模糊聚类技术在甘蔗种植户信息分析中的应用

廖燕玲 (柳州职业技术学院, 广西柳州 545005)

摘要　利用数据挖掘中的模糊聚类分析方法, 获取具有不同种植行为特征的甘蔗种植户群组。这有助于制糖企业对各群组作深入分析及制定相应的关怀策略。

关键词　数据挖掘; 模糊聚类; 模糊等价矩阵; 信息管理
中图分类号　S126　　文献标识码　A　　文章编号　0517 - 6611(2007) 28 - 09098 - 02

Application of Fuzzy Clustering Technique in the Information Analysis of Sugarcane Growers

LIAO Yan-ling (Liuzhou Vocational & Technical College , Liuzhou , Guangxi 545005)

Abstract　Fuzzy clustering analysis method in data mining was used to obtain sugarcane grower groups with different planting behavior characteristics , which was helpful to deep analyzing on all groups by the sugar manufactured enterprises and laying out the corresponding solicitude strategies .

Key words　Data mining ;Fuzzy clustering ;Fuzzy equivalent matrix ;Information management

为了提高我国糖业的国际竞争力, 进一步拓展国际市场, 必然要扩大甘蔗种植面积和制糖生产规模。但是目前我国甘蔗种植呈分散态势, 种植户的种植行为缺乏管理和引导, 直接影响甘蔗的产量和质量。为此, 制糖企业应加强甘蔗种植户的信息管理, 以发现潜在的有价值的种植户信息。该文主要讨论利用模糊聚类技术实现甘蔗种植户分类的整个过程, 以便制糖企业发现有价值的种植户、有潜力的种植户等, 并为之提供有效的奖励或扶持, 从而影响相关种植户的种植行为, 并且最终达到扩大甘蔗种植的目的。

1　原始数据的标准化处理

为了讨论方便, 这里仅采用5 个属性作为分类指标, 依次为各户适种总面积(R_1) 、各户拥有优质土地面积(R_2) 、各户平均产量高于当年总平均产量的发生率(R_3) 、各户全面积种甘蔗发生率(R_4) 、各户平均年种甘蔗面积占比(R_5) 。样本对象数为12 人, 分别为 X_1、X_2、X_3、X_4、X_5、X_6、X_7、X_8、X_9、X_{10}、X_{11}、X_{12}。

表1 中的甘蔗种植户数据均为数值型, 但是量纲和单位不同, 故必须对原始数据进行标准化处理。常用的数据标准化处理方法有多种, 该文采用最小最大标准化法[1 - 2] 。

表1　甘蔗种植户数据

样本	R_1	R_2	R_3	R_4	R_5
X_1	20	10	0 .33	0 .50	0 .80
X_2	31	15	0 .50	0 .33	0 .80
X_3	16	16	0 .83	0 .00	0 .90
X_4	8	0	0 .00	0 .00	0 .30
X_5	3	3	1 .00	0 .50	0 .70
X_6	18	18	0 .67	0 .33	0 .60
X_7	8	4	0 .33	0 .17	0 .40
X_8	11	11	0 .83	0 .67	0 .70
X_9	7	3	0 .67	0 .33	0 .60
X_{10}	4	4	1 .00	0 .50	0 .80
X_{11}	5	0	0 .17	0 .17	0 .30
X_{12}	13	13	0 .67	0 .67	0 .90

假设有 m 项属性作为分类指标, 样本集 $X = \{ X_1 , X_2 , \cdots, X_n\}$, 则可用 m 维向量描述样本, 即 $X_i = \{ x_{i1} , x_{i2} , \cdots, x_{im}\}$ $(i = 1 ,2 , \cdots, n)$

标准化公式为:

$$x_{ik}' = \frac{x_{ik} - \min\limits_{1 \le i \le n}\{ x_{ik}\}}{\max\limits_{1 \le i \le n}\{ x_{ik}\} - \min\limits_{1 \le i \le n}\{ x_{ik}\}} \qquad (1)$$

此时, 有 $x_{ik}' \in [0 ,1]$ 。

令 n(样本个数) = 12 , $k = 1 ,2 , \cdots ,5$, 按照式(1) 计算得标准化数据(表2) 。

表2　标准化数据

样本	R_1	R_2	R_3	R_4	R_5
X_1	0 .61	0 .56	0 .33	0 .75	0 .83
X_2	1 .00	0 .83	0 .50	0 .50	0 .83
X_3	0 .46	0 .89	0 .83	0 .00	1 .00
X_4	0 .18	0 .00	0 .00	0 .00	0 .00
X_5	0 .00	0 .17	1 .00	0 .75	0 .67
X_6	0 .54	1 .00	0 .67	0 .50	0 .50
X_7	0 .18	0 .22	0 .33	0 .25	0 .17
X_8	0 .29	0 .61	0 .83	1 .00	0 .67
X_9	0 .14	0 .17	0 .67	0 .50	0 .50
X_{10}	0 .04	0 .22	1 .00	0 .75	0 .83
X_{11}	0 .07	0 .00	0 .17	0 .25	0 .00
X_{12}	0 .36	0 .72	0 .67	1 .00	1 .00

2　甘蔗种植户模糊聚类分析

在数据的初始化完成后, 可以进行聚类。该文采用等价闭包法进行聚类分析, 先由模糊相似关系矩阵 R 得出模糊等价矩阵 t(R) , 然后求出不同 λ 值的布尔矩阵 t(R) $_\lambda$, 最后得到聚类图[1] 。

2 .1　建立模糊矩阵 R　该文采用最大最小法[2 - 3] 计算模糊相似关系矩阵 R 的元素 r_{ij}。

$$r_{ij} = \begin{cases} 1 & (i = j) \\ \dfrac{\sum\limits_{k=1}^{m} \min(x_{ik}' , x_{jk}')}{\sum\limits_{k=1}^{m} \max(x_{ik}' , x_{jk}')} & (i \ne j) \end{cases} \qquad (2)$$

式中, x_{ik}'为第 i 行第 k 列的属性值; x_{jk}'为第 j 行第 k 列的属性值, 而且 $0 \le r_{ij} \le 1$ 。

令属性个数 m = 5 、i ,j = 1 ,2 , …,12 , 按照式(2) 计算 R 矩阵(表3) 。

2 .2　建立模糊等价矩阵 t(R)　模糊矩阵不具有传递性, 通过褶积求得 R 的传递闭包($R^{16} = R^8$) , 即模糊等价矩阵 t(R) $= R^{16} = R^8$, 如表4 所示。

作者简介　廖燕玲(1965 -) , 女, 广西柳州人, 讲师, 从事数据库研究。
收稿日期　2007-07-20

表3　　模糊相似关系矩阵 R

样本	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	1.00	0.72	0.54	0.06	0.51	0.63	0.37	0.67	0.48	0.58	0.16	0.71
X_2	0.72	1.00	0.62	0.05	0.42	0.72	0.31	0.57	0.47	0.47	0.13	0.65
X_3	0.54	0.62	1.00	0.05	0.39	0.64	0.25	0.55	0.39	0.45	0.07	0.66
X_4	0.06	0.05	0.05	1.00	0.00	0.05	0.15	0.05	0.07	0.01	0.12	0.04
X_5	0.51	0.42	0.39	0.00	1.00	0.50	0.29	0.70	0.65	0.92	0.13	0.60
X_6	0.63	0.72	0.64	0.05	0.50	1.00	0.35	0.63	0.61	0.47	0.14	0.65
X_7	0.37	0.31	0.25	0.15	0.29	0.35	1.00	0.29	0.41	0.26	0.42	0.28
X_8	0.67	0.57	0.55	0.05	0.70	0.63	0.29	1.00	0.57	0.67	0.14	0.84
X_9	0.48	0.47	0.39	0.07	0.65	0.61	0.41	0.57	1.00	0.60	0.20	0.56
X_{10}	0.58	0.47	0.45	0.01	0.92	0.47	0.26	0.67	0.60	1.00	0.15	0.53
X_{11}	0.16	0.13	0.07	0.12	0.13	0.14	0.42	0.14	0.20	0.15	1.00	0.14
X_{12}	0.71	0.65	0.66	0.04	0.60	0.65	0.28	0.84	0.56	0.53	0.14	1.00

表4　　模糊等价矩阵 t(R)

样本	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
X_1	1.00	0.72	0.66	0.15	0.70	0.72	0.41	0.71	0.65	0.70	0.41	0.71
X_2	0.72	1.00	0.66	0.15	0.70	0.72	0.41	0.71	0.65	0.70	0.41	0.71
X_3	0.66	0.66	1.00	0.15	0.66	0.66	0.41	0.66	0.65	0.66	0.41	0.66
X_4	0.15	0.15	0.15	1.00	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
X_5	0.70	0.70	0.66	0.15	1.00	0.70	0.41	0.70	0.65	0.92	0.41	0.70
X_6	0.72	0.72	0.66	0.15	0.70	1.00	0.41	0.71	0.65	0.70	0.41	0.71
X_7	0.41	0.41	0.41	0.15	0.41	0.41	1.00	0.41	0.41	0.41	0.42	0.41
X_8	0.71	0.71	0.66	0.15	0.70	0.71	0.41	1.00	0.65	0.70	0.41	0.84
X_9	0.65	0.65	0.65	0.15	0.65	0.65	0.41	0.65	1.00	0.65	0.41	0.65
X_{10}	0.70	0.70	0.66	0.15	0.92	0.70	0.41	0.70	0.65	1.00	0.41	0.70
X_{11}	0.41	0.41	0.41	0.15	0.41	0.41	0.42	0.41	0.41	0.41	1.00	0.41
X_{12}	0.71	0.71	0.66	0.15	0.70	0.71	0.41	0.84	0.65	0.70	0.41	1.00

2.3　求出不同 λ 值的布尔矩阵 $t(R)_\lambda$　λ 依次从 0.92、0.84、0.72、0.71、0.70、0.66、0.65、0.42、0.41、0.15 取值，分别求 $t(R)_\lambda$。

如，当 $\lambda=0.65$ 时，有

$$t(R)_{0.65}=\begin{bmatrix}1&1&1&0&1&1&0&1&1&1&0&1\\&1&1&0&1&1&0&1&1&1&0&1\\&&1&0&1&1&0&1&1&1&0&1\\&&&1&0&0&0&0&0&0&0&0\\&&&&1&1&0&1&1&1&0&1\\&&&&&1&0&1&1&1&0&1\\&&&&&&1&0&0&0&0&0\\&&&&&&&1&1&1&0&1\\&&&&&&&&1&1&0&1\\&&&&&&&&&1&0&1\\&&&&&&&&&&1&0\\&&&&&&&&&&&1\end{bmatrix}$$

即，X 分为 $\{X_1, X_2, X_3, X_5, X_6, X_8, X_9, X_{10}, X_{12}\}$，$\{X_4\}$，$\{X_7\}$，$\{X_{11}\}$ 4 类。其他同理可求。

2.4　画出聚类图　聚类图见图1。

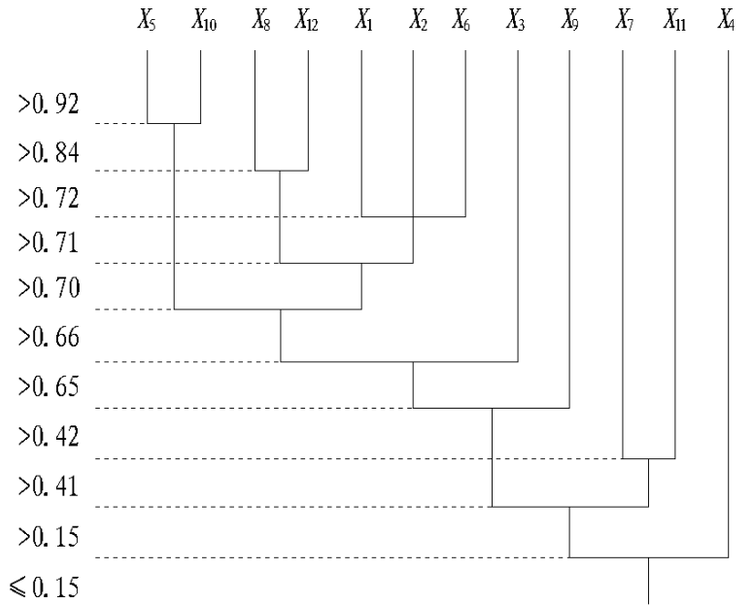

图1　甘蔗种植户聚类图

根据具体情况，可指定一个适当的 λ 值，即得到相应的聚类。如，当 $\lambda=0.66$ 时，X 可分成 $\{X_1, X_2, X_3, X_5, X_6, X_8, X_{10}, X_{12}\}$，$\{X_4\}$，$\{X_7\}$，$\{X_9\}$，$\{X_{11}\}$ 5 类。

3　结语

该文利用模糊聚类技术实现了甘蔗种植户的分类，获得具有不同种植行为特征的甘蔗种植户群组。这有助于制糖企业对各群组作进一步分析，并且制定相应的关怀策略，充分调动甘蔗种植户的积极性，预防甘蔗种植面积流失，为提高我国糖业的国际竞争力创造条件。

参考文献

[1] 李剑峰. 数据挖掘在公司财务分析中的应用[J]. 计算机工程与应用，2005(2):217-219.
[2] 陈安. 数据挖掘技术及应用[M]. 北京：科学出版社，2006.
[3] 蔡秀娟. 模数据挖掘在高校学生管理中的应用[J]. 华南农业大学学报，2006(3):143-147.
[4] 彭云. 基于模糊集的银行个人客户聚类技术[J]. 计算机工程与设计，2006(12):4674-4676.