基于特征加权的混合型模糊聚类分析及其在洪水灾害等级划分中的应用

孔令燕,夏乐天

(河海大学理学院,江苏 南京 210098)

摘要:提出了基于特征加权的混合型模糊聚类分析方法,即在改进的 ISODATA 方法的基础上,运用 Relief F 算法确定样本特征的权重。将这种分析方法应用到新疆‘96.7’洪水灾害等级划分中,结果表明该方法是可行的。

关键词:特征加权;模糊聚类;Relief F 算法;洪水灾害

中图分类号:P426.616 **文献标识码**:A **文章编号**:1003-9511(2009)05-0001-03

模糊聚类分析是用数学方法确定研究对象的亲属关系和相似性,从而客观地对研究对象进行分型划类,具有较强的分辨率和广泛的代表性。目前,应用最为广泛的模糊聚类分析方法从理论上来说主要有两类:第一类是基于模糊等价关系的动态聚类方法,又称为系统聚类法;第二类是基于模糊划分的模糊迭代自组织数据分析法(ISODATA 方法),又称为逐步聚类法。这两种方法在许多领域都得到了广泛的应用。由于这两种方法各有利弊[1],张燕[2]构造了一种混合型模糊聚类分析方法,称为‘改进的 ISODATA 方法’,并在股市分析的股票分类问题中得到了较好的应用。笔者在改进的 ISODATA 方法的基础上,采用 Relief F 算法确定样本特征的权重,并将这种方法应用到洪水灾害等级划分中,以验证该方法的可行性与实用性。

1 基于特征加权的混合型模糊聚类分析

1.1 基于模糊等价关系的动态聚类方法

设分类对象有 n 个样本,论域 $X=\{x_1,x_2,\cdots,x_n\}$,论域中每个样本有 m 个特征。基于模糊等价关系的动态聚类方法的步骤如下:

a. 样本的特征指标标准化。对原始数据进行标准化处理,一般可以采用如下计算公式:

$$\begin{cases}\bar{x}_{ij}=\dfrac{x_{ij}-m}{M-m}\\ m=\min\{x_{1j},x_{2j},\cdots,x_{nj}\}\\ M=\max\{x_{1j},x_{2j},\cdots,x_{nj}\}\end{cases}\quad(i=1,2,\cdots,n;j=1,2,\cdots,m)\qquad(1)$$

式中:x_{ij} 为第 i 个事物的第 j 项特征指标值;m,M 分别为这 n 个事物的第 j 项特征指标值的最小值和最大值;$\bar{x}_{ij}$ 为原始数据标准化后的第 i 个事物的第 j 类特征指标的标准值。

b. 求模糊相似矩阵及其传递闭包过程。采用夹角余弦公式标定事物间的相似系数[3],即

$$a_{ij}=\sum_{k=1}^{m}(x_{ik}x_{jk})\Big/\left(\sqrt{\sum_{k=1}^{m}x_{ik}^2}\sqrt{\sum_{k=1}^{m}x_{jk}^2}\right)\quad(i,j=1,2,\cdots,n)\qquad(2)$$

式中:a_{ij} 表示事物 x_i 与 x_j($x_i=(x_{i1},x_{i2},\cdots,x_{im})$,$x_j=(x_{j1},x_{j2},\cdots,x_{jm})$)之间的相似系数,且 $a_{ij}\in[0,1]$;x_{ik} 为第 i 个事物的第 k 项特征指标值,x_{jk} 为第 j 个事物的第 k 项特征指标值。

由式(2)可得到模糊相似矩阵 $\boldsymbol{A}$,该模糊相似矩阵一般不满足传递性,因而利用逐次‘平方法’改造此矩阵,即求模糊相似矩阵的传递闭包过程:$\boldsymbol{A}\to\boldsymbol{A}^2\to\cdots\to\boldsymbol{A}^{2^k}$,直至出现 k_0 使得 $\boldsymbol{A}^{2^{k_0-1}}=\boldsymbol{A}^{2^{k_0}}$。其中 $\boldsymbol{A}^2=\boldsymbol{A}\circ\boldsymbol{A}$ 是模糊矩阵乘法,即将一般矩阵乘法过程中数的乘法应用为逻辑乘,数的加法应用为逻辑加。

这样,可以取适当的水平截值 α,将所要分类的事物按其相似程度分成确定的几类。但是在传递闭包的过程中,往往会造成‘传递偏差’,使分类结果与实际情况有一定的出入[1],因此,为了纠正这个偏差,在上述传递闭包的结果上进一步改造,采用改进的模糊 ISODATA 方法进行分类,并在采用 Relief F

基金项目:国家自然科学基金(10671032);河海大学自然科学基金(2084408319)

作者简介:孔令燕(1983—),女,山东临沂人,硕士研究生,从事概率论与数理统计研究。

循环算法计算权重因子。

1.2 Relief F 算法

基本的 Relief F 算法是 Kira 等[4]在 1992 年提出的,当时仅局限于解决两类的分类问题。1994 年,Kononenko 扩展了 Relief F 算法,可以解决多类的分类问题,其核心就是给特征集中的每一特征赋予一定的权重。设 $X = \{x_1, x_2, \cdots, x_n\}$ 是样本的全体对象,其中 $\boldsymbol{x}_i = [x_{i1}, x_{i2}, \cdots, x_{im}]^{\mathrm{T}}$,表示第 i 个样本的 m 个特征值。可以先对要分析的样本进行一次聚类。选择隶属度较大的样本 $\boldsymbol{x}_i$,找出 R 个与 $\boldsymbol{x}_i$ 同类的最临近的样本 $\boldsymbol{h}_j, j = 1,2,\cdots,R$,然后在不同类的子集中找出 R 个与 $\boldsymbol{x}_i$ 最临近的样本 $\boldsymbol{m}_{Lj}, j = 1,2,\cdots,R$, $L = 1, 2, \cdots, c$, c 表示分类个数, $L \neq c_{\mathrm{class}}(\boldsymbol{x}_i)$。设 $\boldsymbol{d}_{\mathrm{hit}}$ 为 $n \times 1$ 的矩阵,表示 $\boldsymbol{h}_j$ 与 $\boldsymbol{x}_i$ 在同类样本间特征上的差异,则

$$\boldsymbol{d}_{\mathrm{hit}} = \sum_{j=1}^{R} \frac{|\boldsymbol{x}_i - \boldsymbol{h}_j|}{\max(X) - \min(X)} \tag{3}$$

设 $\boldsymbol{d}_{\mathrm{miss}}$ 为 $n \times 1$ 的矩阵,表示 $\boldsymbol{m}_{Lj}$ 与 $\boldsymbol{x}_i$ 在异类样本特征上的差异,则

$$\boldsymbol{d}_{\mathrm{miss}} = \sum_{L \neq c_{\mathrm{class}}(\boldsymbol{x}_i)} \frac{p(L)}{1 - p(c_{\mathrm{class}}(\boldsymbol{x}_i))} \sum_{j=1}^{R} \frac{|\boldsymbol{x}_i - \boldsymbol{m}_{Lj}|}{\max(X) - \min(X)} \tag{4}$$

式中:$p(L)$ 为第 L 类出现的概率,可以用第 L 类的样本数与数据集中的样本总数相比得到。

设 $\boldsymbol{w}$ 是各维特征的权重值,则 Relief F 算法中的特征权重值 $\boldsymbol{w}$ 更新公式为

$$\boldsymbol{w} = \boldsymbol{w} - \boldsymbol{d}_{\mathrm{hit}}/R + \boldsymbol{d}_{\mathrm{miss}}/R \tag{5}$$

1.3 模糊 ISODATA 算法

在以上分类基础上给出初始分划矩阵,初始分划矩阵仅对应于论域的一种分类,但未必是最佳分类。为了挑出最佳分类,需要计算初始分划矩阵中每一类的理想样本,即聚类中心,它对应于每个特征指标下该类元素的平均值。在一个合理的分类中,每一类的元素与该类的聚类中心的距离应尽可能小,实际应用中采用欧氏距离。如何求解适当的分划矩阵及其聚类中心,是较为困难的,笔者采用 Bezdek[5]的收敛算法来求解适当的分划矩阵及其聚类中心,具体步骤如下:

a. 构造初始分划矩阵。设上述传递闭包过程所得分类数为 $c(2 \leqslant c \leqslant n-1)$,在此基础上构造初始模糊聚类相对隶属度矩阵 $\boldsymbol{A}^{(0)}$,并逐步修正。

b. 计算聚类中心 $\boldsymbol{Q}^{(l)}$。用矩阵的转置表示聚类中心:$\boldsymbol{Q}^{(l)} = (\boldsymbol{Q}_1^{(l)}, \boldsymbol{Q}_2^{(l)}, \cdots, \boldsymbol{Q}_c^{(l)})^{\mathrm{T}}$。其中

$$\boldsymbol{Q}_k^{(l)} = \sum_{i=1}^{n} (A_{ki}^{(l)})^2 \boldsymbol{x}_i \Big/ \sum_{i=1}^{n} (A_{ki}^{(l)})^2 \tag{6}$$

式中,分划矩阵 $\boldsymbol{A}^{(l)} = (A_{ij}^{(l)})_{c \times n}$, $(l = 0, 1, 2, \cdots)$。

c. 计算权重。通过公式(3),(4),(5)循环计算权重 $\boldsymbol{w}$,则 $\boldsymbol{w} = (w_1, w_2, \cdots, w_m) = (w_i)$,且 $\sum_{i=1}^{m} w_i = 1$。

d. 修正模糊聚类相对隶属度矩阵[6] $\boldsymbol{A}^{(l)}$。修正公式为

$$A_{ki}^{(l+1)} = 1 \Big/ \sum_{h=1}^{c} \left(\frac{\| \boldsymbol{w}(\boldsymbol{x}_i - \boldsymbol{Q}_k^{(l)}) \|}{\| \boldsymbol{w}(\boldsymbol{x}_i - \boldsymbol{Q}_h^{(l)}) \|} \right)^2 \quad (k = 1,2,\cdots,c;\ i = 1,2,\cdots,n) \tag{7}$$

从而得到 $\boldsymbol{A}^{(l+1)} = (A_{ki}^{(l+1)})_{c \times n}$

e. 矩阵比较。用矩阵范数 $\| \boldsymbol{A} \| = \max| a_{ij} |$ 来比较 $\boldsymbol{A}^{(l)}$ 与 $\boldsymbol{A}^{(l+1)}$。取误差值 $\varepsilon > 0$,若 $\| \boldsymbol{A}^{(l+1)} - \boldsymbol{A}^{(l)} \| < \varepsilon$,则停止迭代,并使 $\boldsymbol{A}^{*} = \boldsymbol{A}^{(l+1)}$ 作为最终的划分矩阵;否则继续迭代。

2 应 用

表 1 是新疆‘96.7’洪水灾害中 10 个地州市的损失情况。根据文献[7],选取受灾面积、受灾人口、房屋破坏和直接经济损失 4 项指标作为聚类指标。

表 1 新疆‘96.7’洪水灾害指标值

受灾地区编号	受灾面积/万 hm^2	受灾人口/万人	房屋破坏/万 m^2	直接经济损失/万元
01	0.1543	6.0000	20.6900	3480
02	1.3740	5.9700	6.2350	1608
03	0.2601	4.3500	2.8430	177
04	2.3520	9.4000	54.5000	7910
05	1.6673	2.9600	58.7280	4946
06	0.5458	2.6200	5.1050	1826
07	1.0792	4.5400	21.7130	7880
08	0.3410	5.6000	1.5560	395
09	0.2140	20.0000	1.8900	1430
10	4.6026	24.7270	13.5920	6327

下面采用基于特征加权的混合型模糊聚类分析法,对受灾地区进行分类。

a. 求初始相对隶属度矩阵。首先根据式(1)和式(2),采用 Matlab 编程得到模糊相似矩阵,再用 Matlab 编程循环迭代求模糊相似矩阵的传递闭包过程。求得模糊等价矩阵为

$$A = \begin{bmatrix} 1.0000 & 0.8012 & 0.8012 & 0.9384 & 0.9384 & 0.9140 & 0.9140 & 0.8012 & 0.8012 & 0.8012 \\ 0.8012 & 1.0000 & 0.8456 & 0.8012 & 0.8012 & 0.8012 & 0.8012 & 0.8456 & 0.8456 & 0.9672 \\ 0.8012 & 0.8456 & 1.0000 & 0.8012 & 0.8012 & 0.8012 & 0.8012 & 0.9455 & 0.9455 & 0.8456 \\ 0.9384 & 0.8012 & 0.8012 & 1.0000 & 0.9465 & 0.9140 & 0.9140 & 0.8012 & 0.8012 & 0.8012 \\ 0.9384 & 0.8012 & 0.8012 & 0.9465 & 1.0000 & 0.9140 & 0.9140 & 0.8012 & 0.8012 & 0.8012 \\ 0.9140 & 0.8012 & 0.8012 & 0.9140 & 0.9140 & 1.0000 & 0.9786 & 0.8012 & 0.8012 & 0.8012 \\ 0.9140 & 0.8012 & 0.8012 & 0.9140 & 0.9140 & 0.9786 & 1.0000 & 0.8012 & 0.8012 & 0.8012 \\ 0.8012 & 0.8456 & 0.9455 & 0.8012 & 0.8012 & 0.8012 & 0.8012 & 1.0000 & 0.9613 & 0.8456 \\ 0.8012 & 0.8456 & 0.9455 & 0.8012 & 0.8012 & 0.8012 & 0.8012 & 0.9613 & 1.0000 & 0.8456 \\ 0.8012 & 0.9672 & 0.8456 & 0.8012 & 0.8012 & 0.8012 & 0.8012 & 0.8456 & 0.8456 & 1.0000 \end{bmatrix}$$

取适中的水平截值 $\alpha = 0.9140$，由上述模糊等价矩阵 A 得到受灾地区的初始分类：{01,04,05}，{02,10}，{03,08,09}，{06,07}，由此得到初始相对隶属度矩阵

$$A^{(0)} = \begin{bmatrix} 0.7 & 0.1 & 0.1 & 0.7 & 0.7 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.7 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.7 \\ 0.1 & 0.1 & 0.7 & 0.1 & 0.1 & 0.1 & 0.1 & 0.7 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.7 & 0.7 & 0.1 & 0.1 & 0.1 \end{bmatrix}$$

b. 采用 Relief F 算法循环计算指标权重。对受灾地区的 4 项指标取相等权重，得到初始权重向量 $w_0 = (0.25, 0.25, 0.25, 0.25)$。利用式(3)，式(4)和式(5)进行迭代得到指标权重向量 $w = (0.2473, 0.2643, 0.2449, 0.2435)$。根据各个指标与洪水灾害的相关性，与受灾人口指标相比，直接经济损失指标的相对重要程度要大[8]。为了显示出直接经济损失指标的重要性，调整它的权值为 0.3000，这样调整后的指标的权重值为 $w_{调} = (0.2473, 0.2643, 0.2449, 0.3000)$，再通过 Relief F 算法循环迭代，得到 $w_{终} = (0.2338, 0.2637, 0.2296, 0.2728)$。

c. 求解最佳相对隶属度分划矩阵。首先，根据式(6)计算聚类中心 $Q_k^{(l)}$，$k = 1,2,3,4$；再根据式(7)进行迭代计算新分划矩阵 $A^{(l+1)} = (A_{ki}^{(l+1)})_{4\times 10}$；最后，取误差值 $\varepsilon = 0.0001$，采用 Matlab 进行编程计算，直至相邻两次 A 的 $\max\{|A_{ki}^{(l+1)} - A_{ki}^{(l)}|\} < \varepsilon$ 为止，从而得最佳相对隶属度矩阵：

$$A^{*} = \begin{bmatrix} 0.0056 & 0.0031 & 0.0008 & 0.9650 & 0.9719 & 0.0031 & 0.0065 & 0.0021 & 0.0180 & 0.0124 \\ 0.0142 & 0.0201 & 0.0045 & 0.0090 & 0.0069 & 0.0145 & 0.0129 & 0.0137 & 0.7037 & 0.8943 \\ 0.0251 & 0.9439 & 0.9885 & 0.0078 & 0.0068 & 0.9534 & 0.0208 & 0.9685 & 0.1911 & 0.0426 \\ 0.9551 & 0.0328 & 0.0062 & 0.0182 & 0.0144 & 0.0290 & 0.9598 & 0.0157 & 0.0873 & 0.0506 \end{bmatrix}$$

根据各指标的贡献程度和最大隶属度原则，得到最优分类结果，受灾地区洪水灾害程度由重到轻排序为：{07,01}，{10,09}，{05,04}，{03,08,06,02}。根据自然灾害指标和洪水灾害程度等级划分标准，可知本研究中受灾地区分类的有效性。

3 结 语

洪水灾害程度的轻重是一个模糊概念。笔者利用混合型模糊聚类分析方法，求出模糊聚类初始相对隶属度矩阵和模糊聚类中心，在此基础上，根据 Relief F 特征优选权重值算法，求得最优权重值，最后，根据最大隶属度原则和权重因子，清晰地给出样本的分类。将这种方法应用到新疆“96.7”洪水灾害等级划分中，结果表明，基于特征加权的混合型模糊聚类分析法，即 Relief F 算法与混合型模糊聚类分析方法相结合是可行的。

参考文献：

[1] 黄健元.模糊集及其应用[M].银川：宁夏人民教育出版社，1999：112-143.

[2] 张燕.混合型模糊聚类分析方法及其应用[J].河海大学学报：自然科学版，2006，34(3)：353-356.

[3] 曹谢东.模糊信息处理及应用[M].北京：科学出版社，2003：178-184.

[4] KIRA K，RENDELL L A. A practical approach to feature selection[C]//Proceedings of the 9th International Workshop on Machine Leaning. San Francisco，CA：Morgan Kaufmann，1992：249-256.

[5] BEZDEKJ C，ANDERSON I. An application of the varieties clustering algorithm to polygonal curve fitting [J]. IEEE SMC，1985，15(5)：637-641.

[6] 陈守煜.模糊聚类循环迭代理论与模型[J].模糊系统与数学，2004，18(2)：57-61.

[7] 徐海量，陈亚宁.洪水灾害等级划分的模糊聚类分析[J].干旱区地理，2000(4)：350-352.

[8] 陈守煜.复杂水资源系统优化模糊识别理论与应用[M].长春：吉林大学出版社，2002.

(收稿日期：2009-06-21　编辑：彭桃英)