doi:10.16055/j.issn.1672-058X.2015.0011.004

邻域系统的一种粒化方法及应用

罗 来 鹏

(华东交通大学 理学院,南昌 330013)

摘 要:邻域系统是一类具有兼容性与应用性的系统,关于它的粒化一般采用相似度加阈值的方法.基于该方法在阈值确定上的主观性以及不确定计算上结果复杂性,提出一种模糊聚类加统计量的方法,该方法在计算上不仅仍保留 Pawlak 粗糙集系统一些不确定性度量性质,而且在阈值的确定上更为客观;最后示例说明了该方法的有效性.

关键词:邻域系统;粗糙集;粒计算;模糊聚类

中图分类号:TP311.32 **文献标志码**:A **文章编号**:1672-058X(2015)11-0018-04

粒计算是近些年发展起来的一种信息处理方法,是人工智能研究的一个热点,它融合了模糊集理论、粗糙集理论、商空间理论等新型智能信息处理理论,在模式识别、数据挖掘、图像处理等邻域得到了广泛应用,极大地推动了信息科学的发展.粗糙集[1]是 1965 年由波兰学者 Z.Pawlak 提出的一种处理不确定性问题的数学理论,是目前三大粒计算模型之一,在信息科学、管理科学等领域有着广泛应用.它以等价关系为基础,建立论域分类,在此基础上通过计算分类与概念之间的集合关系对数据库进行属性约简,同时利用集合的上、下近似关系获取分类规则.等价关系以及由此决定的等价类是 Pawlak 粗糙集的基础性内容.等价类构成这个系统的基本知识,因而粗糙知识具有粒度性.近几年从粒的角度来理解和发展粗糙集取得了很多成果[2,3].在现实中,由于数据的复杂性、不确定性、不完备性,很多数据很难满足 Pawlak 粗糙集条件,也就是说,很多信息系统很难从等价关系角度建立系统的基本知识.为此,很多学者对 Pawlak 粗糙集进行推广,比如邻域系统就是一种更加贴近实际应用的一种系统.由于邻域系统的属性是连续的,通过等价类建立它的基本知识意义不大,因此开展邻域系统的粒化工作是粗糙集的一个值得研究重要问题,近一些年取得很多成果[4-10].纵观这些方法,发现研究的出发点是从原来的等价关系变为邻域关系,将等价类变为邻域来刻画.此时,一个对象的邻域怎么来计算就显得很重要.大多数情况下是根据相似度加阈值的方法来计算一个对象的邻域,但是该方法阈值选取具有很大的主观性和随机性,而且相似性不满足传递关系,使得计算一个集合的近似刻画结构不清晰.基于这些,此处采用聚类方法计算一个对象的邻域,并通过引入一个统计量的方法计算最佳的阈值方法.

1 Pawlak 粗糙集[1]

定义 1[1] 一个知识表示系统 S 是一个四元组 $S=(U,R,V,F)$,其中 U 是有限对象的集合,$R=A\cup D$ 是有限个属性的非空集合,V 是属性值的集合,F 为信息函数 $F:U\times R\to V$.当子集 A 和 D 分别为条件属性和决策

收稿日期:2015-05-04;**修回日期**:2015-06-10.

作者简介:罗来鹏(1973-),男,江西吉水人,副教授,硕士,从事粗糙集与粒计算研究.

属性时,知识表达系统又称为决策表.对于任何 $B\subseteq R$,都可以决定一个等价关系,也就决定论域上的一个划分,每一个划分小类称为等价类,也称为该系统的一个粒,所有等价类集合表示为 U/B,记为 $U/B=\{X_1,X_2,\cdots,X_m\}$,其中 $X_i\cap X_j=\varnothing,i\neq j,i,j=1,\cdots,m$;并且 $\bigcup_{i=1}^{m}X_i=U$,因此有时用(U,R)表示 Pawlak 近似空间.

定义 2　设(U,R)为 Pawlak 近似空间,对于任意 $X\subseteq U$,有下列定义:$\underline{R}(X)=\{x\in U|:[x]_R\subseteq X\}$.$\overline{R}(X)=\{[x]_R\cap X\neq\varnothing\}$分别称为 X 在近似空间(U,R)的上、下近似集.$\alpha_R(X)=\dfrac{|\underline{R}(X)|}{|\overline{R}(X)|}$称为 X 关于(U,R)的精度,精度越高,X 在关系 R 所决定的基本知识上确定性越高.

Pawlak 粗糙集建立在等价关系上,要求数据是离散型的,而在实际应用中数据比这个复杂得多,比如连续型数据就是一种非常普遍的信息系统,对于这类系统如果从等价类角度来建立基本知识显然不可取,意义不大.因此为了更好拓广应用范围,必须将 Pawlak 粗糙集进行拓展,将等价关系拓展为更为一般的关系.

2　邻域粗糙集

定义 3[4,5]　设 $I=(U,R,V,F,\delta)$为一个邻域信息系统,U 是有限对象的集合,R 为属性集,V 为属性值域,F 为对象在属性上的映射,δ 为邻域阈值,并且 $0\leqslant\delta\leqslant1$,如果 $R=C\cup D$,C 表示条件属性,D 表示决策属性,那么邻域系统又称为邻域决策系统.

定义 4[4,5]　设 $I=(U,R,V,F,\delta)$为一个邻域信息系统,$B\subseteq R$,对于任意 $x\in U$,可定义在 I 上的邻域为 $N_B^{\delta}(x)=\{y|y\in U,D(x,y)<\delta\}$,其中 $D(x,y)$为邻域的计算公式,在实际应用中,可根据具体情况进行合理定义,比如相似度、距离等.显然在一个邻域系统中,一个对象邻域的大小与的 δ 取值有直接关系,δ 越大邻域也越大.

定义 5[4,5]　设 $I=(U,R,V,F,\delta)$为一个邻域信息系统,对于任意 $X\subseteq U$,有下列类似定义:$\underline{R}(X)=\{x\in U|N_R^{\delta}(x)\subseteq X\}$,$\overline{R}(X)=\{N_R^{\delta}(x)\cap X\neq\varnothing\}$分别称为 X 在近似空间(U,R,δ)的上、下近似集.$\alpha_R(X)=\dfrac{|\underline{R}(X)|}{|\overline{R}(X)|}$称为$X$ 关于(U,R,δ)的精度.

3　等价关系与邻域关系

等价关系满足自反性、对称性、传递性,而邻域关系满足自反性、对称性,传递性未必能满足.当 $\delta=0$ 时,邻域关系就退化为等价关系,因此等价关系只是邻域关系的一种特殊情形,而邻域关系是在 Pawlak 系统基础上建立的一种新的更具有普遍性的关系,这种关系极大丰富了粗糙集的理论与应用研究.Pawlak 粗糙集等价类构成论域上的一个划分,而邻域粗糙集邻域构成论域上的一个覆盖.正因为这样,计算一个概念在邻域系统上的上、下近似和精度比 Pawlak 的系统要复杂.

在邻域系统中,阈值 δ 的大小决定一个对象的邻域,从而也就影响到一个概念在邻域系统的上、下近似集以及系统分类精度等问题,通常情况下,δ 越小,邻域系统粒化的粒就越大,δ 越大,邻域系统的粒就越小,显然这两种粒化所得到结果都不能很好刻画系统真实本身,都不利于从粒度角度研究粗糙邻域系统,比如邻域分类精度问题.因此,如何更为客观地确定 δ 的大小,对于一个邻域系统的粒化及不确定度量非常重要.为此,引入模糊聚类加统计量的方法对 δ 的取值进行优化,从而得到系统更好的粒化结果.该方法主要是根据统计学中方差分析理论,对不同的 δ 进行评价,最终得到一个较合理的值.

4 模糊聚类

聚类是数据挖掘中比较重要的一种技术,它是基于所讨论对象相似性大小的一种非监督学习方法,该方法在模式识别等很多领域都有广泛的应用.模糊聚类是一种根据对象的特征,通过建立相似矩阵,计算等价矩阵,最后根据阈值得出聚类结果的一种动态聚类技术.它一般要求数据对象的特征是数值连续型,因此将模糊聚类引入到邻域系统的粒化研究是完全可以的.模糊聚类一般分4步:数据预处理;相似矩阵的建立;模糊等价矩阵的计算;根据阈值,由等价矩阵得出聚类结果.

类中元素相似度大,而类之间元素相异度大,说明分类显著,为此引一个统计量.

设论域 $U=\{x_1,x_2,\cdots,x_n\}$,每个对象 x_i 用 p 维向量进行特征刻画,即记为 $x_i=(c_{i1},c_{i2},\cdots,c_{ip})$,论域 U 的平均中心特征向量记为 $\bar{c}^{(U)}=(\bar{c}_1^{(U)},\bar{c}_2^{(U)},\cdots,\bar{c}_p^{(U)})$,则有 $\bar{c}_k^{(U)}=\frac{1}{n}\sum_{i=1}^{n}c_{ik}(1\leqslant k\leqslant p)$,将 U 划分成 r 类,第 $j(j\leqslant r)$ 类的对象数为 j_n,并表示为 $U_j=\{x_1^{(j)},x_2^{(j)},\cdots,x_{j_n}^{(j)}\}$,每个对象对应一个 p 维特征向量,记为 $x_i^{(j)}=(c_{i1}^{(j)},c_{i2}^{(j)},\cdots,c_{ip}^{(j)})$ $(i\leqslant j_n)$,U_j 的平均中心特征向量为 $\bar{c}^{(j)}=(\bar{c}_1^{(j)},\bar{c}_2^{(j)},\cdots,\bar{c}_p^{(j)})$,其中 $\bar{c}_k^{(j)}=\frac{1}{j_n}\sum_{i=1}^{j_n}c_{ik}^{(j)}$,$F=\frac{\sum_{j=1}^{r}j_n\|\bar{c}^{(j)}-c^{(U)}\|^2/(r-1)}{\sum_{j=1}^{r}\sum_{i=1}^{j_n}\|x_i^{(j)}-\bar{c}^{(j)}\|^2/(n-r)}$ 它是服从自由度为 $r-1,n-r$ 的 F 分布.分子表示类与类之间的距离,分母表示类内样本的距离,显然对于不同的分类,F 值越大,类与类之间的距离也就越大,分类效果就更为明显.这样就可以根据统计方差原理,确定一个比较好的阈值,也就是确定一种最佳的分类方法.

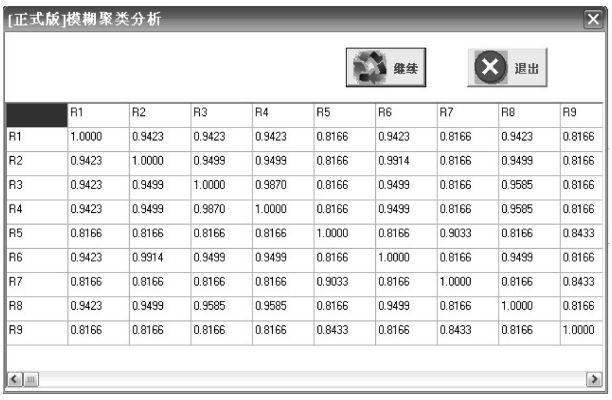
5 实例应用

表1是9个地点的水质检测情况,其中 $x_1,x_2,\cdots,x_9$ 表示9个检测地点,pH*,DO,CODMn,NH3-N为检测指标,显然这是一个邻域系统,下面对其进行粒化处理,所使用的方法是模糊聚类加参数阈值评价方法,具体结果如表1.

(1) 根据模糊聚类方法利用传递闭包得到等价矩阵如下(图1):

表1 水质测量

	pH*	DO	CODMn	NH3-N
四川攀枝花龙洞(x_1)	8.24	8.1	0.5	0.15
重庆朱沱(x_2)	7.73	8.55	1.8	0.2
江西九江河西水厂(x_3)	8.57	6.88	3	0.08
江苏南京林山(x_4)	7.59	6.47	2	0.12
四川乐山岷江大桥(x_5)	7.57	5.49	4.7	1.93
湖北丹江口胡家岭(x_6)	7.72	9.03	2.5	0.07
湖南长沙新港(x_7)	6.29	4.34	2.9	0.92
湖北武汉宗关(x_8)	7.91	5.87	3.6	0.25
江西南昌滁槎(x_9)	6.78	5.35	1.7	2.18



	R1	R2	R3	R4	R5	R6	R7	R8	R9
R1	1.0000	0.9423	0.9423	0.9423	0.8166	0.9423	0.8166	0.9423	0.8166
R2	0.9423	1.0000	0.9499	0.9499	0.8166	0.9914	0.8166	0.9499	0.8166
R3	0.9423	0.9499	1.0000	0.9870	0.8166	0.9499	0.8166	0.9585	0.8166
R4	0.9423	0.9499	0.9870	1.0000	0.8166	0.9499	0.8166	0.9585	0.8166
R5	0.8166	0.8166	0.8166	0.8166	1.0000	0.8166	0.9033	0.8166	0.8433
R6	0.9423	0.9914	0.9499	0.9499	0.8166	1.0000	0.8166	0.9499	0.8166
R7	0.8166	0.8166	0.8166	0.8166	0.9033	0.8166	1.0000	0.8166	0.8433
R8	0.9423	0.9499	0.9585	0.9585	0.8166	0.9499	0.8166	1.0000	0.8166
R9	0.8166	0.8166	0.8166	0.8166	0.8433	0.8166	0.8433	0.8166	1.0000

图1 等价矩阵

(2) 根据表1具体值粒化分类,有几种结果:当阈值 $\lambda=0.9$ 时,系统可以粒化为三类:$\{x_1,x_2,x_3,x_4,x_6,$

$x_8\}$,$\{x_5,x_7\}$,$\{x_9\}$;$\lambda=0.94$ 时,系统可以粒化为 4 类:$\{x_1,x_2,x_3,x_4,x_6,x_8\}$,$\{x_5\}$,$\{x_7\}$,$\{x_9\}$;$\lambda=0.95$时,系统可以粒化为 6 类:$\{x_1\}$,$\{x_2,x_6\}$,$\{x_3,x_4,x_8\}$,$\{x_5\}$,$\{x_7\}$,$\{x_9\}$.

(3) 将不同的 λ 值所对应的分类,设置显著性水平为 0.25 进行 F 统计方差分析,各临界值分别为 $F_{0.25}(\lambda=0.9)=7.16$,$F_{0.25}(\lambda=0.94)=7.76$,$F_{0.25}(\lambda=0.95)=14.9$,$F$ 检验值分别为 $F_{\lambda=0.9}=4.46$,$F_{0.94}=7.78$,$F_{0.95}=9.47$.

(4) 综合上述分析,当 $\lambda=0.94$ 时分类特别显著,因此系统最好的粒化结果应该为$\{x_1,x_2,x_3,x_4,x_6,x_8\}$,$\{x_5\}$,$\{x_7\}$,$\{x_9\}$.

6 结　论

主要讨论了邻域系统的一种粒化方法,该方法是根据以往常见的方法,针对阈值确定问题就如何提高它的有效性进行展开的,该方法比相似度加阈值方法具有更好的优越性,为进一步拓展邻域系统的粒化提供了另外一种思路.从应用实例来看,所得结果基本反映实际情况.

参考文献:

[1] PAWLAK Z.Rough Sets[J].International Journal of Computer and Information Science,1982,11(5):341-356

[2] 郭翠峰,胡鹏,胡展阂.区间值模糊目标信息系统的规则提取[J].重庆工商大学学报:自然科学版,2011,28(5):509-512

[3] 孙文.基于双论域的一般多粒度模糊粗糙集[J].重庆工商大学学报:自然科学版,2015,32(3):12-15

[4] 唐朝辉,陈玉明.邻域系统的不确定性度量方法[J].控制与决策,2014,29(4):691-695

[5] 杨习贝,杨静.邻域系统粗糙集模型[J].南京理工大学学报,2012,36 (2):291-295

[6] LIN G P,QIAN Y H,LI J J.NMGRS: Neighborhood-based Multigranulation Rough Sets[J].International Journal of Approximate Reasoning,2012(53):1080-1093

[7] WANG L J,YANG X B,YANG J Y,et al.Relationships among Generalized Rough Sets in Six Coverings and Pure Reflexive Neighborhood System[J].Information Sciences,2012(207):66-78

[8] CHEN Y M,WUA K SH,CHEN X H,et al.An Entropy-based Uncertainty Measurement Approach in Neighborhood Systems[J].Information Sciences,2014(279):239-250

[9] YANG X B,ZHANG M,DOU H L,et al.Neighborhood Systems-based Rough Sets in Incomplete Information System[J].Knowledge-based Systems,2011(24):858-867

[10] SYAU Y R,LIN E B.Neighborhood Systems and Covering Approximation Spaces[J].Knowledge-based Systems,2014(66):61-67

A Granulation Approach and Application in Neighborhood System

LUO Lai-peng

(School of Sciences,East China Jiaotong University,Nanchang 330013,China)

Abstract: Neighborhood system is a system with more compatibility and application.Its granulation generally adopts the method of similarity and the threshold.The main shortcomings of the method are the subjectivity of threshold and the complex of calculation results.In this paper, a new approach that adopts fuzzy clustering and statistics is supposed.The approach can not only keep some uncertainty measurement properties of Pawlak rough set system,but also is more objective in the threshold.The example shows the effectiveness of the approach.

Key words: neighborhood system; rough set; granular computing; fuzzy clustering