区间值的聚类方法研究及应用

陈东升[1]　张利利[2]

（1. 郑州轻工业学院数学与信息科学系　河南郑州 4500021；2. 黄河科技学院信息工程学院　河南郑州 450006；）

摘要：基于区间值的模糊聚类分析方法，可以更大程度地保留信息。本文建立了三种相似系数的构造方法，在获取对象的特征指标值为区间值时，利用最大最小法构造相似系数并进行聚类，通过实例说明区间值退化为点时的聚类过程。

关键词：直接聚类；相似系数；最大最小法

The Research and Application of Interval Value Clustering

ChenDongSheng[1]　ZhangLiLi [2]

(1. Department of Information and Computation Science,Zhengzhou Institute of Light Industry , Zhengzhou 450002;2.Information Technology Institute, Huanghe S&T College,Zhengzhou,450006.)

Abstract：Fuzzy clustering analysis based on interval value can preserve information in a more great extent. In this paper， three methods of getting similarity coefficient are proposed， then when the data of expressing characters of things is interval value， the method of fuzzy clustering analysis based on maximum-minimum method is given ，At last， the new clustering method as proposed are applied to practice in interval value degenerating point.

Key words：direct clustering method; Possibility Degree; Maximum-minimum Method;

1. 引言

聚类[1]是作为一种基本的数据挖掘方法，广泛应用于相似搜索、模糊识别、趋势分析等领域，其中聚类算法在金融投资、地理信息系统、卫星图像和信息检索等领域[2]也有着广泛的应用。在许多实际问题中，我们得到的表征事物特征的数据往往不是一些确切的有效数值，而是一些区间值。本文给出相似系数的三种构造方法：距离倒数法、距离补法和最大最小法，其中用距离倒数法、距离补法构造的相似矩阵以数值为元素，可以利用[3]中的方法进行模糊聚类；其中用最大最小法构造的相似矩阵以区间值为元素，本文提出了基于最大最小法的区间值聚类分析方法，最后，举例说明了所给出的聚类方法在实际中的应用。

2.1 区间值的加权标准化

2. 基于区间值多重指标的直接聚类法

设具有区间值多重指标信息的聚类分析问题的聚类对象集为，特征指标集为 $X=\{x_1,x_2,\cdots,x_n\}$，$Q=\{Q_1,Q_2,\cdots,Q_m\}$，对第 i 个样本的第 k 个指标进行测度，得到 x_i 关于 Q_k 的特征指标值为区间值形式，

$$\tilde{x}_{ik}{}' = [x_{ik}{}'^{-}, x_{ik}{}'^{+}] \in I(R^+)$$

$$\text{即} = \left\{ \tilde{a} \,\middle|\, \tilde{a} = [a^-, a^+], a^-, a^+ \in R \right\}.$$

设指标 Q_j 的权重为区间值形式，记为

基金项目：河南省科技计划基金资助项目（0413031920）；河南省教育厅基金资助项目（2008A110021）
Foundation: The Science and Technology Plan of Henan Province(0413031920); Supported by Foundation of He'nan Educationl Committee(2008A110021)

978-1-935068-01-3 © 2009 SciRes.

$$\tilde{w}_j=[w_j^-,w_j^+]\in I[0,1]$$
$$=\left\{\tilde{a}\,\middle|\,\tilde{a}=[a^-,a^+],a^-,a^+\in[0,1]\right\},\ j=1,2,\cdots,m,$$

将 x_{ij} 加权标准化为 $\tilde{x}_{ij}$，$i=1,2,\cdots,n$；$j=1,2,\cdots,m$，其中

$$\tilde{x}_{ij}=[\bar{x}_{ij},x_{ij}^+]=\tilde{w}_j\tilde{x}_{ij}'=[\bar{w}_j\tilde{x}_{ij}'^-,w_j^+\tilde{x}_{ij}'^+]$$

2.2 相似系数的定义及性质

令相似矩阵 $\boldsymbol{R}=(\tilde{r}_{ij})_{n\times n}$，用以下方法确定 x_i 与 x_j 的相似系数 $\tilde{r}_{ij}$

1)距离倒数法

$$\tilde{r}_{ij}=\begin{cases}\dfrac{M}{\sum_{k=1}^{m}\left|\tilde{x}_{ik}^{+}-\tilde{x}_{jk}^{+}\right|+\left|\tilde{x}_{ik}^{-}-\tilde{x}_{jk}^{-}\right|} & \tilde{x}_{ik}\neq\tilde{x}_{jk}\\ 1 & \tilde{x}_{ik}=\tilde{x}_{jk}\end{cases}$$

2)距离补法

$$\tilde{r}_{ij}=1-Md(\tilde{x}_{ik},\tilde{x}_{jk})$$

其中

$$d(\tilde{x}_{ik},\tilde{x}_{jk})=\sqrt{\sum_{i=1}^{m}\left(\left|\tilde{x}_{ik}^{+}-\tilde{x}_{jk}^{+}\right|^2+\left|\tilde{x}_{ik}^{-}-\tilde{x}_{jk}^{-}\right|^2\right)},$$

M 适当选取，使得 $\tilde{r}_{ij}\in I[0,1]$.

3）最大最小法

$$\tilde{r}_{ij}=\begin{cases}1 & i=j\\ M\sum_{k=1}^{m}\dfrac{\tilde{x}_{ik}\wedge\tilde{x}_{jk}}{\tilde{x}_{ik}\vee\tilde{x}_{jk}} & i\neq j\end{cases}$$

其中，当可能度[4] $P_{\tilde{x}_{ik}>\tilde{x}_{jk}}\geq 0.5$ 时，$\tilde{x}_{ik}\wedge\tilde{x}_{jk}=\tilde{x}_{jk}$，$\tilde{x}_{ik}\vee\tilde{x}_{jk}=\tilde{x}_{ik}$，当 $P_{\tilde{x}_{ik}>\tilde{x}_{jk}}<0.5$ 时，$\tilde{x}_{ik}\wedge\tilde{x}_{jk}=\tilde{x}_{ik}$，$\tilde{x}_{ik}\vee\tilde{x}_{jk}=\tilde{x}_{jk}$，$M$ 适当选取，使得 $\tilde{r}_{ij}\in I[0,1]$。

若用距离倒数法和距离补法构造相似系数 $\tilde{r}_{ij}$，显然 $\boldsymbol{R}$ 是以数值为元素的矩阵，且有以下性质：

(1) $\boldsymbol{R}>E$ （自反性）；

(2) $\boldsymbol{R}^T=\boldsymbol{R}$ (对称性)。故 $\boldsymbol{R}=(\tilde{r}_{ij})_{n\times n}$ 是相似矩阵，可以利用原有的模糊聚类方法[3][7]。

若用最大最小法构造相似系数 $\tilde{r}_{ij}$，则显然 $\boldsymbol{R}$ 是以区间值为元素的矩阵，由[4]中的区间值的大小，比较可知 $\boldsymbol{R}$ 有以下性质：(1) $\boldsymbol{R}>E$ （自反性）；(2) $\boldsymbol{R}^T=\boldsymbol{R}$ (对称性)，故 $\boldsymbol{R}=(\tilde{r}_{ij})_{n\times n}$ 是相似矩阵。下面给出基于最大最小法的区间值的直接聚类法。

2.3 基于区间值的三种直接聚类法

对传统的聚类分析，文[3]提出了三种直接聚类法：直接聚类法、最大树法和编网法。现在把这三种方法推广，可得到相应的基于区间值的直接聚类法。下面给出基于区间值的编网法聚类的步骤。

1.编网。确定水平截割水平 $\tilde{\lambda}=[\lambda^-,\lambda^+]$，在模糊相似矩阵 $\boldsymbol{R}(\tilde{r}_{ij})$ 的主对角线上填入对象的符号。在主对角线的下方，若 $\tilde{r}_{ij}\geq\tilde{\lambda}$，则将 $\tilde{r}_{ij}$ 用“*”代替；若 $\tilde{r}_{ij}<\tilde{\lambda}$，则将 $\tilde{r}_{ij}$ 用空格代替。再由“*”所在的位置向上引纵线，向右引横线。

2.分类。把经过“*”可以联系的点归为一类，即可得到 $\tilde{\lambda}$ 水平上的等价分类。

3. 区间值聚类方法的应用举例

表 1

聚类对象	x_1	x_2	x_3	x_4	x_5
Q_1	[10,20]	[15,20]	[10,15]	[10,20]	[10,15]
Q_2	[12,16]	[10,15]	[12,18]	[18,20]	[10,20]
Q_3	[45,50]	[20,45]	[10,25]	[20,25]	[30,50]

表 2

聚类对象	x_1	x_2	x_3	x_4	x_5
Q_1	[0.05, 0.2]	[0.075, 0.2]	[0.05, 0.15]	[0.05, 0.2]	[0.05, 0.15]
Q_2	[0.3, 0.56]	[0.25, 0.525]	[0.3, 0.63]	[0.45, 0.7]	[0.25, 0.7]
Q_3	[0.18, 0.3]	[0.08, 0.27]	[0.04, 0.15]	[0.08, 0.15]	[0.12, 0.3]

例 1 假设由 5 个被聚类对象（x_1,x_2,x_3,x_4,x_5），3 个特征指标（Q_1,Q_2,Q_3），且这些指标值均为效益型指标，权重分别为 $\tilde{w}_1=[0.1,0.2],\tilde{w}_2=[0.5,0.7],\tilde{w}_3=[0.2,0.3]$, 已知原始的特征值如表 1。

1)用本文 2.1 的方法对表 1 中的区间值加权标准化得表 2。

2）由本文 2.2 中区间值的最大最小法构造相似系数（其中 $M=0.1429$），可得对称的相似矩阵 $\boldsymbol{R}(\tilde{r}_{ij})$ 为

$$\begin{pmatrix}1 & [0.138,0.833] & [0.123,0.814] & [0.135,0.868] & [0.154,0.986]\\ & 1 & [0.114,0.804] & [0.129,0.816] & [0.125,0.907]\\ & & 1 & [0.135,0.896] & [0.128,0.967]\\ & & & 1 & [0.125,0.829]\\ & & & & 1\end{pmatrix}$$

3）用文[4]中的方法对区间值 $\tilde{r}_{ij}$ 排序。求出相似系数 $\tilde{r}_{ij}$ 两两比较的可能度矩阵 $\boldsymbol{P}$ 如下：

	$\tilde{r}_{12}$	$\tilde{r}_{13}$	$\tilde{r}_{14}$	$\tilde{r}_{15}$	$\tilde{r}_{23}$	$\tilde{r}_{24}$	$\tilde{r}_{25}$	$\tilde{r}_{34}$	$\tilde{r}_{35}$	$\tilde{r}_{45}$
$\tilde{r}_{12}$	0.5	0.5123	0.4888	0.4447	0.5191	0.5094	0.4794	0.4794	0.4596	0.5061
$\tilde{r}_{13}$	0.4877	0.5	0.4768	0.4334	0.5069	0.4971	0.4678	0.4676	0.4484	0.4939
$\tilde{r}_{14}$	0.5112	0.5232	0.5	0.4562	0.5299	0.5204	0.4904	0.4906	0.4707	0.5170
$\tilde{r}_{15}$	0.5553	0.5666	0.5438	0.5	0.5729	0.5642	0.5335	0.5342	0.5135	0.5605
$\tilde{r}_{23}$	0.4809	0.4931	0.4701	0.4271	0.5	0.4902	0.4613	0.4611	0.4421	0.4871
$\tilde{r}_{24}$	0.4906	0.5029	0.4796	0.4358	0.5098	0.5	0.4704	0.4703	0.4509	0.4968
$\tilde{r}_{25}$	0.5206	0.5322	0.5096	0.4665	0.5387	0.5296	0.5	0.5003	0.4806	0.5262
$\tilde{r}_{34}$	0.5206	0.5324	0.5094	0.4658	0.5389	0.5297	0.4997	0.5	0.4800	0.5263
$\tilde{r}_{35}$	0.5404	0.5516	0.5293	0.4865	0.5579	0.5491	0.5194	0.5200	0.5	0.5457
$\tilde{r}_{45}$	0.4939	0.5061	0.4830	0.4395	0.5129	0.5032	0.4738	0.4737	0.4543	0.5

令 $r_{ij}(1 \le i < j \le 5)$ 表示 $\tilde{r}_{ij}$ 大于 $\tilde{r}_{12}$，$\tilde{r}_{13}$，$\tilde{r}_{14}$，$\tilde{r}_{15}$，$\tilde{r}_{23}$，$\tilde{r}_{24}$，$\tilde{r}_{25}$，$\tilde{r}_{34}$，$\tilde{r}_{35}$，$\tilde{r}_{45}$ 可能度的和，由矩阵 $\boldsymbol{P}$ 得表 3。

表 3

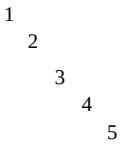
r_{12}	r_{13}	r_{14}	r_{15}	r_{23}	r_{24}	r_{25}	r_{34}	r_{35}	r_{45}
4.8	4.7	5.0	5.4	4.7	4.8	5.1	5.1	5.2	4.8

由表 3 可得

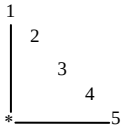
$r_{23} < r_{13} < r_{24} < r_{45} < r_{12} < r_{14} < r_{34} < r_{25} < r_{35} < r_{15}$，故 $\tilde{r}_{23} < \tilde{r}_{13} < \tilde{r}_{24} < \tilde{r}_{45} < \tilde{r}_{12} < \tilde{r}_{14} < \tilde{r}_{34} < \tilde{r}_{25} < \tilde{r}_{35} < \tilde{r}_{15}$.

4)用编网法聚类

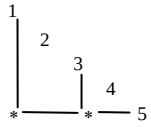
若取 $\tilde{\lambda} > \tilde{r}_{15}$，得编网图 1，可得 $\tilde{\lambda}$ 上的水平分类 $\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\}$；若取 $\tilde{r}_{15} \ge \tilde{\lambda} > \tilde{r}_{35}$，得编网图 2，可得 $\tilde{\lambda}$ 上的水平分类 $\{x_1,x_5\},\{x_2\},\{x_3\},\{x_4\}$；若取 $\tilde{r}_{35} \ge \tilde{\lambda} > \tilde{r}_{25}$，得编网图 3，可得 $\tilde{\lambda}$ 上的水平分类 $\{x_1,x_3,x_5\},\{x_2\},\{x_4\}$；若取 $\tilde{r}_{25} \ge \tilde{\lambda} > \tilde{r}_{34}$，得编网图 4，可得 $\tilde{\lambda}$ 上的水平分类 $\{x_1,x_2,x_3,x_5\},\{x_4\}$；若取 $\tilde{\lambda} \le \tilde{r}_{34}$，可得 $\tilde{\lambda}$ 上的水平分类 $\{x_1,x_2,x_3,x_4,x_5\}$。



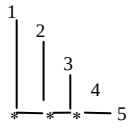
编网图 1



编网图 2



编网图 3



编网图 4

例 2 文[6]提出按科研规模划的大小,将我国的大学分为研究型、研究教学型、教学研究型、教学型等四型。我们选取 10 所交通院校,它们的评价数据及类型如表 4。

对表 4 中的 10 所学校在教学(人才培养)、科研两方面分别进行综合评价，得到他们的教学、科研得分及总分。用 $\tilde{x}_{i1}', \tilde{x}_{i2}'$(i = 1,2, …,10) 分别表示上述 10 所学校的教学得分、科研得分。显然，本例中的指标值均为退化区间值形式。我们假设教学、科研的权重相同。

教学和科研均为效益性指标，故可令 $\tilde{x}_{i1} = \dfrac{\tilde{x}_{i1}'}{x_1^{\max}}$，$\tilde{x}_{i2} = \dfrac{\tilde{x}_{i2}'}{x_2^{\max}}$，$i = 1,2,\cdots,10$，

其中

$x_1^{\max} = \max\{x_{11}'^{+}, x_{21}'^{+}, \cdots, x_{10,1}'^{+}\} = 58.61$

$x_2^{\max} = \max\{x_{12}'^{+}, x_{22}'^{+}, \cdots, x_{10,2}'^{+}\} = 84.42$。可得标准化聚类对象特征指标值 $\tilde{x}_{ij}(i = 1,2,\cdots,10; j = 1,2)$ 矩阵如下：

聚类对象	x_1	x_2	x_3	x_4	x_5
教学	1	0.7371	0.2119	0.1643	0.0795
科研	1	0.5849	0.1173	0.1225	0.0117

聚类对象	x_6	x_7	x_8	x_9	x_{10}
教学	0.0519	0.0601	0.0534	0.0450	0.0345
科研	0.0180	0.0049	0.0062	0.0072	0.0017

利用本文 2.2 中的区间值的最大最小法构造相似系数（其中 $M = 0.57$），可得相似矩阵 $\boldsymbol{R}(\tilde{r}_{ij})$ 如下：

1	0.7535	0.1876	0.1635	0.0520	0.0398	0.0371	0.0340	0.0298	0.0207
	1	0.2782	0.2465	0.0729	0.0577	0.0512	0.0473	0.0418	0.0283
		1	0.9878	0.2707	0.2271	0.1854	0.1737	0.1560	0.1011
			1	0.3302	0.2638	0.2313	0.2141	0.1896	0.1276
				1	0.7426	0.6696	0.6849	0.6734	0.3301
					1	0.6474	0.7503	0.7222	0.4327
						1	0.9569	0.8147	0.5249
							1	0.9712	0.5246
								1	0.5716
									1

978-1-935068-01-3 © 2009 SciRes.

表 4

校名	排名	总分	人才培养			科学研究			学校参考类型
			得分	研究生	本科生	得分	自然科学	社会科学	
上海交通大学 (x_1)	4	140.03	58.61	43.50	15.11	84.42	77.72	6.70	研究型
西安交通大学 (x_2)	10	92.58	43.20	26.45	16.75	49.38	37.03	12.35	研究型
西南交通大学 (x_3)	55	22.32	12.42	5.14	7.28	9.90	8.18	1.72	研教型
北京交通大学 (x_4)	60	19.97	9.63	5.47	4.15	10.34	9.30	1.05	研教型
兰州交通大学 (x_5)	207	5.64	4.66	0.44	4.21	0.99	0.88	0.11	教研型
石家庄铁道学院 (x_6)	246	4.55	3.04	0.45	2.58	1.52	1.42	0.09	教学型
华东交通大学 (x_7)	284	3.94	3.52	0.14	3.39	0.41	0.29	0.12	教学型
重庆交通大学 (x_8)	309	3.65	3.13	0.91	2.94	0.52	0.40	0.12	教研型
大连交通大学 (x_9)	349	3.25	2.64	0.21	2.43	0.61	0.55	0.07	教研型
山东交通学院 (x_{10})	458	2.16	2.02	0.00	2.02	0.14	0.13	0.01	教学型

由于本例中的相似矩阵 $\boldsymbol{R}(\tilde{r}_{ij})$ 中的元素均为退化区间值，故可直接得各元素的大小排序。取 $\lambda=0.74$，由一般数值的编网方法，可得将这 10 所交通院校分为四类

$$\{x_1,x_2\},\{x_3,x_4\},\{x_5,x_6,x_7,x_8,x_9\},\{x_{10}\}。$$

从表 4 可以看出，x_6 在教学方面比 x_5 少 1.62 分，在科研方面比 x_5 多 0.53 分，但是在文[9]中，作者却将 x_6 划分为教学型，x_5 划分为教研型；x_8 和 x_9 教学方面相差 0.49 分，科研方面相差 0.09 分，总分相差 0.4 分，它们同为教研型，而 x_8 和 x_7 教学方面相差 0.39 分，科研方面相差 0.11 分，总分相差 0.29 分，它们却被为分成两种类型。这种分法显然是不合理的。

其实，若要把这 10 所学校分为研究型，研教型，教研型和教学型四类，当 $\lambda=0.74$，$\{x_1,x_2\},\{x_3,x_4\},\{x_5,x_6,x_7,x_8,x_9\},\{x_{10}\}$ 这种分类是比较合理的。其中，$\{x_1,x_2\}$ 研究型，$\{x_3,x_4\}$ 为研教型，$\{x_5,x_6,x_7,x_8,x_9\}$ 为教研型，$\{x_{10}\}$ 为教学型。

4 结束语

由于信息的不完全性,我们得到的表示事物特征的数值往往是一些区间值。基于区间值的模糊聚类分析方法，可以更大程度地保留信息。本文给出基于区间值的相似系数的构造方法，并给出基于区间值的聚类方法。从应用角度来讲，这些方法简单，易于理解和操作，可以有效解决通信领域中的分类问题。

用距离倒数法、距离补法构造的相似矩阵以数值为元素，可以利用原有的方法进行模糊聚类，计算简单，但分类比较粗糙，在某些要求分类不是特别精细的系统中采用此法，可以收到事半功倍的效果；用最大最小法构造的相似矩阵以区间值为元素，本文提出了基于最大最小法的区间值直接聚类方法。并且由本文的例 2 可知，当所有的区间值退化为一个实数时，与原有的模糊类分析方法是一致的[3][7]。从理论上来讲，这些方法也可看成是对原来的模糊聚类分析方法的一个推广。但是，如何将基于区间值的信息处理方法应于数据挖掘等诸多领域，仍然是值得我们研究的一个课题。

References （参考文献）

[1] 史忠植. 知识发现[M]. 北京:清华大学出版社, 2002.

[2] 陈东升. FUZZY 图最大树聚类方法及其应用[J], 运筹与管

理, 2007, 16(3):69-73.

[3] 谢季坚, 刘承平. 模糊数学方法及其应用[M]. 武汉:华中科技大学出版社, 2000, 5.

[4] 徐泽水, 达庆利. 区间数的排序方法研究[J]. 系统工程, 2001, 19 (6) : 94～ 96.

[5] 于春海, 樊治平. 一种基于区间数多指标信息的聚类方法[J]. 东北大学大学学报, 2004, 25(2):183-186.

[6] 武书连. 挑大学选专业[M]. 北京:中国统计出版社, 2008.

[7] 陈东升, 模糊聚类方法在高校分类中的应用[J]. 数学的实践与认识, 2005, 35(4):75-81.

978-1-935068-01-3 © 2009 SciRes.