文章编号:1004-5694(2006)05-0650-03

基于模糊 C 均值聚类的粗集理论连续属性的离散化新算法*

黄晓莉, 曾黄麟, 王秀碧
(四川理工学院,自贡 643000)

摘 要:讨论模糊 C 均值聚类算法在决策表条件属性对决策属性的相容程度的指导下对粗集理论中的连续属性进行离散化的一种新算法。该算法充分考虑属性之间的相关性,将所有连续属性转化为矩阵同时处理,能明显提高传统动态层次聚类算法离散化过程的速度。算法测试结果表明,新算法能较好地保留有效属性,提高离散化精度。

关键词:模糊 C 均值聚类;粗集理论;连续属性;离散属性
中图分类号:TP18 文献标识码:A

0 引 言

粗集(RS)的突出优点是具有很强的定性分析能力,不需要预先给定属性的数量描述(如统计学中的概率分布),直接从给定问题的描述集合出发找出问题的内在规律。但粗集适于处理离散数据,对于连续数据的处理能力有限,大大限制了其应用范围,因而连续属性的离散化是粗集理论中急待解决的问题。在保证信息系统分辨关系的条件下,采用基数最小的断点集合对系统进行的离散化就是质量最高的离散化[1]。目前,粗集中连续属性的离散化方法多数是监督启发式算法,其中包括利用专家领域知识进行简化、概念树法、基于等间隔的划分质量期望值法[2]及传统动态层次聚类算法[3]等。前面几种算法的缺点在于候选断点的选择缺乏统一理论指导,没有考虑连续属性在决策中的相关性,从而产生不合理或者多余的断点。而最后一种算法虽然利用了决策表的相关性信息,但它寻找断点的算法使用的是传统聚类分析方法(硬划分),在聚类过程中没有考虑各属性值对区间的隶属度及属性区间的长度,容易陷入局部极值点使算法不能收敛,而且聚类过程只能针对单一属性,存在 NP 难题。本文中我们提出的离散化新算法将连续属性值看作区间数,通过模糊 C 均值聚类算法(FCM)将样本划分为若干等价波段组,然后根据最大隶属度原则,只保留每组中具有代表性的波段。新算法与其他算法的不同之处在于:

1) FCM 算法利用决策表给出的连续属性的信息构建聚类原型矩阵和模糊划分矩阵,将所有连续属性同时转化为矩阵处理[4],充分考虑了属性之间的相关性,将聚类归结成一个带约束的非线性规划问题,通过优化求解获得数据集的模糊划分和聚类;

2) 聚类过程中考虑区间大小的影响,将聚类对象用区间长度矩阵加权;

3) 将每个条件属性对所有决策属性的相容程度 df_i 作为该属性对聚类的贡献率对聚类对象加权,从而避免了纯粹从数学角度将样本聚类,在有效减少冗余属性的情况下,使离散化尽可能少地损失有用信息;

4) 利用所有条件属性和所有决策属性的相容程度 df 作为衡量离散化质量的依据,df 越大,离散化后的决策表相容性越好,离散化的质量越高。

由以上看出,新算法在离散化过程中充分利用了粗集理论的系统信息,考虑了属性间的相互影响,具有较强的合理性。测试结果表明新算法的综合性能优于参考算法,能有效提高离散化精度,增强系统鲁棒性,提高离散化速度。

1 基本概念

1.1 粗集基本概念:

定义 1[5] 定义 $S=<U,C,D,V,f>$ 是知识系统,其中 U 是研究论域,$C\cup D=\mathbf{R}$ 是属性(特征)集合,子集 C 和 D 分别表示条件属性集和决策属性集,$V=v_r, r\in\mathbf{R}$ 是属性值集合,v_r 表示属性 r 的属性值,f 定义一个信息函数 $f:U\times\mathbf{R}\rightarrow V$,它指定 U 中每一对象 x 的属性值。S 的表格形式称为决策表。本文中我们讨论的 C 和 D 为连续属性,即 v_r 是区间数。

* 收稿日期:2006-02-14
基金项目:四川省教育厅应用研究基础项目资助(2005A140)
作者简介:黄晓莉(1972-),女,四川全天县人,硕士研究生,主要研究方向为模式识别与智能系统,E-mail:hxlpiaoran@sina.com;曾黄麟,教授。

定义2 设子集 $X,Y \subset U$,若根据决策属性 D,X 和 Y 不可分辨时,称其为 $\mathrm{ind}(D)$,它代表这2个子集都属于 D 中的一个范畴,表示为 $U \mid \mathrm{ind}(D)$。C 的下近似集定义成 $CX_* = \bigcup \{Y \in U \mid \mathrm{ind}(C) : Y \subseteq X\}$,表示根据条件属性 C,U 中一定能归入 X 的元素的集合。定义 $\mathrm{POS}_C(D) = \bigcup_{X \in U \mid \mathrm{ind}(D)} CX_*$ 为与 C 相关的等价类 $U \mid \mathrm{ind}(D)$ 的正域,它表示 U 中所有通过 C 能唯一归入 $U \mid \mathrm{ind}(D)$ 的元素的集合。定义 D 对 C 的依赖程度 $df = |\mathrm{POS}_C(D)| / |U|$[6],其中 $|\cdot|$ 表示集合中元素的基数目,df 表示所有利用 C 将 U 中的元素能唯一归入 $U \mid \mathrm{ind}(D)$ 的程度[5],即条件属性和决策属性的相容程度。

1.2 模糊划分基本概念

定义3 $U = (\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_k, \cdots, \boldsymbol{X}_n)$ 是研究论域,$\boldsymbol{X}_k = (x_{k1}, x_{k2}, \cdots, x_{kj}, \cdots, x_{ks})$ 为具有 s 个属性的样本 $\boldsymbol{X}_k$ 的特征矢量,对应特征空间中的一个点。x_{kj} 为特征矢量 $\boldsymbol{X}_k$ 第 j 个属性上的属性值。

定义4 设 A 是 U 的一个子集,定义 $\mu_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$ 为集合 A 的特征函数。如果 U 上的集合 $\widetilde{A}$ 由特征函数 $\mu_{\widetilde{A}}(x)$ 表征,而 $\mu_{\widetilde{A}}(x)$ 在闭区间 $[0,1]$ 上取值,则 $\widetilde{A}$ 称为模糊集合。$\mu_{\widetilde{A}}(x)$ 称为隶属函数,反映了 U 中的元素 x 对于 $\widetilde{A}$ 的隶属程度。对任意 $x \in U$,$\widetilde{A}$ 也表示为 $\mu_{\widetilde{A}}(x) : U \to [0,1]$。

定义5 设 $\mathbf{R}^+$ 为正实数集,$P(\mathbf{R}^+)$ 为 $\mathbf{R}^+$ 的幂集,令 $I(\mathbf{R}^+) = \{\overline{x} \mid \overline{x} = [x^-, x^+] \subset \mathbf{R}^+\}$,则集合 $I(\mathbf{R}^+)$ 中的元素 $\overline{x} \in I(\mathbf{R}^+)$ 为区间数,其中 x^- 称为左区间值,x^+ 称为右区间值。区间中值 $\dot{x} = \frac{x^+ + x^-}{2}$,区间大小 $\hat{x} = \frac{|x^+ - x^-|}{2}$。

区间数做预处理,即二次特征提取,提取区间数 $\overline{X}_k$ 的区间中值 $\dot{X}_k$ 和区间大小 $\hat{X}_k$,把区间数 $\overline{X}_k$ 投影到 $\dot{\boldsymbol{X}}_k$ 和 $\hat{\boldsymbol{X}}_k$ 张成的空间 $\mathrm{span}(\dot{\boldsymbol{X}}_k, \hat{\boldsymbol{X}}_k)$ 中。设 η 为加权因子,用来控制区间大小对聚类的影响,$\boldsymbol{E}_{s\times s}$ 为单位矩阵,$\boldsymbol{W}$ 为投影变换加权矩阵,即

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{E}_{s\times s} & 0 \\ 0 & \eta \cdot \boldsymbol{E}_{s\times s} \end{bmatrix}_{2s\times 2s}$$

投影变换后的矢量 $\widetilde{\boldsymbol{X}}_k = (\dot{\boldsymbol{X}}_k^{\mathrm{T}}, (\eta \cdot \hat{\boldsymbol{X}}_k)^{\mathrm{T}})^{\mathrm{T}} \in \mathbf{R}^{2S}$。但由于每个条件属性对聚类的贡献不一样,我们用各个条件属性 c_i 对所有决策属性 D 的依赖程度 df_i 构成各属性的贡献加权矩阵 ω^*,则加权后的矢量 $\boldsymbol{X}_k$ 为:

$$\boldsymbol{X}_k = \omega^* \cdot \widetilde{\boldsymbol{X}}_k \qquad (k = 1,2,\cdots,n)$$

设 $\omega = \begin{bmatrix} df_1 & 0 & \cdots & 0 \\ 0 & df_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & df_s \end{bmatrix}_{s\times s}$,则

$$\omega^* = \begin{bmatrix} \omega & 0 \\ 0 & \omega \end{bmatrix}_{2s\times 2s}$$

则区间数的模糊 C 均值聚类算法化为传统模糊 C 均值聚类算法(即 FCM 算法)。

定义6 定义 FCM 算法的模糊 C 划分矩阵为[7]:

$$\boldsymbol{M}_f = (\mu_{ik} \mid \mu_{ik} \in [0,1], \forall i,k;$$
$$\sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{n} \mu_{ik} < n, \forall i)。$$

令 $\boldsymbol{P}_i = (p_{i1}, p_{i2}, \cdots, p_{ij}, \cdots, p_{ic})$ 为传统 FCM 算法的聚类原型矢量,即模糊聚类中心矢量,其中 p_{ij} 为第 i 个样本第 j 个属性区间的中心值或均值,则 $\boldsymbol{P} = (\boldsymbol{P}_1, \cdots, \boldsymbol{P}_i, \cdots, \boldsymbol{P}_c)$ 为传统 FCM 算法聚类原型矩阵,即模糊聚类中心矩阵。样本 $\boldsymbol{X}_k$ 与第 i 类的聚类原形 $\boldsymbol{P}_i$ 使用下面的距离度量:

$$(d_{ik})^2 = \| \boldsymbol{X}_k - \boldsymbol{P}_i \|_{\boldsymbol{A}} = (\boldsymbol{X}_k - \boldsymbol{P}_i)^{\mathrm{T}} \boldsymbol{A} (\boldsymbol{X}_k - \boldsymbol{P}_i)$$

其中,$\boldsymbol{A}$ 为 $s \times s$ 阶的对称正定矩阵,当 $\boldsymbol{A}$ 取单位阵 $\boldsymbol{I}$ 时,$(d_{ik})^2$ 为欧几里德距离。$\| \cdot \|$ 为求矩阵范数。

定义 $J_m(\boldsymbol{M}_f, \boldsymbol{P}) = \sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ik})^m (d_{ik})^2$ 为各类中样本 $\boldsymbol{X}_k$ 与其聚类原形矢量 $\boldsymbol{P}_i$ 的类内加权误差平方和目标函数,其中 m 为加权指数,即平滑参数。则聚类准则为寻求最佳组对 $(\boldsymbol{M}_f, \boldsymbol{P})$,在满足约束 $\mu_{ik} \in \boldsymbol{M}_f$ 的条件下,$J_m(\boldsymbol{M}_f, \boldsymbol{P})$ 为最小。

定理1 对 $\forall k$,定义集合 $I_k = \{i \mid 1 \leqslant i \leqslant c, d_{ik} = 0\}$,$\bar{I}_k = \{1,2,\cdots,c\} - I_k$,则使得 $J_m(\boldsymbol{M}_f, \boldsymbol{P})$ 为最小的 μ_{ik} 值为[7]:

$$\begin{cases} \mu_{ik} = \dfrac{1}{\sum_{j=1}^{c} \left(\dfrac{d_{ik}}{d_{jk}} \right)^{\frac{2}{m-1}}} & \text{当 } I_k = \varnothing \\ \mu_{ik} = 0, \forall i \in \bar{I}_k, \text{及} \sum_{i \in I_k} \mu_{ik} = 1 & \text{当 } I_k \neq \varnothing \end{cases} \tag{1}$$

使 $J_m(\boldsymbol{M}_f, \boldsymbol{P})$ 为最小的 P_i 值为

$$P_i = \frac{1}{\sum_{k=1}^{n} (\mu_{ik})^m} \sum_{k=1}^{n} (\mu_{ik})^m X_k \tag{2}$$

证明:因矩阵 $\boldsymbol{M}$ 的各列都是独立的,故

$$\min\{J_m(\boldsymbol{M}_f, \boldsymbol{P})\} =$$
$$\min\{\sum_{k=1}^{n} \sum_{i=1}^{c} (\mu_{ik})^m (d_{ik})^2\} =$$
$$\sum_{k=1}^{n} \min\{\sum_{i=1}^{c} (\mu_{ik})^m (d_{ik})^2\}$$

的极值约束条件为:$\sum_{i=1}^{c} \mu_{ik} = 1$。由拉格朗日乘数法,令

$$F = \sum_{i=1}^{c} (\mu_{ik})^m (d_{ik})^2 + \lambda (\sum_{i=1}^{c} \mu_{ik} - 1)$$

最优化的必要条件为:

$$\frac{\partial F}{\partial \lambda} = (\sum_{i=1}^{c} \mu_{ik} - 1) = 0 \tag{3}$$

$$\frac{\partial F}{\partial \mu_{jk}} = [m(\mu_{jk})^{m-1}(d_{jk})^2 + \lambda] = 0 \quad (4)$$

由(4)式解得：

$$\mu_{jk} = \left[\frac{-\lambda}{m(d_{jk})^2}\right]^{\frac{1}{m-1}} \quad (5)$$

将(5)式代入(3)式得：

$$\sum_{j=1}^{c}\mu_{jk} = \sum_{j=1}^{c}\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}}\left[\frac{1}{(d_{jk})^2}\right]^{\frac{1}{m-1}} = 1 \quad (6)$$

解(6)式得：

$$\left(\frac{\lambda}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{j=1}^{c}\left[\frac{1}{(d_{jk})^2}\right]^{\frac{1}{m-1}}} \quad (7)$$

将(7)式代入(3)式得证定理 1 中使 $J_m(\boldsymbol{M}_f, \boldsymbol{P})$ 取最小值的 μ_{ik}。同理可证得定理 1 中使 $J_m(\boldsymbol{M}_f, \boldsymbol{P})$ 取最小值的 P_i。

2　离散化新算法的实现步骤

离散化新算法的实现步骤如下：

第 1 步　令 $c = 2$；初始化各条件属性的贡献加权矩阵 $\omega^* = \boldsymbol{E}_{2s\times 2s}$，即假定各个属性对分类具有同等贡献；初始化区间大小对聚类的影响因子 η；

do

{ 按照定义 5 将区间数 $\overline{X}_k$ 投影加权成 $\widetilde{\boldsymbol{X}}_k = (\hat{\boldsymbol{X}}_k^{\mathrm{T}}, (\eta \cdot \hat{\boldsymbol{X}}_k)^{\mathrm{T}})^{\mathrm{T}} \in \mathbf{R}^{2S}$，并将 $\widetilde{\boldsymbol{X}}_k$ 用 ω^* 加权成 $\boldsymbol{X}_k = \omega^* \cdot \widetilde{\boldsymbol{X}}_k$；

由定义 6 的方法初始化聚类原型矩阵；设定停止迭代误差 β；设定迭代计数器 $m = 0$；

do

{用(1)式更新划分矩阵 $\boldsymbol{M}_f$；

用(2)式更新聚类原型矩阵 $\boldsymbol{P}$；

用 df_i 更新 ω^*

$m = m + 1$；

} While $\| \Delta \boldsymbol{P} \| > \beta$　（$\| \Delta \boldsymbol{P} \|$ 为矩阵 $\boldsymbol{P}$ 迭代前后变化量的范数）

否则传统 FCM 算法停止，输出 $\boldsymbol{M}_f$ 和 $\boldsymbol{P}$

$dfo = 0$；$mum = 0$；$df = 0$；

根据 $\boldsymbol{M}_f$ 得到连续属性的模糊划分，对离散化后的决策表按定义 2 计算 df；

if $df > dfo$

$dfo = df$；

$mum = c$；

else $mum = c - 1$；

$c = c + 1$；

} while $c \leqslant n$（n 为样本数）

$c = mum$。

第 2 步　根据得到的 c 输出相应的 $\boldsymbol{M}_f$ 对连续属性的区间进行划分，将连续属性离散化。

第 3 步　对离散后的属性值用 0，1，2，… 进行编码。

3　算法测试

采用 UCI 机器学习数据库的 3 个数据集：Iris（4 个条件属性、1 个决策属性、样本数 150），Glass（9 个条件属性、1 个决策属性、样本数 214），Ecoli（7 个条件属性、1 个决策属性、样本数 336），验证本文讨论的连续属性离散化新算法、传统动态层次聚类算法、基于等间隔的划分质量期望值法的区别。实验将各个数据集中的样本随机抽取 70% 作为训练子集，剩下的 30% 作为测试子集对离散化后的属性获得的分类规则进行分类精度测试。

算法测试在 Matlab6.5 中进行。本文中我们讨论的新算法主要使用其中的 stepfcm.m 函数；传统动态层次聚类算法使用其中的 kmeans.m 函数；基于等间隔的划分质量期望值法则使用其中的 statistics 工具箱。3 种算法的分类精度测试分别在 3 个数据集中进行 20 次。以各个数据集为横坐标，取每种算法分别在各个数据集中的 20 次分类精度的平均值为纵坐标，得到图 1 所示曲线图。

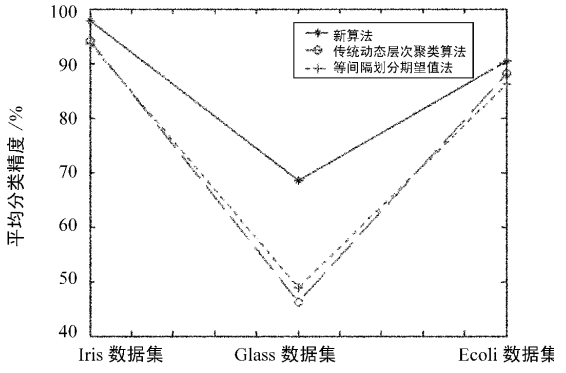

图 1　3 种离散化方法的分类精度比较图

Fig. 1 Discretization results of different algorithms

从图 1 可以看出，新算法由于利用相容程度指导离散化过程，考虑了属性间的相互影响，所以能有效提高分类精度，可以解决单纯的聚类分析很难解决的问题。而且离散化的过程只利用了决策表中数据的统计信息，不涉及决策表的领域知识，是一种领域独立的连续属性离散化算法，具普遍适用性。

参考文献：

[1] NGUYEN S H, SKOWRON A. Quantization of real values attributes, rough set and boolean reasoning approaches[C]//. JCIS 1995: Proc. of the 2nd Joint Annual Conf. on Information Sci., Wrightsville Beach, NC, USA, September 28-October 1, 1995, NC: [s. n], 1995: 34-37.

（下转 678 页）