混合型模糊聚类分析方法及其应用

张 燕

(河海大学理学院,江苏 南京 210098)

摘要:在动态聚类方法和模糊 ISODATA 方法的基础上,提出了混合型模糊聚类分析方法.该方法首先利用传统的传递闭包方法得到 1 个初始分类,并在此基础上提出初始分划矩阵,根据考虑权重因子的模糊 ISODATA 方法对相关数据进行迭代计算,从而对数据进行有效分类.以股票分类为例对该方法进行实证分析,分析结果表明,应用该方法可以对股票进行有效分类优选.

关键词:模糊聚类;传递闭包;模糊 ISODATA 方法;股票分类

中图分类号:O159 **文献标识码**:A **文章编号**:1000-1980(2006)03-0353-04

事物的分类方法多种多样,最常见的是单层次分类方法和多层次分类方法.但是这些传统的分类方法往往是根据事物的某一类或某几类特征而建立的,这种分类的类别界限明确,属于硬划分.然而,现实世界中事物的特征之间往往不存在绝对清晰的界限,如好与坏、高与矮、冷与热等.因此,人们很难用数学语言对事物特征进行精确描述和严格区分.所以,对于事物的中介性态和归属的划分,往往只能采用软划分.Zadeh 提出的模糊集理论就为这种软划分提供了有力的分析工具,人们开始用模糊的方法来处理聚类问题,并称之为模糊聚类分析方法[1].由于这种方法更好地表达了事物的中介属性,因而,它比一般的简单聚类方法更能合理、客观地反映世界.

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 2 类:(a)基于模糊等价关系的动态聚类方法;(b)基于模糊划分的模糊迭代自组织数据分析法(ISODATA 方法).这 2 种方法在解决实际问题时各有利弊[2].首先,基于模糊等价关系的动态聚类方法,实际上是按模糊相似关系的传递闭包进行分类的,因而在改造模糊相似矩阵的过程中容易造成所谓的"传递偏差",这会造成分类结果与实际情况严重不符的后果.其次,基于模糊划分的 ISODATA 方法的初始分划矩阵的给定必须具有一定的依据,而且在实际应用中,各个特征指标对分类的影响并不相同,因此,在传统的 ISODATA 方法中必须考虑各指标的权重因子.

本文考虑的混合型模糊聚类分析方法将上述 2 种方法结合起来,首先利用传统的传递闭包方法得到一个初始分类,在此基础上提出 ISODATA 方法中必需的初始分划矩阵,并考虑特征指标的不等权重因子,从而使分类结果更符合实际情况.

1 算法介绍

1.1 传递闭包过程

设有论域 $X=\{x_1,x_2,\cdots,x_n\}$,论域中的元素代表所要分类的事物,每一个事物抽取 m 个特征,即 $x_i=(x_{i1},x_{i2},\cdots,x_{im})$,$i=1,2,\cdots,n$.算法的主要步骤如下.

a. 标准化事物样本的特征指标.对原始数据进行处理,一般可以采用如下公式计算[3]:

$$\bar{x}_{ij}=\frac{x_{ij}-m}{M-m}\quad(i=1,2,\cdots,n;j=1,2,\cdots,m) \tag{1}$$

$$m=\min\{x_{1j},x_{2j},\cdots,x_{nj}\}\qquad M=\max\{x_{1j},x_{2j},\cdots,x_{nj}\}$$

式中:x_{ij}——第 i 个事物的第 j 项特征指标值;m,M ——这 n 个事物的第 j 项特征指标值的最小值和最大值;$\bar{x}_{ij}$——标准化后的第 i 个事物的第 j 类特征指标的标准值.

收稿日期:2004-12-10

作者简介:张燕(1978—),女,江苏溧阳人,讲师,硕士研究生,主要从事金融数学研究.

b. 用 a_{ij} 表示事物 x_i 与 x_j 之间的相似系数,其中:

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{im}), x_j = (x_{j1}, x_{j2}, \cdots, x_{jm}) \text{ 且 } a_{ij} \in [0,1] \ (i,j = 1,2,\cdots,n)$$

根据以下实证分析例子的实际情况,本文将采用夹角余弦公式进行标定[4],即有

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

从而得到模糊相似矩阵 A.

c. 由于此模糊相似矩阵不满足传递性,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:$A \to A^2 \to \cdots \to A^{2^k}$,直至出现 k_0,使得 $A^{2^{k_0-1}} = A^{2^{k_0}}$. 其中 $A^2 = A \circ A$ 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 $\wedge$(min),数的加法应用为逻辑加 $\vee$(max).

这样,可以取适当的水平 μ,将所要分类的事物按其相似程度分成确定的几类. 但是从上文知道,在传递闭包的过程中,往往会造成"传递偏差",它会造成分类结果与实际情况有一定的出入. 因此,为了纠正这个误差,下面将在上述传递闭包的结果上进一步改造,即采用模糊 ISODATA 方法进行分类,并在其中考虑权重因子.

1.2 考虑权重因子的模糊 ISODATA 算法

下面将在以上分类基础上给出初始分划矩阵,它仅对应于论域的一种分类,但未必是最佳分类. 为了能从所有分类中挑出最佳分类,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该尽可能的小,本文实际例证中将采用欧氏距离. 如何求解适当的软分划矩阵及聚类中心,使此距离达到最小,是较为困难的. 本文将采用 Bezdek[5] 的收敛算法,具体步骤如下:

a. 设上述传递闭包过程所得分类数为 s,$(2 \leqslant s \leqslant n-1)$,在此基础上构造初始分划矩阵 $P^{(0)}$,并逐步修正.

b. 计算聚类中心 $Q^{(l)}$,$Q^{(l)} = (Q_1^{(l)}, Q_2^{(l)}, \cdots, Q_s^{(l)})^{\mathrm{T}}$,其中

$$Q_k^{(l)} = \sum_{i=1}^{n} (p_{ki}^{(l)})^2 x_i \Big/ \sum_{i=1}^{n} (p_{ki}^{(l)})^2 \qquad (k = 1,2,\cdots,s) \tag{3}$$

$$P^{(l)} = (p_{ij}^{(l)})_{s \times n} \ (l = 0,1,2,\cdots) \qquad x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$$

c. 考虑权重因素 λ,修正分划矩阵 $P^{(l)}$[6],即有:$P^{(l+1)} = (p_{ki}^{(l+1)})_{s \times n}$,其中

$$p_{ki}^{(l+1)} = \left[\sum_{h=1}^{s} \left(\frac{\| \lambda (x_i - Q_k^{(l)}) \|}{\| \lambda (x_i - Q_h^{(l)}) \|} \right)^2 \right]^{-1} \qquad (k = 1,2,\cdots,s;\ i = 1,2,\cdots n) \tag{4}$$

d. 用矩阵范数 $\| C \| = \max\limits_{ij} | c_{ij} |$ 来比较 $P^{(l)}$ 与 $P^{(l+1)}$,取定 $\varepsilon > 0$,若 $\| P^{(l+1)} - P^{(l)} \| < \varepsilon$,则迭代停止,并取 $P^* = P^{(l+1)}$ 作为最终分类矩阵;否则继续迭代.

2 实例分析

为了验证混合型模糊聚类分析方法的实用性,将该方法应用于股市分析的股票分类问题中. 众所周知,股票市场具有高收益与高风险并存的特性,因而,如何对股市进行有效分析和预测是人们关注的核心. 在众多股票行情的优劣分析报告中,证券公司往往根据现有的资料(如收益、主营业务收入、现金流量等)进行简单的直观排序,但某一项因素的绝对优势并不代表整个股票的行情优劣. 本文给出的混合型模糊聚类分析方法就是一种行之有效的分类方法.

本文所采用的原始数据均来源于 www.stockstar.com 数据中心,随机抽取 10 只股票作为研究对象,特征指标值分别取为每股收益、每股经营现金流量和每股主营业务收入,原始数据如表 1

表 1 每股经营现金流量和主营业务收入原始数据

Table 1 Chart of operational cash flow and earnings of core business per share

股票编号	2003 年年报			2004 年中报		
	每股收益/元	每股经营现金流量/元	每股主营业务收入/元	每股收益/元	每股经营现金流量/元	每股主营业务收入/元
01	0.41	0.92	3.09	0.22	0.39	1.50
02	0.51	0.98	4.49	0.06	0.18	1.38
03	1.20	2.04	9.82	0.66	0.63	3.58
04	0.44	0.74	5.79	0.13	0.02	2.04
05	1.08	2.04	15.32	0.35	0.34	4.85
06	0.49	0.20	4.72	0.12	-0.20	1.16
07	0.47	1.70	4.57	0.04	-0.92	0.54
08	0.38	0.72	4.16	0.15	0.01	2.11
09	0.43	-0.27	6.36	0.13	0.30	2.04
10	0.28	0.16	2.02	0.08	0.10	0.63

所示.

2.1　传递闭包过程

a. 计算模糊相似矩阵. 首先利用公式(1)得到标准化特征指标值,此时分类事物的个数 $n=10$,特征指标数 $m=6$. 然后利用公式(2)进行标定,得到如下模糊相似矩阵:

$$\boldsymbol{A}=\begin{bmatrix}1.0000 & 0.9520 & 0.8497 & 0.9406 & 0.7688 & 0.9257 & 0.5046 & 0.9641 & 0.8259 & 0.9205\\ 0.9520 & 1.0000 & 0.8375 & 0.9605 & 0.8303 & 0.9424 & 0.6327 & 0.9493 & 0.8019 & 0.8648\\ 0.8497 & 0.8375 & 1.0000 & 0.8951 & 0.9413 & 0.9092 & 0.5929 & 0.8769 & 0.7361 & 0.6107\\ 0.9406 & 0.9605 & 0.8951 & 1.0000 & 0.9174 & 0.9565 & 0.5723 & 0.9876 & 0.8687 & 0.8028\\ 0.7688 & 0.8303 & 0.9413 & 0.9174 & 1.0000 & 0.8852 & 0.6309 & 0.8710 & 0.7480 & 0.5215\\ 0.9257 & 0.9424 & 0.9092 & 0.9565 & 0.8852 & 1.0000 & 0.4641 & 0.9274 & 0.9131 & 0.8292\\ 0.5046 & 0.6327 & 0.5929 & 0.5723 & 0.6309 & 0.4641 & 1.0000 & 0.5400 & 0.1133 & 0.2572\\ 0.9641 & 0.9493 & 0.8769 & 0.9876 & 0.8710 & 0.9274 & 0.5400 & 1.0000 & 0.8509 & 0.8319\\ 0.8259 & 0.8019 & 0.7361 & 0.8687 & 0.7480 & 0.9131 & 0.1133 & 0.5809 & 1.0000 & 0.8224\\ 0.9205 & 0.8648 & 0.6107 & 0.8028 & 0.5215 & 0.8292 & 0.2572 & 0.8319 & 0.8224 & 1.0000\end{bmatrix}$$

b. 上述矩阵是一个模糊相似矩阵,为了得到模糊等价矩阵,构造它的传递闭包,利用上述算法中给出的公式及 matlab 程序计算,得到了迭代 11 次之后的模糊等价矩阵:

$$\bar{\boldsymbol{A}}=\begin{bmatrix}1.0000 & 0.9605 & 0.9174 & 0.9641 & 0.9174 & 0.9565 & 0.6327 & 0.9641 & 0.9131 & 0.9205\\ 0.9605 & 1.0000 & 0.9174 & 0.9605 & 0.9174 & 0.9565 & 0.6327 & 0.9605 & 0.9131 & 0.9205\\ 0.9174 & 0.9174 & 1.0000 & 0.9174 & 0.9413 & 0.9174 & 0.6327 & 0.9174 & 0.9131 & 0.9174\\ 0.9641 & 0.9605 & 0.9174 & 1.0000 & 0.9174 & 0.9565 & 0.6327 & 0.9876 & 0.9131 & 0.9205\\ 0.9174 & 0.9174 & 0.9413 & 0.9174 & 1.0000 & 0.9174 & 0.6327 & 0.9174 & 0.9131 & 0.9174\\ 0.9565 & 0.9565 & 0.9174 & 0.9565 & 0.9174 & 1.0000 & 0.6327 & 0.9565 & 0.9131 & 0.9205\\ 0.6327 & 0.6327 & 0.6327 & 0.6327 & 0.6327 & 0.6327 & 1.0000 & 0.6327 & 0.6327 & 0.6327\\ 0.9641 & 0.9605 & 0.9174 & 0.9876 & 0.9174 & 0.9565 & 0.6327 & 1.0000 & 0.9131 & 0.9205\\ 0.9131 & 0.9131 & 0.9131 & 0.9131 & 0.9131 & 0.9131 & 0.6327 & 0.9131 & 1.0000 & 0.9131\\ 0.9205 & 0.9205 & 0.9174 & 0.9205 & 0.9174 & 0.9205 & 0.6327 & 0.9205 & 0.9131 & 1.0000\end{bmatrix}$$

取适中的 $\mu_0=0.9025$,则由模糊等价矩阵得到相应的分类结果:{01,02,04,06,08,10},{03,05},{07},{09}.

2.2　ISODATA 算法过程

a. 在上述分类结果上,首先构造初始分划矩阵:

$$\boldsymbol{P}^{(0)}=\begin{array}{c}\begin{matrix}x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10}\end{matrix}\\ \begin{bmatrix}0.7 & 0.7 & 0.1 & 0.7 & 0.1 & 0.7 & 0.1 & 0.7 & 0.1 & 0.7\\ 0.1 & 0.1 & 0.7 & 0.1 & 0.7 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1\\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.7 & 0.1 & 0.1 & 0.1\\ 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.7 & 0.1\end{bmatrix}\begin{matrix}\text{Ⅰ}\\ \text{Ⅱ}\\ \text{Ⅲ}\\ \text{Ⅳ}\end{matrix}\end{array}$$

b. 求解聚类中心,并计算新的分划矩阵进行迭代运算. 首先按式(3)计算聚类中心 $Q_k^{(l)}(l=0,1,2,\cdots)$,其中 $k=1,2,3,4$, $i=1,2,\cdots,10$, $j=1,2,\cdots,6$, x_{ij} 为表中原始数据;然后再按式(4)迭代计算新的分划矩阵 $\boldsymbol{P}^{(l+1)}=(p_{ki}^{(l+1)})_{4\times 10}$,其中考虑欧氏范数. 但是权重因子的确定关键而困难,本文根据要处理的实际情况,随机咨询了 18 位资深专家的意思,综合考虑所采用的特征指标[7],得到指标权向量 $\lambda=(0.25,0.10,0.15,0.25,0.10,0.15)$.

取 $\varepsilon=0.001$,则直至相邻 2 次 $\boldsymbol{P}$ 的 $\max\{|p_{ki}^{(l+1)}-p_{ki}^{(l)}|\}<\varepsilon$ 为止,采用 matlab 编程计算,迭代 45 次之后,得到最佳分划矩阵为

$$\boldsymbol{P}^{*}=\begin{bmatrix}0.9115 & 0.1173 & 0.0001 & 0.0717 & 0.0000 & 0.0818 & 0.2988 & 0.2626 & 0.1288 & 0.9036\\ 0.0020 & 0.0028 & 0.0001 & 0.0071 & 1.0000 & 0.0025 & 0.0102 & 0.0058 & 0.0197 & 0.0042\\ 0.0803 & 0.8694 & 0.0001 & 0.8877 & 0.0000 & 0.9065 & 0.6560 & 0.7107 & 0.7546 & 0.0810\\ 0.0062 & 0.0104 & 0.9997 & 0.0335 & 0.0000 & 0.0092 & 0.0350 & 0.0209 & 0.0969 & 0.0112\end{bmatrix}$$

根据最大隶属度原则和本文设定的权重因子,得到相应的分类结果,按从优到劣排列为:{03},{05},{02,04,06,

07,08,09}, {01,10}.

因此,根据上述分类可以将股票分为4类:编号为03的股票属于最优股,05股票处于第2位,其余股票按此划分归为2类,属于表现较差的股票.从收益、现金流量和业务收入3个方面来综合观察,这个分类结果符合股票的真实情况.证券公司或股民在选择股票的时候对股票优劣程度的判定就有了比较合理而准确的参考标准,并且这个参考标准比直接按某一方面简单排序得到的标准更为综合和全面.

可见2种方法结合的模糊聚类分析具有较好的实用效果,但改进的ISODATA方法必须选取合适的权重因子.如何结合实际选择合理的权重,使分类结果更为理想,有待进一步探讨.

参考文献:

[1] 高新波,谢维信. 模糊聚类理论发展及应用的研究进展[J].科学通报,1999,44(21):2241-2250.
[2] 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.
[3] 许仁忠. 模糊数学及其在经济管理中的应用[M].成都:西南财经大学出版社,1987:217-224.
[4] 曹谢东. 模糊信息处理及应用[M].北京:科学出版社,2003:178-184.
[5] BEZDEK J C,ANDERSON I. An application of the c-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] 李希灿,解明东,许德生,等. 模糊聚类与模糊识别理论模型研究[J].模糊系统与数学,2002,16(2):58-64.

Mixed fuzzy clustering analysis and its application

ZHANG Yan

(*College of Sciences*, *Hohai University*, *Nanjing* 210098, *China*)

Abstract: By combination of the dynamic clustering method with ISODATA method, the mixed fuzzy clustering analysis method was put forward. According to the method, an initial classification was firstly performed by use of the conventional transitive closure method, and then an initial division matrix was constructed. With the weighted fuzzy ISODATA method for iterative computation of the related data, the data were classified effectively. As a case study, the method was applied to stock classification, and the result shows that the method is effective for stock classification and optimal selection.

Key words: fuzzy clustering; transitive closure; fuzzy ISODATA method; stock classification