第34卷 第10期 计 算 机 工 程 2008年5月
Vol.34 No.10 Computer Engineering May 2008
·软件技术与数据库· 文章编号：1000—3428(2008)10—0086—03 文献标识码：A 中图分类号：TP182

基于聚类有效性分析的模糊粗糙集归纳学习方法

印 勇，孙如英
(重庆大学通信工程学院，重庆 400044)

摘 要：引入模糊 C 均值聚类算法进行连续属性模糊化，通过聚类有效性分析来确定最佳分类数目，克服了属性模糊化方法需要人为确定划分类数的缺点。用属性模糊化得到的属性隶属度矩阵约简模糊粗糙属性，由此提出一种基于模糊粗糙集的属性约简算法。实例验证了该方法的可行性和有效性。

关键词：模糊粗糙集；模糊聚类；有效性分析；决策表；属性约简

Inductive Learning Approach of Fuzzy-rough Set Based on Clustering Valid Analysis

YIN Yong, SUN Ru-ying
(College of Communication Engineering, Chongqing University, Chongqing 400044)

【Abstract】 Fuzzy C means clustering is introduced to fuzzify the continuous attribute, and the best minute class number is obtained by the valid analysis of clustering. It has overcome the disadvantage of determining artificially the class number for fuzzifing attribute approach. The attribute degree of membership matrix which obtained by attribute fuzzified is used to attributes reduction, and attributes reduction algorithm based on fuzzy rough sets is given. An example is illustrated to prove its feasibility and effectiveness.

【Key words】 fuzzy-rough set; fuzzy clustering; valid analysis; decision table; attribute reduction

1 概述

应用粗糙集理论进行知识获取的一个重要步骤是对信息系统进行约简处理。属性约简之前必须对连续属性进行离散化，由于离散化后的属性值没有保留属性值在实数值上存在的差异，这一过程将造成某种程度的信息损失。文献[1]为了解决粗糙集离散化过程中的信息损失问题，将模糊集理论引入粗糙集中，对信息系统中的对象不再进行离散化，而讨论对象间的关系时也用对象的相似关系而非粗糙集中的等价关系。正是由于模糊粗糙集理论引入的模糊概念易于保留连续属性值的信息，因此使用该理论处理数据集更能保留原始数据集所包含的信息。相关研究表明[2-4]，应用模糊粗糙集得到的模糊规则或基于案例的推理系统以及原始数据集的约简比粗糙集具有更高的准确度。但目前属性模糊化方法需要人为地规定划分的类数，几乎不考虑信息系统具体属性值的特征，方法往往过于主观、不够合理、可操作性不强。

2 模糊聚类分析

在模糊粗糙集的应用中，模糊等价类的划分(即模糊聚类分析)是必须考虑的问题。在粗糙集中，属性对应的等价类是普通集合，而在模糊粗糙集中，属性对应的等价类是模糊集，因此，往往把属性的等价类划分过程称为属性模糊化过程。在粗糙集中，每个对象属于且仅属于一个等价类，在模糊粗糙集中，每个对象可以属于多个模糊等价类。

2.1 模糊 C 均值聚类

令 $X=\{x_1,x_2,\cdots,x_n\}\subset R^p$ 是特征空间中的一个有限数据集合，其中，n是数据集合中的元素个数，c为样本的分类数，$2\le c\le n$；$R^{c\times n}$ 是所有实 $c\times n$ 矩阵的集合；$V=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_c\}\subset R^p$ 为特征空间 R^p 上的矢量集合，有c个聚类中心向量；μ_{ij} 是第j个样本属于第i个中心的隶属度；$\boldsymbol{U}=[\mu_{ij}]\in R^{c\times n}$ 是 1 个 $c\times n$ 矩阵。模糊C均值聚类分析的目标函数定义为

$$J_m(\boldsymbol{U},V,X)=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m\left\|x_j-\boldsymbol{v}_i\right\|_A^2 \tag{1}$$

(1) $\mu_{ij}\in[0,1],1\le i\le c,1\le j\le n$ ；

(2) $\sum_{i=1}^{c}\mu_{ij}=1,1\le j\le n$ ；

(3) $0<\sum_{j=1}^{n}\mu_{ij}<n,1\le i\le c$ 。

其中，

$$\boldsymbol{v}_i=\frac{\sum_{j=1}^{n}{u_{ij}}^m x_j}{\sum_{j=1}^{n}{u_{ij}}^m},1\le i\le c \tag{2}$$

$$\mu_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{1/m-1}} \tag{3}$$

$d_{ij}=\|x_j-\boldsymbol{v}_i\|_A^2$，$1\le i\le c,1\le j\le n$；$m\in[1,+\infty]$是一个加权指数(又称为平滑因子)，控制模式在模糊类之间的分享程度，m的一个经验范围为$1.1\le m\le 5$；J_m是类内误差的加权平方和目标函数。模糊 C 均值聚类算法通过式(2)和式(3)达到目标函数 J_m 的最小化，此时得到 X 的一个最优模糊 C 划分：$\boldsymbol{U}^*=[\mu_{ij}^*]$。

基金项目：重庆市应用基础研究基金资助项目(6976)
作者简介：印 勇(1963－)，男，副教授、博士，主研方向：数据挖掘，智能信息处理，图像处理；孙如英，硕士研究生
收稿日期：2007-07-08 **E-mail：**yy@ccee.cqu.edu.cn

2.2 聚类有效性分析

分析给定数据集的聚类结果是否合理属于聚类的有效性研究。对聚类分析而言，有效性问题又可以转化为最佳类别数的确定。本文运用基于可能性分布的聚类有效性函数确定各个属性划分的最佳类别数[5]，引入一个新的可能性划分系数$P(\boldsymbol{U};c)$，结合模糊划分系数$F(\boldsymbol{U};c)$和$P(\boldsymbol{U};c)$，定义一个新的聚类有效性函数。

对于给定的聚类中心数c和隶属矩阵$\boldsymbol{U}$,划分系数$F(\boldsymbol{U},c)$定义为

$$F(\boldsymbol{U};c)=\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}{}^{2} \tag{4}$$

其中，n为待分析的样本数据的个数。

对每一个样本x_j，有$\sum_{i=1}^{c}\mu_{ij}=1$。这可看成是对模糊C均值聚类算法的一个概率约束。$F(\boldsymbol{U};c)$也可写成

$$F(\boldsymbol{U};c)=\frac{1}{n}\sum_{j=1}^{n}(\sum_{i=1}^{c}\mu_{ij}^{2}/\sum_{i=1}^{c}\mu_{ij}) \tag{5}$$

从可能性分布的角度来看，$F(\boldsymbol{U};c)$可解释成：每一个样本点相对于c个聚类中心都有一个可能性分布，$F(\boldsymbol{U};c)$是n个可能性分布描述因子的平均值。对偶地，对每一个聚类中心，n个样本点的隶属度构成一个可能性分布。因此，对于给定的聚类数c和隶属度矩阵$\boldsymbol{U}$，可能性划分系数$P(\boldsymbol{U};c)$定义为

$$P(\boldsymbol{U};c)=\frac{1}{c}\sum_{i=1}^{c}(\sum_{j=1}^{n}\mu_{ij}^{2}/\sum_{j=1}^{n}\mu_{ij}) \tag{6}$$

对于给定的聚类数c和隶属度矩阵$\boldsymbol{U}$，聚类有效性函数定义为

$$FP(\boldsymbol{U};c)=F(\boldsymbol{U};c)-P(\boldsymbol{U};c) \tag{7}$$

令Ω_c表示$\boldsymbol{U}$的“最优”有限集合，若存在$(\boldsymbol{U}^*,c^*)$满足

$$FP(\boldsymbol{U}^*,c^*)=\min_{c}\{\min_{\Omega_c}FP(\boldsymbol{U};c)\} \tag{8}$$

则以$(\boldsymbol{U}^*,c^*)$为“最优”的有效性聚类，c^*为最佳的分类数目。

2.3 复合属性模糊化

粗糙集中，复合属性A对论域$\boldsymbol{U}$的划分可以表示为

$$\boldsymbol{U}/A=\otimes\{U/a\mid a\in A\} \tag{9}$$

其中，$\boldsymbol{U}/a$是属性a对论域$\boldsymbol{U}$划分形成的等价类集合。

$\otimes$算子定义如下：

$$S_1\otimes S_2=\{X\cap Y:\forall X\in S_1,\forall Y\in S_2,X\cap Y\neq\varnothing\}$$

若$A=\{a_1,a_2,\cdots a_n\}$，则式(9)可以写成

$$\boldsymbol{U}/A=\{X_{1_{i1}}\bigcap X_{2_{i2}}\bigcap\cdots\bigcap X_{n_{in}}\mid X_{1_{i1}}\in\boldsymbol{U}/a_1,X_{2_{i2}}\in\boldsymbol{U}/a_2,\cdots,X_{n_{in}}\in\boldsymbol{U}/a_n\} \tag{10}$$

用$<F_1,F_2,\cdots,F_n>$表示式(10)中$\boldsymbol{U}/A$的等价类$F_1\bigcap F_2\bigcap\cdots\bigcap F_n$，即

$$<F_1,F_2,\cdots,F_n>=F_1\bigcap F_2\bigcap\cdots\bigcap F_n \tag{11}$$

根据公式$(A\bigcap B)(x)=A(x)\wedge B(x)=\min\{A(x),B(x)\}$，对象$x\in\boldsymbol{U}$对该等价类的隶属度为

$$\mu_{<F_1,F_2,\cdots,F_n>}(x)=\mu_{F_1}(x)\wedge\mu_{F_2}(x)\wedge\cdots\wedge\mu_{F_n}(x) \tag{12}$$

在模糊粗糙集中，复合属性A对论域$\boldsymbol{U}$的划分的表示方法与式(9)一致，只是$\otimes$的定义稍有不同，要求在每个模糊等价类中，至少存在一个对象对该模糊等价类的隶属度大于0：

$$S_1\otimes S_2=\{X\cap Y:\forall X\in S_1,\forall Y\in S_2,height(X\cap Y)\neq 0\} \tag{13}$$

类似地，对象$x\in\boldsymbol{U}$对复合属性的模糊等价类的隶属度为

$$\mu_{<F_1,F_2,\cdots,F_n>}(x)=\mu_{F_1}(x)\wedge\mu_{F_2}(x)\wedge\cdots\wedge\mu_{F_n}(x)=\min(\mu_{F_1}(x),\mu_{F_2}(x),\cdots,\mu_{F_n}(x)) \tag{14}$$

为简单起见，式(14)用Zadeh定义的min运算表示了模糊交运算[6]。

3 模糊粗糙属性约简算法

在模糊粗糙集中，每个等价类都是模糊的，它的上、下近似可以定义如下：

$$\mu_{\underline{P}X}(F_i)=inf_x\max\{1-\mu_{F_i}(x),\mu_X(x)\},\forall i$$

$$\mu_{\overline{P}X}(F_i)=sup_x\min\{\mu_{F_i}(x),\mu_X(x)\},\forall i$$

其中，F_i是一个模糊等价类；$\mu_X(x)$是对象x属于$\boldsymbol{U}$上任意模糊集合X的程度。

对于$F_i\in\boldsymbol{U}|C$的模糊正域定义为

$$\mu_{POS_c}(F_i)=\sup_{X\in U|D}\mu_{\underline{X}}(F_i)$$

其中，X是决策属性D的模糊等价类。

$x\in\boldsymbol{U}$对模糊正域的隶属度为

$$\mu_{POS_c}(x)=\sup_{F_i\in\boldsymbol{U}|C}\min\{\mu_{F_i}(x),\mu_{POS_c}(F_i)\}$$

根据模糊正域的定义，可以求出模糊粗糙集条件下决策属性D对条件属性集合C的依赖度：

$$\gamma_C(D)=\frac{|\mu_{POS_c}(x)|}{|\boldsymbol{U}|}=\frac{\sum_{x\in\boldsymbol{U}}\mu_{POS_c}(x)}{|\boldsymbol{U}|}$$

属性约简是指在保持决策表分类或决策能力不变的前提下，删除冗余属性。

设C和D分别是决策表的条件属性集合和决策属性集合，对于C的子集C'，若满足：(1)$\gamma_C(D)=\gamma_{C'}(D)$；(2)从$C'$中删除任何属性$a$后都有$\gamma_{C'-\{a\}}(D)<\gamma_{C'}(D)$，则称$C'$是$C$相对于决策属性$D$的一个约简。

为降低计算复杂度，本文提出属性递减的约简算法。该算法不需要对条件属性的每个子集进行验证，而是从条件属性全集出发，逐步从条件属性集合中减去那些不会对决策表系统造成信息损失的属性。在每次试图减去一个属性时，先对剩下的属性逐个测试，若去掉某个属性不改变条件属性集合的依赖性，删除其中依赖性最小的属性。算法如下：

```
R ← C;
do
S ← {};   //可删除集合初始化
∀x ∈ R;
if γ_{R-{x}}(D) = γ_R(D)   S ← S⋃{x};
if s = {}   //R 中已没有可删除属性
return R;
γ。← γ_C(D); T ← {}; //否则选择一个依赖性最小的属性删除
∀x ∈ S;
if γ_x(D) < γ。   T ← {x};
γ。← γ_x(D);
R ← R − T
```

4 基于聚类有效性分析的模糊粗集归纳学习方法

根据上述分析，本文提出的基于聚类有效性分析的模糊粗集归纳学习方法如下：(1)构造决策表。对样本数据集进行遗失数据完备化，删除重复对象，构造决策表。(2)属性模糊化。用模糊C均值聚类算法模糊化连续属性，利用可能性分布聚类有效性函数分析得到最佳聚类数，并获得属性隶属于各个类的隶属度矩阵及属性在论域的最佳划分。(3)决策表简化。利用隶属度矩阵求出条件属性对决策属性的依赖度，通过属性约简算法消掉冗余属性；对消掉冗余属性后的新决策

表进行属性值的约简，得到决策表的最小约简。(4)规则获取。根据条件属性的最小约简，过滤掉冗余及等价的决策规则，提取出有价值的规则。

5 实例分析

为了证明此方法的有效性，选择如表 1 所示的气象信息决策表，该决策表含有 4 个条件属性 $C=\{a1,a2,a3,a4\}$ (晴朗度指数 $a1$、温度指数 $a2$、湿度指数 $a3$、风况指数 $a4$)和一个决策属性 $D=\{d\}$。

表 1 气象信息决策表

对象	晴朗度指数 $a1$	温度指数 $a2$	湿度指数 $a3$	风况指数 $a4$	d
1	0.80	0.90	0.90	0.30	0
2	0.75	0.85	0.88	0.50	0
3	0.50	0.95	0.75	0.20	1
4	0.20	0.60	0.80	0.40	1
5	0.15	0.95	0.50	0.00	0
6	0.25	0.30	0.55	0.60	0
7	0.45	0.20	0.55	0.65	1
8	0.78	0.70	0.85	0.10	0
9	0.90	0.40	0.45	0.40	1
10	0.20	0.65	0.90	0.45	1
11	0.95	0.55	0.60	0.55	1
12	0.50	0.55	0.80	0.60	1
13	0.60	0.90	0.60	0.20	1
14	0.10	0.45	0.85	0.50	0

首先进行属性模糊化。采用模糊 C 均值聚类分别求取 4 个属性的模糊划分矩阵 $\boldsymbol{U}$。采用基于可能性划分的聚类有效性函数分析，得出 $a1$, $a2$ 划分成 3 类最有效；$a3$, $a4$ 划分成 2 类最有效(如表 2 所示)。

表 2 气象数据聚类有效性分析结果

类数	$FP(\boldsymbol{U},c)$			
	$a1$	$a2$	$a3$	$a4$
2	0.035 8	0.097 4	-0.003 4	-0.036 9
3	-0.006 9	0.033 4	0.008 4	-0.024 1
4	0.015 6	0.043 4	0.002 8	0.035 6
5	0.007 3	0.066 3	0.005 2	-0.018 2

得到论域在属性上的划分如下：

$\boldsymbol{U}/a1=\{\{1,2,8,9,11\},\{4,5,6,10,14\},\{3,7,12,13\}\}$

$\boldsymbol{U}/a2=\{\{1,2,3,5,13\},\{6,7,9\},\{4,8,10,11,12,14\}\}$

$\boldsymbol{U}/a3=\{\{1,2,3,4,8,10,12,14\},\{5,6,7,9,11,13\}\}$

$\boldsymbol{U}/a4=\{\{2,4,6,7,9,10,11,12,14\},\{1,3,5,8,13\}\}$

利用模糊粗糙集的属性约简算法计算得出

$\gamma_C(d)=\gamma_{C-\{a3\}}(d)=0.7998$

即属性 $a3$ 对于决策属性是冗余的。对约去 $a3$ 的决策表进行属性值的约简，消去每条决策规则中属性的冗余值，得到决策规则的一种最小解并提取出如下决策规则：

(1)if $0.75\le a1\le 0.95$ and $0.85\le a2\le 0.95$ then $d=0$；

(2)if $0.45\le a1\le 0.60$ then $d=1$；

(3)if $0.45\le a2\le 0.70$ and $0.40\le a4\le 0.65$ then $d=1$；

(4)if $0.10\le a1\le 0.25$ and $0.40\le a4\le 0.65$ then $d=0$；

(5)if $0.75\le a1\le 0.95$ and $0.00\le a4\le 0.30$ then $d=0$；

(6)if $0.75\le a1\le 0.95$ and $0.20\le a2\le 0.40$ then $d=1$。

6 结束语

本文提出一种基于聚类有效性分析的模糊粗糙集归纳学习方法，使用模糊 C 均值聚类算法模糊化连续属性，通过聚类有效性分析自动确定最佳的分类数目，克服了目前属性模糊化方法需要人为确定划分的类数、几乎不考虑信息系统具体属性值的缺点。该方法为解决粗糙集中连续属性的规则获取问题提供了一个有效的途径。

参考文献

[1] Dubois D. Putting Rough Sets and Fuzzy Sets Together[M]// Intelligent Decision Support: Handbook of Applications and Advanced of the Rough Set Theory. Boston: Kluwer Academic Publishers, 1992: 203-222.

[2] Wang Yifan. Mining Stock Price Using Fuzzy Rough Set System[J]. Expert System with Application, 2003, 24(1): 13-22.

[3] Jensen R, Shen Q. Semantics-preserving Dimensionality Reduction: Rough and Fuzzy-rough-based Approaches[J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1457-1471.

[4] 聂作先, 刘建成. 一种面向连续属性空间的模糊粗糙约简[J]. 计算机工程, 2005, 31(6): 163-165.

[5] 高新波. 模糊聚类分析及其应用[M]. 西安: 西安电子科技大学出版社, 2004.

[6] 汪培庄. 模糊集合论及其应用[M]. 上海: 上海科学技术出版社, 1983.

(上接第 85 页)

5 结束语

本文在索引器模块引入后缀数组技术。用后缀数组实现的索引器采用了全文检索，解决了目前短语查准率低的问题，并且巧妙地避开了中文分词。实现全文索引有以下优势：

(1)良好引用：能找到比一般搜索引擎更加有用的信息。

(2)易于使用：对于提升全民互联网使用水平非常重要(方便用户输入有效的关键字)。

(3)数据挖掘：索引的全面性，可以把数据挖掘中成熟的算法和思路应用于搜索引擎中，进而提高查准率。

由于采用后缀数组全文索引，占用的空间 $O(5n)$要比词索引(使用倒排文件方式)大很多，因此引入压缩后缀数组技术解决该问题，给出了应用在搜索引擎中的算法伪代码，把索引空间压缩到了 $O(n)$。从而大大开辟了后缀数组技术在搜索引擎中的应用。本文的实验与统计数据表明短语查询的准确率提高了近 20%。

参考文献

[1] 姚全珠, 丁晓剑, 任雪利, 等. 一种新的基于 XML 的索引机制[J]. 计算机工程, 2006, 32(15): 90-92.

[2] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval: 现代信息检索[M]. 王知津, 贾福新, 郑红军, 等, 译. 北京: 机械工业出版社, 2005.

[3] 徐宝文, 张卫丰. 搜索引擎与信息获取技术[M]. 北京: 清华大学出版社, 2003.

[4] McCreight E M. A Space-economical Suffix Tree Construction Algorithm[J]. Journal of the ACM, 1976, 23(12): 262-272.

[5] Grossi R, Vitter J S. Compressed Suffix Arrays and Suffix Trees with Application to Text Indexing and String Matching[C]//Proc. of the 32nd ACM Symposium on Theory of Computing. Los Angeles, California, USA: [s. n.], 2000.

[6] Manber U, Myers G. Suffix Arrarys: A New Method for On-line String Searches[J]. SIAM Journal on Computing, 1993, 24(5): 935-948.