺点,但有序近邻聚类方法需要事先确定阈值,并根据不同的阈值进行聚类划分,该方法简单,在海量数据处理时具有明显的优势,但由于需要事先设定阈值,在具体数据分析中很不方便。所以本文提出有序秩聚类分析,克服以上两种方法在海量数据处理时的缺陷。并应用该方法对地震发生频繁的地震带,长期观测所生成的大量数据进行分析,不仅明确了我国地震的空间分布,而且对地震带的活跃期进行客观地划分。

本文所述的有序聚类方法无论是对样品聚类还是对变量聚类都是基于相似指标。假设用 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 表示 $n$ 个有序的样品,每个样品有 $p$ 个变量。把 $n$ 个样品看成 $p$ 维空间中的 $n$ 个点,则两个样品间相似程度就可用 $p$ 维空间中的两点距离公式来度量。变量的相似指标主要以相关程度来度量。

## 二、Fisher 最优求解有序聚类法

Fisher 最优求解法是英国统计学家 R. A. Fisher

*国家教育部社科研究规划项目(06JA910003)资助。

于 1958 年最先提出,在不打乱顺序基础上将样品(变量)客观地分成若干类。其基本思想是,设有序样品依次是 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$($X_{(i)}$ 为 $p$ 维向量),寻找将其分成 $k$ 类的最佳分割点,使类内的误差平方和最小。Fisher 最优求解法的计算步骤:

1. 计算类的直径。设某一类 $G$ 包含的样品是 $X_{(i)}, X_{(i+1)}, \cdots, X_{(j)}$,该类的直径为:

$$D(i,j) = \sum_{t=i}^{j} (X_{(t)} - \bar{X}_G)' (X_{(t)} - \bar{X}_G) \qquad (1)$$

其中 $\bar{X}_G = \frac{1}{j-i+1} \sum_{t=i}^{j} X_{(t)}$,即为类均值。

2. 计算类的误差函数。用 $P(n,k)$ 表示将 $n$ 个有序样品分为 $k$ 类的某一种分法:

$G_1 = (i_1, i_1+1, \cdots, i_2-1)$,

$G_2 = (i_2, i_2+1, \cdots, i_3-1)$, $\cdots$,

$G_k = (i_k, i_k+1, \cdots, n)$,

其中 $i_1 = 1 < i_1 < \cdots < i_k \leq n$。则上述分类法的误差函数为:$e[P(n,k)] = \sum_{t=1}^{k} D(i_t, i_{t+1}-1)$

对于给定的 $n$ 和 $k$,$e(P(n,k))$ 越小,表示各类的离差平方和越小,分类越有效。

3. 求最优求解法的递推公式。

根据式(1),Fisher 最优求解法的递推公式为:

$$e[p(n,2)] = \min_{2 \leq j \leq n} \{D(1,j-1) + D(j,n)\}$$

$$e[p(n,k)] = \min_{k \leq j \leq n} \{e[p(j-1,k-1)] + D(j,n)\} \qquad (2)$$

4. 求精确最优解。从递推公式(2)可知,要得到分点 $j_k$,使得

$$e[(p(n,k)] = e[p(j_k-1,k-1)] + D(j_k,n)$$

从而获得第 $k$ 类:$G_k = \{j_k, \cdots, n\}$,必须先计算 $j_{k-1}$ 使得

$$e[(p(j_k-1,k-1)] = e[p(j_{k-1}-1,k-2)] + D(j_{k-1}, j_k-1)$$

从而获得第 $k-1$ 类:$G_{k-1} = \{j_{k-1}, \cdots\cdots, j_k-1\}$。依此类推,$\cdots$,要得到分点 $j_3$,使得

$$e[(p(j_4-1,3)] = e[p(j_3-1,2)] + D(j_3, j_4-1)$$

从而获得第 3 类:$G_3 = \{j_3, \cdots, j_4-1\}$,必须先计算 $j_2$

$$e[(p(j_3-1,2)] = \min_{2 \leq j \leq j_3-1} \{D(1,j-1) + D(j,j_3-1)\}$$

从而获得第 2 类:$G_2 = \{j_2, \cdots, j_3-1\}$。这时自然获得 $G_1 = \{1, \cdots, j_2-1\}$。

最后获得最优分割为:$G_1, G_2, \cdots, G_k$。

## 三、有序近邻聚类和有序秩聚类法

Fisher 最优求解方法需要计算每种可能的聚类结果,然后选取误差平方和最小的聚类划分。计算量很大,随着样品数 $n$ 的增加,或聚类数 $k$ 的增加,Fisher 最优求解的计算量将呈几何级数增加,尤其是对于海量数据的处理,Fisher 最优求解需要耗费大量的计算资源,有些甚至根本无法得到聚类结果。所以对于海量数据有序聚类时,迫切需要新的方法,有序近邻聚类和有序秩聚类可以很好地解决这个问题。

### (一)有序近邻聚类

有序近邻聚类是在充分体现样品(变量)顺序的基础上构建的一种聚类方法,其基本思想是,先计算相邻样品(变量)的相似指标,并确定阈值,然后根据有序样品(变量)的相似指标值与阈值的比较进行聚类划分。有序近邻聚类方法的计算步骤:

1. 将 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ 看成一个有向链,并构成一个大类。

2. 计算样品(或变量)的相似指标。由于分析的是有序样品,构建相邻样品相似指标向量为:

$$\Gamma = (\gamma_{1,2}, \gamma_{2,3}, \cdots, \gamma_{n-1,n})' \qquad (3)$$

其中,$\gamma_{i,i+1}$ 表示 $X_{(i)}, X_{(i+1)}$ 的相似指标。

3. 根据数据情况或聚类要求,确定一个阈值 $\alpha$。这里需要注意的是,也可以根据聚类个数及 $X_{(i)}$,$X_{(i+1)}$ 的相似指标来确定阈值 $\alpha$。

4. 如果 $\gamma_{i,i+1} > \alpha$,这个有向链就从 $i$ 处断开,$i$ 之前(包括 $i$)归为一类,$i$ 之后归为一类。

在此需要说明的是,如存在 $\gamma_{i,i+1} = \cdots = \gamma_{j,j+1} = \cdots = \gamma_{s,s+1}$ 时,无法确定该在 $i, j$ 还是 $s$ 等处划分时,则在 $i, j, s$ 处计算误差函数 $e(P(n,k))$,找出一个使得误差函数达到最小处断开。

### (二)有序秩聚类

在具体问题分析过程中,由于有序近邻聚类需要预先确定阈值,这样对问题的解决从逻辑上带来了困难,因为有序近邻聚类本质是将大类分成小类的思维过程,如果在聚类之前预先确定阈值,不仅破坏了系统整体聚类的逻辑关系,而且还在设计计算程序和海量数据处理时带来不便。为此,本文在有序近邻聚类方法基础上提出有序秩聚类方法。有序秩聚类方法的思想就是根据有序样品(或变量)数据

相邻相似指标的秩大小进行划分聚类,具体步骤为:

1. 针对有序样品 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$,并将其看成一个有向链,构成一个大类。根据式(3)计算相邻样品的相似指标向量 $\Gamma = (\gamma_{1,2}, \gamma_{2,3}, \cdots, \gamma_{n-1,n})'$。

2. 计算 $\gamma_{i,i+1}$ 的秩,得到秩向量 $R = (r_{12}, r_{23}, \cdots, r_{i,i+1}, \cdots, r_{n-1,n})'$。也就是说最大的相似指标记为 1,最小的相似指标记为 $n-1$。

3. 根据秩向量 $R$ 进行聚类划分。例如聚成两类,就在秩为 1 的地方划分。如聚成三类,在聚成两类划分的基础上,再在秩为 2 的地方划分,这样一直到聚成 $n$ 类为止。

这里需要提及的是,如存在 $r_{i,i+1} = \cdots = r_{j,j+1} = \cdots = r_{s,s+1}$ 时,无法确定该在 $i, j$ 还是 $s$ 等处划分时,则需要对聚类方法进行修正,即在 $i, j, s$ 等处计算误差函数 $e(P(n,k))$,找出一个使误差函数达到最小处断开。这样可以按照相似指标刻划的聚类要求,逐步实现由大类分成小类过程。

有序秩聚类方法的本质是,要得到使相似指标生成的误差函数达到最小的聚类结果。也就是误差函数最小原则,其表现为:以相似指标向量产生的秩向量 $R$ 为基础,首先在秩为 1 的地方划分,生成 2 类的聚类结果,在这样的分类结果中,各类分别计算出的误差函数的总和,是任何分成两类的聚类结果中最小的。依此类推,在秩为 $k$ 的地方划分,即可产生 $k+1$ 类的聚类结果,这一结果是任何划分成 $k+1$ 类时,各类误差函数总和最小的聚类结果。当秩出现相等的情形时,找出一个使得误差函数达到最小处断开。这样一直到聚成 $n$ 类为止。

实现了有序秩聚类后,面临的一个主要问题就是聚类数目的确定。这一问题的解决可以从两方面考虑:一方面从理论上,可以根据相似指标向量和误差函数所提供的信息,确定一个阈值 $\alpha$,通过有序秩聚类分析可以得到聚类数目;另一方面从实践上,根据要解决的实际问题所描述的群体特征的基本要求,确定聚类的数目,得到可以刻划实际状况的聚类结果。

这里需要强调的是,有序近邻聚类中的阈值是事先确定,即在聚类之前首先要确定出阈值,因为在聚类过程中需要将相似指标和阈值比较来生成聚类结果,这也是有序近邻聚类分析在计算设计上遇到的困难之一。而有序秩聚类中的阈值是事后确定,即在由大类分成小类的分析结果之后,根据理论和实际的情况确定阈值,为海量数据的处理提供了行之有效的方法。

## 四、三种有序聚类方法比较分析

在海量数据的分析中,分析方法的好坏不仅体现在结果上,而且反映在运算速度上。下面比较分析这三种方法的计算次数和时间复杂度。

### (一) Fisher 最优求解计算次数

根据 Fisher 最优求解的基本过程,我们可以归纳出其主要计算部分为:

1. 计算类的直径 $D(i,j)$,$1 \leq i \leq j \leq n$,将运算 $\frac{n(n-1)}{2}$ 次,每次都包含了多次的加法与乘法运算。

2. 当 $j \geq 2$ 时,计算 $\min e(P(i,j))$,$1 \leq i \leq n$,$2 \leq j \leq k$,将运算 $\frac{(n-1)(n-2)}{2}$ 次,同样,每次都包含了多次的加法与乘法运算。

3. 根据给定的聚类数目 $k$,确定聚类划分点,最多计算 $(n-2)(n-3)\cdots(n-k)$ 次。

### (二) 有序近邻聚类计算次数

针对有序近邻聚类分析而言,其计算主要分为两部分:

1. 计算近邻相似指标 $\gamma_{i,i+1}$,$1 \leq i < n$,将运算 $n-1$ 次,每次的加法与乘法运算都少于 $D(i,j)$ 中的加法和乘法运算量。

2. 根据给定的阈值 $\alpha$ 确定分类情况,如果 $\gamma_{i,i+1} > \alpha$,有向链就在 $i$ 处断开,这最多只要 $n-1$ 次。

### (三) 有序秩聚类计算次数

有序秩聚类方法是对前两种方法进一步改进产生的,其计算主要分为三部分:

1. 计算近邻相似指标 $\gamma_{i,i+1}$,$1 \leq i < n$,将运算 $n-1$ 次,每次的加法与乘法运算都少于 $D(i,j)$ 中的加法和乘法运算量。

2. 计算 $\gamma_{i,i+1}$ 的秩,将计算 $n-1$ 次。

3. 根据给定的分类数目 $k$,最多将计算 $n-k$ 次。

下面将三种聚类算法的效率用时间复杂度来评价,其结果见表 1 所示。

**表 1　三种有序聚类方法的时间复杂度**

| 聚类方法 | 时间复杂度 |
|---|---|
| Fisher 最优求解聚类 | $o(n^{k-1})$ |
| 有序近邻聚类 | $o(n)$ |
| 有序秩聚类 | $o(n)$ |

通过上面在计算速度上的比较分析,可以看出,

有序近邻聚类和有序秩聚类的运算量远远小于 Fisher 最优求解有序聚类方法运算量,随着样品数 $n$ 的增加,或聚类数 $k$ 的增加,Fisher 最优求解的计算量将呈几何级数增加,而有序近邻聚类和有序秩聚类计算量只是呈线性增加。所以对于海量数据的有序聚类,有序近邻聚类和有序秩聚类相对于 Fisher 最优求解有序聚类在运算速度上具有明显的优势。实际上当 $n > 100$ 时,Fisher 最优求解需要耗费很长时间,而有序近邻聚类和有序秩聚类几秒钟内就可以计算出来;当 $n > 300$ 时,Fisher 最优求解将耗费庞大的计算资源,PC 机几乎无法得到聚类结果,而有序近邻聚类和有序秩聚类可以很快得到聚类结果。

## 五、地震活跃期实证分析

地震具有一定的时空分布规律,不仅在空间上呈现出条带性、分区性。而且从时间上看,不同的地震活动区在长期的地震活动中还表现出平静期与活跃期交替的现象,这种盛衰交替的现象称之为地震活动的阶段性或间歇性,频度高、强度大的时期称为活跃期或活跃幕,频度低、强度低的时期称之为平静期或平静幕,对某个地震区或地震带的活跃期进行客观地划分,有助于地震的预测预报。

本文对地震高发地带的中国南北地震带(包括云南、四川、重庆、甘肃、陕西和宁夏)2001 年 1 月 1 日至 2007 年 12 月 31 日震级大于 M3.5 共 375 个的地震数据进行分析。从图 1 可以看出,不同时间段地震发生的强度和频率都不同,有些阶段地震发生强度大,频率高,而有些阶段地震发生频率低,而且强度也低。通过图 1,可以大致地看出,2001 - 2001 年底地震活动比较活跃,而且 3 次地震接近 M6.0,可认为地震活跃期;2001 - 2003 年中地震发生频率比较低,且绝大多在 M4.5 以下,可认为地震平静期;2003 年中 - 2003 年底,地震发生频率非常高,且震级都比较高,两次地震大于 M6.0,进入短暂的强活跃期;2003 - 2004 年中,频率较低,震级低,进入短暂的平静期;2004 年中 - 2006 年中,频率相对较高,震级相对较高,进入活跃期;2006 - 2007 年底,又进入平静期。但这样划分很主观,有很大的随意性。

由于我们分析的数据为震级大于 M3.5 共 375 个的地震数据,根据表 1 中各种聚类方法时间复杂度的描述,Fisher 最优求解有序聚类计算量非常庞大,如果我们将得到聚类数为 8 的聚类结果,其时间复杂度将为 $o(375^{8-1})$,这样无法得到聚类结果。而有序秩聚类方法的时间复杂度将为 $o(375)$,其计算程序的设计又优于有序近邻聚类,因此本文使用有序秩聚类方法,用 R 语言编程实现该算法,并将前 8 类的聚类结果整理成表 2。

**表 2 中国南北地震带 2001 - 2007 年有序聚类部分结果**

| 分类数 | 序号分割点 | 时间新增分隔点 |
|---|---|---|
| 2 | 109 | 2003-7-21 |
| 3 | 108 109 | 2003-7-21 |
| 4 | 45 108 109 | 2001-10-27 |
| 5 | 45 108 109 139 | 2003-10-16 |
| 6 | 45 108 109 139 200 | 2004-8-10 |
| 7 | 45 108 109 139 200 305 | 2006-6-21 |
| 8 | 21 45 108 109 139 200 305 | 2001-5-24 |

表 2 中,序号分割点是指各样本按时间顺序排列的序号,时间新增分割点指在上一类分割点基础上新增分割点所对应的地震发生时间,例如聚成两类时,分隔点是 109,表示在第 109 个样本前划分为两类,对应的地震发生时间为 2003 年 7 月 21 日 22 时 27 分 8 秒。当聚成三类时,在聚成两类时的分割点基础上新增分割点 108,表示再在 108 号前划分,对应时间为 2003 年 7 月 21 日 16 时 28 分 8 秒。由于 2003 年 7 月 21 日,云南大姚共发生了三次地震,可以考虑合为一类,再结合实际情况把中国南北地震带 2001 - 2007 年划分为六期比较合理,详见表 3。

**表 3 中国南北地震带 2001 - 2007 年地震活跃期划分**

| 中活跃期 | 平静期 | 强活跃期 | 平静期 | 中活跃期 | 平静期 |
|---|---|---|---|---|---|
| 2001-1-1 - 2001-10-26 | 2001-10-27 - 2003-7-20 | 2003-7-21 - 2003-10-15 | 2003-10-16 - 2004-8-9 | 2004-8-10 - 2006-6-20 | 2006-6-21 - 2007-12-31 |

## 六、结论

通过上面的分析可以看出,当样品数比较大时(比如 $n > 300$),Fisher 最优求解法几乎无法得到聚类结果,而有序秩聚类可以很快得到聚类结果,尤其在处理海量数据时具有明显的优势。所以,当对海量数据有序聚类分析时,本文建议使用有序秩聚类方法。有序秩聚类分析方法可以应用到很多领域,比如考古学、地震学、计量经济学、生物学等,本文地震阶段划分实证分析只是有序秩聚类的一方面应用而已。由于客观原因,本文实证分析收集的数据时间段跨度不是很大,地震阶段性划分还不够明显,如果可以获得更长时间跨度的地震数据,可以更好发现我国地震的活动规律,对地震预测预报可以起到更好的作用。

# 我国企业不同生命周期阶段竞争力演化模式实证研究*

曹 裕 陈晓红 王傅强

**内容提要**:本文通过问卷调查方式实证研究了企业竞争力在不同生命周期阶段的差异及其在资源、能力和动态机制三个方面的演化模式。研究表明,我国企业不同生命周期阶段企业竞争力存在显著的差异,随着企业从初创期到成熟期的成长,企业竞争力不断增强,但衰退期企业竞争力大幅减弱。各阶段企业竞争力的构成特点是:初创期企业资源比较缺乏,主要依靠能力进行竞争;成长期企业资源日渐丰富,在竞争力的构成要素中资源与能力并举;成熟期企业人、财、物等资源全面丰富,但能力的作用开始减退;衰退期企业资源开始耗竭,重新回到了依靠能力进行竞争的状态,但该阶段企业学习能力、创新能力和动态机制表现最差,这将影响企业的变革,并导致企业的死亡。

**关键词**:企业竞争力;生命周期;演化模式;实证研究
**中图分类号**:C812 **文献标识码**:A **文章编号**:1002 - 4565(2008)12 - 0087 - 09

# The Empirical Research of the Evolution of Firm Competitiveness of China at Different Life Cycle Stages

Cao Yu Chen Xiaohong Wang Fuqiang

**Abstract**: This paper investigates the firm competitiveness at difference stages of the life cycle and the evolution mode. The results are as followed: there is significant difference at different life cycle stages of firm competitiveness and the feature of competitiveness in China. That is to say, at the firm's born stage, the firm is lack of resources and its competition is mainly rely on the ability; in the firm's growth stage; the firm resources become richer and as important as firm's ability in the comprises of firm competence; in the maturity stage, the firm's resources become very rich, but the role of capability is beginning to decrease; at the firm's recession stage, the firm resource are beginning to be exhausted, and the firm's competitiveness returns to rely on capability.

**Key words**: Firm competitiveness; Life cycle; Evolution model; Empirical research

*本文受教育部长江学者和创新团队发展计划"复杂经济环境下不确定性问题决策理论研究"(批准号 IRT0761)资助;同时,本文还得到湖南省中小企业研究中心资助。

**参考文献**

[1]朱建平. 数据挖掘的统计方法及实践[M]. 北京:中国统计出版社,2005.

[2]张润楚编著. 多元统计分析[M]. 北京:科学出版社,2006.

[3] Johnson, R. A. and Wichern, D. W. (1998), Applied Multivarite Statistical Analysis[M], 4th ed, Prentice-Hall, Inc.

[4] Kaufan L, Rousseeuw PJ. Finding Groups in Data: an Introduction to Cluster Analys[M]is. New York: John Wiley & Sons, 1990.

[5]朱建平,杨贵军,张润楚(2002). 列联资料的有向聚类分析及其应用[J]. 数据统计与管理,Vol. 21, NO. 4, 28 - 33.

[6]杨向东,秦乃岗. 使用有序样品聚类分析分析法划分东南沿海地震活跃幕[J]. 华南地震,2004,24(2).

[7]方开泰. 几个有序样品的聚类方法[J]. 应用数学学报,1982,5(1).

[8]王斌会,方匡南,谢佳斌. R 语言统计分析软件教程[M]. 北京:中国教育文化出版社,2007.1.

**作者简介**

朱建平,男,1962 年生,2003 年毕业于南开大学数学科学学院统计学系,获理学博士学位,现为厦门大学经济学院教授,博士生导师,计划统计系主任,主要研究方向为数理统计、数据挖掘。

方匡南,男,1983 年生,厦门大学经济学院计划统计系博士生,主要研究方向为数理统计与数据挖掘。

(责任编辑:李峻浩)