模糊聚类法在专题数据分级中的应用

蔡 畅, 刘 川, 吴国荣

(信息工程大学 测绘学院,河南 郑州 450052)

摘 要: 在对专题数据的处理中,运用模糊聚类分析法对其建立相似矩阵,以截集法进行分类分级,并对加权聚类法的应用进行了详细的探讨。

关键词: 模糊聚类; 数据分级; 截集; 加权聚类

中图分类号: P283 文献标识码: B 文章编号: 1672- 5867(2008)06- 0011- 03

Application of Fuzzy Clustering Method in the Classification of Thematic Data

CAI Chang LIU Chuan, WU Guo- rong

(Institute of Surveying and Mapping Information and Engineering University, Zhengzhou 450052, China)

Abstract During the process of the thematic data, this paper built up the similar matrix using the Fuzzy Clustering Method Then it carried on the classification of rating by Level- Set Method It also discussed the application of Weighted Clustering Method.

Key words Fuzzy Clustering; data classification of rating; Level- Set; Weighted Clustering

0 引 言

随着计算机技术对地图制图学的作用越来越明显,一些现代数学模型、数学方法也在其中发挥了越来越重要的作用,为解决地图制图数据的广泛性和复杂性提供了依据。例如,在专题数据的分级过程中,应用模糊聚类法可以更合理地对数据进行分类分级。

1 数据分级的意义和标准

分级问题是专题地图制图学中的一个重要问题,许多实践证明,未经分类、分级的制图数据所表现的地理要素的分布特征缺乏可读性,其解释功能很弱,当数据分成具有相同数字特征的各类或各级时,其传播综合信息的功能会大大增强。

在进行数据分级时,一般需要按照下列原则确定:

1)保持数据的分布特征,在分级数一定的条件下,使各等级内部的差异尽可能小,各级数据应聚集在该级代表值周围[4]。

2)任何一个等级内部必须有数据,任何一个数据都必须归属于相应的等级[4]。

3)受人们视觉条件限制,在一幅地图上,一般把级别定为 4~ 7级,7级或 8级被认为是所能用的最大分级数。

2 常见的分级算法

本文把各种数列和级数分级方法称为传统分级算法。

1)等差数列

设有一组数据 $X_1, X_2, \cdots, X_n$, L为数据的最小值,H 为数据的最大值,K 为分级数,则第 i级的分级界线 A_i 为:

$$A_i = L + \frac{i-1}{K}(H - L)$$

2)等比数列

设有一组数据 $X_1, X_2, \cdots, X_n$, L为数据的最小值,H 为数据的最大值,K 为分级数,则第 i级的分级界线 A_i 为:

$$A_i = L(H/L)^{\frac{i-1}{K-1}}$$

级数分级方法中,制图数据处理主要应用算术级数分级和几何级数分级。

1)算术级数分级

$$B_i = a + (i-1)d$$

式中, a为首项的值, d为公差, i为要确定项的序数

2)几何级数分级

$$B_i = gr^{i-1},\ i = 1, 2, \cdots$$

式中, g 为第一个非零的值, r为公比, i为要确定项的

收稿日期: 2007- 09- 20

作者简介: 蔡 畅 (1983-),男,湖南长沙人,地图制图学与地理信息工程专业在读博士研究生,主要研究方向为地理信息系统的开发与应用。

序数。

传统分级方法的优点是: 计算简单, 分级有规律可循,便于读者理解和进行对比分析。但其分级界线的确定脱离数据分布特征, 造成对原始数据信息的某种歪曲[3]。

3 基于截集的模糊聚类算法

聚类分析是按一定原则研究事物分类的一种多元统计分析方法, 它根据样本的多指标、多个观察样品、定量的确定样品、指标之间的相似性和亲疏关系, 其目的是对空间物体的集群性进行分析, 将其分为几个不同的子群(类)。子群的形成是 GIS运作的结果, 根据此可揭示某种地理机制。

模糊聚类的思路: 设有待分类的样本集 $X = \{x_1, x_2, \cdots, x_m\}$, 其中 m 为样本容量, 样本可以是任何一个待分类专题数据集。模糊混合聚类法的基本思想是: 先将待分类的样本 X 按模糊聚类最大矩阵元原理选择合理的初始分类数,然后根据最小二乘法最优准则与模糊决策原理进行修改. 直至得到最优分类为止。

设 x_i, x_j 分别为样本集中的两个样本, 它们之间的相似程度 r_{ij}可以用以下几种方法来衡量[2]:

1)最大最小法

$$r_{ij} = \frac{\min\{x_i, x_j\}}{\max(x_i, x_j)};$$

2)算术平均最小法

$$r_{ij} = \frac{\min\{x_i, x_j\}}{(x_i + x_j)/2};$$

3)几何平均最小法

$$r_{ij} = \frac{\min\{x_i, x_j\}}{\sqrt{x_i \times x_j}}。$$

用上述 3种方法中的任何一种计算各样本之间的相似程度,得到以 r_{ij}为单元的相似矩阵, 并注意到 $r_{ij} = r_{ji}$, $r_{ij} = 1$, 故得到一个上三角矩阵:

$$U = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_1m \\ & \ddots & r_{23} & \cdots & r_{2m} \\ & & \cdots & \cdots & \cdots \\ & & & 1 & r_{m-1m} \\ & & & & 1 \end{bmatrix}$$

选择每行除对角元素外的最大元素作为截集水平 λ, λ按照从大到小的顺序排列, 使值相近的样本聚为一类。根据截集水平 λ的不同, 可以以不同的水平将样本集分为不同数目的初始类。对于剩余的样本利用最小二乘原理, 进行模糊识别归类。

在此过程中将出现以下情况:

1)对于某一截集水平 λ, 若 $r_{ik} > \lambda$, $r_{jk} > \lambda$, 当 $r_{ij} > \lambda$ 时, 毫无疑问样本 i, j, k 可以聚为一类。对于 3个以上的样本, 条件成立时仍有此结论。

2)对于某一截集水平 λ, 若 $r_{ik} > \lambda$, $r_{ik} > \lambda$, 当 $r_{ij} < \lambda$ 时, 尚不能对样本 i, j, k 进行聚类, 此时应降低截集水平, 并对在此截集水平下可能同时属于两个子类的样本进行模糊识别, 具体方法是利用最小二乘法计算样本与某子类之间的距离, 选择距离最近的子类, 进行合并[1]。

表 1是一组专题数据, 表示世界主要国家的国民生产总值。

表 1 世界主要国家 GDP值列表

Tab. 1 GDP value of the major countries in the world

国家	国民生产总值	国家	国民生产总值
英国	1.6	韩国	0.4704
加拿大	0.727	法国	1.565
德国	2.4495	印度	0.505
美国	10.4	西班牙	0.6934
日本	4.2	俄罗斯	1.0069
中国	1.411	墨西哥	0.6468
意大利	1.3	澳大利亚	0.4041

在这里用最大最小法描述各国家之间 GDP值的相似程度, 经过计算得到相似矩阵:

$$\begin{bmatrix} 1 & 0.45 & 0.63 & 0.15 & 0.38 & 0.43 & 0.65 & 0.88 & 0.97 & 0.29 & 0.81 & 0.32 & 0.40 & 0.25 \\ & 1 & 0.72 & 0.07 & 0.17 & 0.95 & 0.29 & 0.52 & 0.47 & 0.65 & 0.56 & 0.70 & 0.89 & 0.56 \\ & & 1 & 0.10 & 0.24 & 0.69 & 0.41 & 0.71 & 0.64 & 0.47 & 0.77 & 0.50 & 0.64 & 0.40 \\ & & & 1 & 0.40 & 0.07 & 0.24 & 0.14 & 0.15 & 0.05 & 0.13 & 0.05 & 0.06 & 0.04 \\ & & & & 1 & 0.17 & 0.58 & 0.34 & 0.37 & 0.11 & 0.31 & 0.12 & 0.15 & 0.10 \\ & & & & & 1 & 0.28 & 0.49 & 0.44 & 0.68 & 0.53 & 0.73 & 0.93 & 0.58 \\ & & & & & & 1 & 0.58 & 0.64 & 0.19 & 0.53 & 0.21 & 0.26 & 0.16 \\ & & & & & & & 1 & 0.90 & 0.33 & 0.92 & 0.36 & 0.46 & 0.29 \\ & & & & & & & & 1 & 0.30 & 0.83 & 0.32 & 0.41 & 0.26 \\ & & & & & & & & & 1 & 0.36 & 0.93 & 0.73 & 0.86 \\ & & & & & & & & & & 1 & 0.39 & 0.50 & 0.31 \\ & & & & & & & & & & & 1 & 0.78 & 0.80 \\ & & & & & & & & & & & & 1 & 0.62 \\ & & & & & & & & & & & & & 1 \end{bmatrix}$$

依次以截集水平 λ= 0.97, λ= 0.95, λ= 0.93对矩阵 进行截集分类。

得到的分类为 {英国, 法国}, {加拿大, 西班牙}, {西班牙, 墨西哥}, {韩国, 印度}, 而其余各国各自为一类, 显然分级数过多, 没有达到分级的要求。

再次以截集水平 λ= 0. 92, λ= 0. 90对矩阵进行截集分类, 得到新的聚类 {中国, 意大利}, {中国, 法国}, 此时检验 {英国, 中国}, {意大利, 法国}的相似程度没有达到截集水平, 故需要再次降低截集水平。

λ= 0. 80时, 得到较优的分类为: {中国, 英国, 法国, 意大利}, {加拿大, 西班牙, 墨西哥}, {韩国, 印度, 澳大利亚}, 其余各国各自为一类。

λ= 0. 71时, 得到的较优的分类为: {美国}, {日本, 德国}, {中国, 英国, 法国, 意大利}, {加拿大, 西班牙, 墨西哥, 韩国, 印度, 澳大利亚}。

对于俄罗斯, 在此截集水平下, 可以归类到 {中国, 英国, 法国, 意大利}中, 也可以归类到 {加拿大, 西班牙, 墨西哥, 韩国, 印度, 澳大利亚}, 此时需要计算并比较俄罗斯 GDP值与两个子类中各国 GDP值的最小二乘距离的均值, 选择较小的距离值, 认为该国家属于相应子类群体。

$$d = \frac{\sum_{i=1}^{m} (G_{俄} - G_i)^2}{m}$$

计算得知俄罗斯归类到 {中国, 英国, 法国, 意大利}中。

因此得到的最终分类为: {美国}, {日本, 德国}, {中国, 英国, 法国, 意大利, 俄罗斯}, {加拿大, 西班牙, 墨西哥, 韩国, 印度, 澳大利亚}。

4　结束语

许多聚类方法的做法是: 对 n个子群, 首先选择最近的两个子群 (点)归为一个新的子群, 这样就得到 $n-1$个子群, 接下去重新计算 $n-1$个子群两两之间的聚类统计量, 再得到 $n-2$个子群 ……, 依次类推, 直至满足所要求的分级数。这是一个迭代的过程, 计算量较大。本文中采用截集法一次对多个子群进行了聚类, 简化了计算过程。

参考文献:

[1] 韩中庚. 数学建模方法与应用 [M]. 北京: 高等教育出版社, 2005.

[2] 何宗宜. 地图数据处理模型的原理与方法 [M]. 武汉: 武汉大学出版社, 2004.

[3] 崔纪锋. 统计专题地图的设计与实现 [D]. 郑州: 信息工程大学测绘学院, 2005.

[4] 李铭. 专题地图统计数据分级的模式识别方法的研究 [J]. 常德师范学院学报 (自然科学版), 2000, 12(1): 78-81.

[责任编辑: 王丽欣]

(上接第 3页)

[6] 刘南, 刘仁义, 谢炯, 等. 基于实体对象层次模型的海量空间数据管理 [J]. 浙江大学学报 (工学版), 2004, 11(38): 1391-1397.

[7] 文艺, 朱欣焰, 袁道华. 面向对象的空间数据组织与管理 [J]. 四川大学学报 (自然科学版), 2000, 3(37): 373-378.

[8] 孔冬艳. 基于对象关系型空间数据库理论的 GIS实现 [D]. 北京: 中国地质大学博士论文, 2006.

[9] 高原, 耿国华, 董乐红. 基于关系数据库的空间对象处理技术研究 [J]. 计算机应用与软件, 2007, 6(24): 12-13.

[责任编辑: 王丽欣]

(上接第 10页)

市地下管线普查中, 电 (磁)法技术不但用于金属管线的探测, 还可用于非金属管线的探测, 且探测精度较高。随着电子技术和计算机技术的发展, 电 (磁)法技术在地下管线探测中的应用更加广泛, 具有很大的应用前景。

参考文献:

[1] 周凤林, 洪立波, 邢方亮. 城市地下管线探测技术手册 [M]. 北京: 中国建筑工业出版社, 1998.

[2] 区福邦. 城市地下管线普查技术研究与应用 [M]. 南京: 东南大学出版社, 1999.

[3] 杨进. 环境地球物理教程 [M]. 北京: 中国地质大学 (北京)出版社, 2004.

[4] 贺颖, 王振东. 工程物探在我国的发展前景 [J]. 水文地质工程地质, 1998, (2): 54-55.

[5] 刘万恩, 蔡克俭. 利用高密度电法探测城市地下管道 [J]. 物探设备, 2003, (12): 260-262.

[6] 刘晓东, 张虎生, 朱伟忠. 高密度电法在工程物探中的应用 [J]. 工程勘察, 2001, (4): 64-66.

[7] 建设部地下管线专业委员会. 城市地下管线探测技术规程 (CJJ-2003) [S]. 北京: 中国建筑工业出版社, 2004.

[8] 杨向东. 地下管线综合探测技术在道路改造中的应用 [J]. 物探与化探, 2001, (12): 477-479.

[9] 姜文青. 探测煤气 PE管的 Noggin500地质雷达 [J]. 地质与勘探, 2004, (10): 152-154.

[10] 陈军, 赵永辉, 万明浩. 地质雷达在地下管线探测中的应用 [J]. 工程地球物理学报, 2005, 2(4): 260-263.

[11] 吴畏, 郭华, 张井, 等. 小波变换在探地雷达探测地下管线信号处理中的应用 [J]. 地球物理学进展, 2003, 18(3): 493-496.

[责任编辑: 王丽欣]