概念形成学习系统中数值属性的表示与聚类*

张华杰 墙芳躅 袁国斌

(中国地质大学(武汉)计算机系 武汉 430074)

摘要 本文在对 COBWEB、CLASSIT 等概念聚类系统研究的基础上,提出了一种用数值属性的聚类分划来表示数值属性的方法.这种表示的核心是基于数值属性的取值分布.对于在这种表示下数值属性聚类的相关问题及性质,本文进行了较为详细的讨论.在此基础上,本文给出了一个能统一处理数值属性和符号属性的聚类评价函数.一个基于聚类分划表示方法的概念形成学习系统 CFLS(concept formation learning system)已在微机上实现,并被应用于地质学领域的三叶虫分类问题.本文对 CFLS 的设计和实现进行了介绍.

关键词 人工智能,知识获取,机器学习,学习系统,概念形成.

中图分类号 TP18

自 Michalski 提出概念聚类以来,概念聚类成为机器学习的热点问题之一.[1]概念形成(concept formation)是在概念聚类基础上的发展.按照 Gennari 的观点[2],概念形成即增量式概念聚类,它可以形式定义为:

给定:一组顺序出现的例子及其相关的描述.

求解:将这些例子聚合为类的聚类,综合每一类中例子得到的该类的定义,以及这些类的一个层次组织形式,即概念树.

在概念形成的研究中,一些代表性的工作有 Feigenbaum 的 EPAM, Lebowitz 的 UNIMEM[3],Fisher 的 COBWEB[4],Gennari 的 CLASSIT[2]和 Martin 的 OLOC[5]等.在概念形成中,例子和概念大多以属性/值对的集合形式描述.COBWEB 中提出了一个处理符号属性的较有代表性的模型.对于数值属性的处理,CLASSIT 的方法较为典型.

在 CLASSIT 中,一个概念 C 的一项数值属性表示为(μ,σ)的形式,其中 μ 为分入该类中例子的该项属性的平均值,σ 为其均方差.CLASSIT 将 COBWEB 中适用于符号属性的评价函数扩展到连续域上,推导出了以每一类中数值属性的 σ 值为基础的评价函数.CLASSIT 只能处理数值属性.

---

* 本文研究得到地矿部高科技项目"矿产资源与地质环境专家系统"基金资助.作者**张华杰**,1963 年生,副教授,主要研究领域为机器学习,专家系统.**墙芳躅**,1940 年生,教授,主要研究领域为知识工程.**袁国斌**,1967 年生,讲师,主要研究领域为专家系统.

本文通讯联系人:张华杰,武汉 430074,中国地质大学(武汉)计算机系

本文 1996-06-17 收到修改稿

通过研究发现,CLASSIT 对数值属性的处理方法存在一些不足,主要表现为:

(1) CLASSIT 中以 σ 为基础的数值属性评价函数与以概率为基础的符号属性的评价函数不容易统一.

(2) CLASSIT 中对一项数值属性仅保留(μ,σ)两个值,没有考虑该项数值属性取值的分布. 例如,属性 A 取值为{10,9,8,9,10,100,99,98,100},显然该项属性在 10 和 100 二个聚焦点附近取值,若简单地求出均值和均方差,$\mu(A)=49.2$,$\sigma(A)=14.92$,不能较好反映 A 的取值情况. 这种情形在概念树的较高层结点经常出现,CLASSIT 的处理方法使得在较高层结点搜索时,数值属性的启发性较差.

(3) 目前的概念形成理论主要是以概率统计为基础的,数值属性表示为(μ,σ),难以区分一项属性取某个值或在某个区间取值的频率与属性取值之间的差异这两个不同问题.

(4) 当 $\sigma=0$ 和 $\sigma\in(0,1)$时,$\frac{1}{\sigma}$的处理比较复杂.

我们提出了数值属性的一种新的表示方法,这种表示方法根据属性取值的分布来描述数值属性. 本文对这种表示下的聚类问题进行探讨,并给出以此为基础的数值属性与符号属性统一的框架. 在概念形成学习系统 CFLS(concept formation learning system)实现了这一框架,本文也将对 CFLS 的实现情况作介绍.

1　数值属性的聚类分划表示及其相关聚类问题

1.1　数值属性的聚类分划表示方法

考虑这样一类聚类问题,聚类系统的输入是一个一个顺序输入的实例,输出是在不断修正的概念树. 实例描述为属性/值对的集合,即

实例 $e=\{att_1/value_1,att_2/value_2,\ldots,att_n/value_n\}$,其中 att_i 为属性名,$value_i$ 为实例 e 在该项属性的取值.

定义 1. 对一个概念 C,依次分入 C 中的实例 $e_1,e_2,\ldots,e_m$ 组成的集合称为 C 的覆盖实例集,记为 $E(C)$,$e_i(1\leqslant i\leqslant m)$. 在属性 att_j 上的取值组成的集合称 C 在 att_j 上的分布集,记为 $D_j(C)$.

在下面的讨论中,我们只考虑数值属性.

定义 2. 假设概念 C 在属性 att_j 上的分布集 $D_j(C)=\{x_1,x_2,\ldots,x_m\}$,$x_i\in R$,$i=1,2,\ldots,m$. 对 $D_j(C)$的一个分划 $\pi(D_j)=\{\{x_{11},x_{12},\ldots,x_{1n_1}\},\ldots,\{x_{M1},x_{M2},\ldots,x_{Mn_M}\}\}$,若满足对任一分划块 $S_r(\pi)=\{x_{r1},x_{r2},\ldots,x_{rn_r}\}(1\leqslant r\leqslant M)$,有对任一 $x_{rl}\in S_r(\pi)$均有 $x_{rl}\in\mu(S_r(\pi))\pm\Delta$,这里 $\mu(S_r(\pi))$为 $S_r(\pi)$的最大元素与最小元素的平均值,称为中心均值,Δ 为一常数. 则 $\pi(D_j)$称为 D_j 在 Δ 下的一个 M 维聚类分划,$\pi(D_j)$的每一个分划块称为 D_j 在 $\pi(D_j)$下的一个聚类块. 第 l 个分划块记为 $S_l(\pi)$.

这里常数 Δ 实际上对应某项属性的属性值分入同一类允许的最大差异值,它与属性的性质密切相关,直接影响到聚类结果.

有了上述定义,我们就可以给出概念 C 的表示形式.

$$C=(character_1,character_2,\ldots,character_n)$$

其中 $character_i(1\leqslant i\leqslant n)$对应一项数值属性 att_i,其形式为 $character_i=\{(\mu_{i_1},(x_{i_11},\ldots,$

$x_{i_1m_1}),P_{i_1}),\ldots,(\mu_{i_r},(x_{i_r1},\ldots,x_{i_rm_r}),P_{i_r})\}$，这里$\{\{x_{i_11},\ldots,x_{i_1m_1}\},\ldots,\{x_{i_r1},\ldots x_{i_rm_r}\}\}$是$C$在分布集$D_i(C)$上的一个聚类分划$\pi(D_i)$，$\mu_{i_j}$对应该聚类分划的第$j$个聚类块$S_j(\pi)$的中心均值，$P_{i_j}=P(x\in S_j(\pi)|x\in D_i(C))$.

在上述表示形式中，概念C的一项数值属性表示为根据它的取值分布得到的一组聚类块及每个聚类块对应的条件概率，我们称这种表示方法为数值属性的聚类分划表示法. 如果符号属性采用COBWEB中的模式，数值属性与符号属性可以较好地统一. 在这种表示中，属性取值的差异与属性在某一聚类块中出现的频率被明显地区别出来. 由于表示形式考虑到了属性取值的分布，因而在概念树较高层的搜索应比CLASSIT方法更具有启发性. 另外，不存在对$\sigma\in[0,1]$区间的情形处理复杂的问题.

根据属性取值的分布来表示它带来的2个问题是：在增量式聚类过程中如何形成一项数值属性的聚类分划，对于同一属性分布集什么样的聚类分划对概念形成过程具有较好的启发性. 在下面的讨论中，我们重点讨论后一个问题，对于前者只在1.3节中作一般性讨论.

1.2 聚类分划的评价函数与最佳聚类分划

设D是概念C在属性att_i上的分布集，$\pi(D)$是D在Δ下的任一聚类分划.

定义 3. 对$\pi(D)$的任一聚类块$S_j(\pi)$，定义$l(S_j)$为该聚类块的聚类频率，这里

$$l(S_j)=\frac{|S_j|}{|D|}$$

其中$|S_j|$为集合$S_j(\pi)$的基数，$|D|$为集合D的基数.

定义 4. 定义函数$F_c(D,\pi(D))$为D在聚类分划$\pi(D)$下的聚类评价函数，这里

$$F_c(D,\pi(D))=\sum_{j=1}^{M}l^2(S_j(\pi))$$

其中$S_j(\pi)$为$\pi(D)$的第j个聚类块，M为$\pi(D)$的维数.

显见$F_c\in[0,1]$. 实际上，在Δ确定的情况下，F_c的大小对应聚类分划聚类质量的好坏.

定理 1. 设$D=\{x_1,x_2,\ldots,x_N\}$，$\pi(D)$是D在Δ下的一个聚类分划，当$\pi(D)=\{\{x_1,x_2,\ldots,x_N\}\}$，$F_c(D,\pi(D))$取最大值，即$F_c(D,\pi(D))=1$. (证明略).

定理 2. 设$D=\{x_1,x_2,\ldots,x_N\}$，$\pi(D)$是D在Δ下的一个聚类分划. 当$\pi(D)=\{\{x_1\},\{x_2\},\ldots,\{x_N\}\}$，$F_c(D,\pi(D))$取最小值，$F_c(D,\pi(D))=\frac{1}{N}$. (证明略).

定理 3. 设$\pi_1(D)$，$\pi_2(D)$均为D的聚类分划，且$\pi_1(D)$是$\pi_2(D)$的细分，则

$$F_c(D,\pi_1(D))\leqslant F_c(D,\pi_2(D))$$

若$\pi_1(D)$是$\pi_2(D)$的真细分，则 $F_c(D,\pi_1(D))<F_c(D,\pi_2(D))$.

(证明由定义4易证).

定理 4. 设$\pi(D)$是D在Δ下一个聚类分划，$S_i(\pi)$、$S_j(\pi)$是π的2个不同聚类块$(i\neq j)$，若有

(1) $OV(S_i,S_j)\neq\varnothing$. 这里$OV(S_i,S_j)=ID(\mu(S_i)\pm\Delta)\cap ID(\mu(S_j)\pm\Delta)$，其中$ID(\mu(S_i)\pm\Delta)=\{x|x\in D$ 且 $x\in[\mu(S_i)-\Delta,\mu(S_i)+\Delta]\}$，$ID(\mu(S_j)\pm\Delta)$类似

(2) $(OV(S_i,S_j)-S_i)\cap S_j\neq\varnothing$

(3) $(OV(S_i,S_j)-S_j)\cap S_i\neq\varnothing$

则存在 D 的另一个聚类分划 $\pi'(D)$,使得 $F_c(D,\pi'(D))>F_c(D,\pi(D))$.

证明:设 $\pi(D)=\{S_1,\ldots,S_i,\ldots,S_j,\ldots,S_m\}$.

且 $|S_i|\geqslant|S_j|$

构造 D 的另一个聚类分划 $\pi'(D)$

$\pi'(D)=\{S_1,\ldots,S_{i-1},S_i\cup OV(S_i,S_j),S_{i+1},\ldots,S_{j-1},S_j-OV(S_i,S_j),\ldots,S_m\}$

因为 $|S_i\cup OV(S_i,S_j)|+|S_j-OV(S_i,S_j)|=|S_i|+|S_j|$

所以 $l(S_i\cup OV(S_i,S_j))+l(S_j-OV(S_i,S_j))=l(S_i)+l(S_j)$ (1)

因为 $|OV(S_i,S_j)|>0$,且 $|S_i|\geqslant|S_j|$

所以 $l(S_i\cup OV(S_i,S_j))>l(S_i)\geqslant l(S_j)$

因为 $l(S_i\cup OV(S_i,S_j))>l(S_i)\geqslant l(S_j)>l(S_j-OV(S_i,S_j))$

又因为(1),所以 $l^2(S_i\cup OV(S_i,S_j))+l^2(S_j-OV(S_j,S_j))>l^2(S_i)+l^2(S_j)$

故有 $F_c(D,\pi'(D))=\sum_{r=1}^{i-1}l^2(S_r)+l^2(S_i\cup OV(S_i,S_j))+\sum_{r=i+1}^{j-1}l^2(S_r)+l^2(S_j-OV(S_i,S_j))+\sum_{r=j+1}^{m}l^2(S_r)>\sum_{r=1}^{m}l^2(S_r)=F_c(D,\pi(D))$. □

定理 5. 设 $\pi(D)$ 是分布集 D 在 Δ 下的一个聚类分划,S_i,S_j 是 $\pi(D)$ 的 2 个聚类块 $(i\neq j)$,若有

(1) $|S_i|\geqslant|S_j|$

(2) $ID(\mu(S_i)\pm\Delta)\cap ID(\mu(S_j)\pm\Delta)\neq\varnothing$

(3) 存在 μ'_i,使得 $S_i\subseteq ID(\mu'_i\pm\Delta)$ 且 $ID(\mu'_i\pm\Delta)\cap ID(\mu(S_j)\pm\Delta)\neq\varnothing$

则存在 D 在 Δ 下的一个聚类分划 $\pi'(D)$,使得 $F_c(D,\pi'(D))>F_c(D,\pi(D))$.

证明:设 $\pi(D)=\{S_1,S_2,\ldots,S_i,\ldots,S_j,\ldots,S_m\}$

构造 $\pi'(D)=\{S_1,\ldots,S_{i-1},S'_i,S_{i+1},\ldots,S_{j-1},S'_j,S_{j+1},\ldots,S_m\}$

其中 $S'_i=S_i\cup(ID(\mu'_i\pm\Delta)\cap ID(\mu(S_j)\pm\Delta))$,$S'_j=S_j-(ID(\mu'_i\pm\Delta)\cap ID(\mu(S_j)\pm\Delta))$

同定理 4,易证 $l^2(S'_i)+l^2(S'_j)>l^2(S_i)+l^2(S_j)$

类似有 $F_c(D,\pi'(D))>F_c(D,\pi(D))$. □

定理 5 是对定理 4 的补充,它表示当一个聚类块的聚类频率大于另一聚类块时,应将实例尽可能地聚入聚类频率高的聚类块中.

定义 5. 分布集 D 在 Δ 下的一个聚类分划 $\pi(D)$ 称为 D 在 Δ 下的一个最佳聚类分划,当且仅当对 D 的任一个在 Δ 下的聚类分划 $\pi'(D)$ 有 $F_c(D,\pi(D))\geqslant F_c(D,\pi'(D))$.

由定义 5,分布集 D 的最佳聚类分划不一定唯一. 例如,令 $D=\{0,0.5,1,1.5\}$,$\Delta=0.6$. 分划 $\pi_1(D)=\{\{0,0.5,1\},\{1.5\}\}$,$\pi_2(D)=\{\{0\},\{0.5,1,1.5\}\}$,显见 π_1,π_2 均为 D 的最佳聚类分划,而 $\pi_1\neq\pi_2$.

定义 6. 设 $\pi(D)$ 为分布集 D 在 Δ 下的一个聚类分划,S_i,S_j 是 π 的 2 个聚类块 $(i\neq j)$,$\mu(S_i)<\mu(S_j)$. 若存在一个 μ_m,$\mu(S_i)\leqslant\mu_m\leqslant\mu(S_j)$,使得 $ID(\mu_m\pm\Delta)\cap S_i\neq\varnothing$,$ID(\mu_m\pm\Delta)\cap S_j\neq\varnothing$,则称 S_i 和 S_j 在 Δ 下邻接.

聚类块的邻接概念表述的是为了寻找 D 的最佳聚类分划存在的聚类块可能重新划分的问题,即对最佳聚类分划的搜索实际上可以通过对邻接聚类块之间可能的各种划分的组

合进行搜索来完成.在最理想的情况下,我们有定理6.

定理6. 设π是分布集D在Δ下的一个聚类分划,若π中不存在邻接聚类块,则π一定是D的最佳聚类分划.

证明:设π是D的一个聚类分划,π中任意2个聚类块都是不邻接的.

假设π不是D的最佳聚类分划,则对D的任一最佳聚类分划$\pi_{max}(D)$,有$\pi\neq\pi_{max}$.

因为$\pi\neq\pi_{max}$

所以至少存在π中的一个聚类块$S_i(\pi)$与π_{max}中的一个分划块$S_j(\pi_{max})$,使得$S_i(\pi)\neq S_j(\pi_{max})$且$S_i(\pi)\cap S_j(\pi_{max})\neq\varnothing$.

设$x_c\in S_i(\pi)\cap S_j(\pi_{max})$,存在2种情形.

(I) $S_j(\pi_{max})\not\subset S_i(\pi)$,即$S_j(\pi_{max})$中存在$-x_j\in S_i(\pi)$

在π中必有另一聚类块$S_l(\pi)(l\neq i)$,使$x_j\in S_l(\pi)$

因为$x_j,x_c\in S_j(\pi_{max})$.

所以存在$-\mu_m$,使$x_j,x_c\in ID(\mu_m\pm\Delta)$

而$x_c\in S_i(\pi)$,$x_j\in S_l(\pi)$,由定义6知$S_i(\pi)$与$S_l(\pi)$是邻接的,与条件矛盾.

(II) $S_j(\pi_{max})\subset S_i(\pi)$

存在$x_i\in S_i(\pi)$,但$x_i\notin S_j(\pi_{max})$

则存在π_{max}中的分划块$S_r(\pi_{max})(r\neq j)$,使$x_i\in S_r(\pi_{max})$

因为$x_i\in S_i(\pi)$,且$S_i(\pi)$不与π中的任何其它聚类块邻接.

所以对于任一$x_q\in S_r(\pi_{max})$,必有$x_q\in S_i(\pi)$(否则x_i所在的$S_i(\pi)$与x_q所在的另一聚类块邻接).

即$S_r(\pi_{max})\subset S_i(\pi)$

因为$S_j(\pi_{max})\subset S_i(\pi)$,$S_r(\pi_{max})\subset S_i(\pi)$

所以$S_j(\pi_{max})\cup S_r(\pi_{max})\subseteq S_i(\pi)$

由定理3知,π_{max}不是D的最佳分划,矛盾.

综合(I)(II)知假设不成立,即π一定是D的最佳聚类分划.□

定理6给出了最佳聚类分划的一个充分条件,显见它不是必要条件.

1.3 增量式数值属性聚类算法讨论

上节我们讨论了根据数值属性的取值分布情况进行聚类的一些性质,另一个重要问题是如何在增量式聚类学习过程中形成一项数值属性的最佳聚类分划.

实际上,在增量式学习过程中一项数值属性根据取值分布情况会形成一些聚类块,按照1.2节,若这些聚类块任意两个都不邻接,则已得到最佳聚类分划.如果存在邻接的聚类块,就存在聚类块的重新划分问题,最佳聚类分划包含在重新划分的组合中.因此给出邻接聚类块的最佳划分算法是问题的关键.

定义7. 设π是分布集D的一个聚类分划,若由π的聚类块组成的向量$(S_i,S_{i+1},\ldots,S_{i+M-1})$,$S_i,\ldots,S_{i+M-1}\in\pi$,$M\geqslant 2$,满足对任意$j\in\{i,\ldots,i+M-1\}$,$S_j$与$S_{j+1}$是邻接的,则称$(S_i,S_{i+1},\ldots,S_{i+M-1})$为$\pi$的一个$M$维邻接块.

对于M维邻接块的重新划分算法我们进行研究,给出了一个完备的重新划分算法M-REDIVIDE,它对任一M维邻接块$(S_1,S_2,\ldots,S_M)$均能给出最佳划分,其时间开销为

$|S_1|\times|S_2|\times\ldots\times|S_M|$. 限于篇幅,这里不再讨论.

在构造实际的学习系统时,完备的重新划分算法时间开销太大,故无实际意义. 实际上数值属性的分布集 D 是一维的,相邻的两个聚类块的重新划分算法是一种近似但实用的算法. 在 CFLS 中,我们就采用的这种算法.

2　一个概念形成学习系统 CFLS

CFLS(concept formation learning system)是我们设计和实现的一个增量式概念聚类学习系统. 它采用数值属性的聚类分划表示方法及相关聚类原理,并实现了与符号属性表示与处理的统一. 我们把它应用于地质学领域的古生物三叶虫分类问题,取得了一定成效.

2.1　CFLS 的聚类评价函数

CFLS 中符号属性的表示与 COBWEB 相同,数值属性的表示采用聚类分划表示法,如 1.1 所述. 在增量式概念聚类过程中,决定聚类质量的关键问题是评价函数的启发性. 下面讨论 CFLS 评价函数的定义及性质.

设概念树中结点 C_p 的全部子结点为 $C_1,C_2,\ldots,C_k$. 对于已分入 C_p 的实例 e,将 e 分入某一 C,$C\in\{C_1,C_2,\ldots,C_k\}$,其评价函数 $E(e,C_p,C)$ 为用 e 的取值修改概念 C 的各项统计值后,由下式得出.

$$E(e,C_p,C)=E_S(e,C_p,C)+E_N(e,C_p,C)$$

$$E_S(e,C_p,C)=\frac{1}{K}\Big(\sum_k P(C_k)\sum_{i\in I_S}\sum_j P(A_i=V_{ij}|C_k)^2-\sum_{i\in I_S}\sum_j P(A_i=V_{ij})^2\Big)$$

$$E_N(e,C_p,C)=\frac{1}{K}\Big(\sum_k P(C_k)\sum_{i\in I_N}\sum_j P(A_i\in S_{ij}^k|C_k)^2-\sum_{i\in I_N}\sum_j P(A_i\in S_{ij}^p|C_p)^2\Big)$$

这里 I_S 为所有符号属性的下标集,I_N 为所有数值属性的下标集;$S_{i1}^k,S_{i2}^k,\ldots,S_{im}^k$ 为概念 C_k 中数值属性 A_i 的全部聚类块,$k=1,2,\ldots,K$. $S_{i1}^p,S_{i2}^p,\ldots S_{in}^p$ 为在概念 C_p 中数值属性 A_i 的全部聚类块.

对 CFLS 的评价函数 $E(e,C_p,C)$ 的几点讨论:

(1) $E_N(e,C_p,C)$ 对数值属性聚类具有较好的启发性. $\sum_{i\in I_N}\sum_j(A_i\in S_{ij}^p|C_p)^2$,反映了在父结点数值属性 A_i 的聚类情况. 把 C_p 继续分为 $C_1,C_2,\ldots,C_k$ 等 k 个子类. 易见,如果在每一个子类 $C_j(1\leqslant j\leqslant k)$ 中,A_i 的取值越集中而且相近,则 $\sum_t P(A_i\in S_{it}^k|C_k)^2$ 值越大,$E_N(e,C_p,C)$ 的值越大,而这正对应着对属性 A_i 的较好聚类,即同一类中该项属性的取值集中且相近. 反之,$E_N(e,C_p,C)$ 的值越小.

(2) $E_N(e,C_p,C)$ 的值受聚类分划的影响.

(3) $E(e,C_p,C)$ 在概念树的高层结点上具有较好的启发性. 概念树的高层结点数值属性取分散,但 $E(e,C_p,C)$ 能从它的取值分布情况中找到继续分类的启发信息. 对高层结点,σ 的启发性较差.

(4) 数值属性的聚类分划表示是动态的,聚类分划根据该项属性当前的取值分布而变化. 这与将数值属性的连续域按区间离散化不同,后者是静态的.

2.2　CFLS 的实现

我们在486微机上用C语言实现了CFLS,并以古生物三叶虫的分类为背景,进行了测试.由领域专家提供的多组实例,经CFLS聚类得到的结果与人类专家分类大体近似.CFLS系统已通过了地质矿产部组织的部级鉴定.

3 结 论

本文提出的用聚类分划描述数值属性以及数值和符号属性统一的评价函数在以下几个方面较CLASSIT有了一些进展:(1)将符号和数值属性的处理统一在以条件概率为基础的评价函数中;(2)用聚类分划表示数值属性主要考虑到了其取值分布情况,使评价函数在高层概念结点的启发性增强,并且区分了取值的频率与取值之间的差异.

这种方法存在的不足是使聚类过程变得更为复杂,重新划分邻接聚类块的开销较大.本文还介绍了实现这种方法的一个聚类学习系统CFLS的情况.

参考文献

1 Michalski R S, Stepp R. Learning from observation: conceptual clustering. In: Michalski R S, Carbonell J G, Mitchell T M eds. Machine Learning: An Artificial Intelligence Approach, Los Altos: Morgan Kaufmann Publishers, 1983. 471～497.

2 Gennari J H. Models of incremental concept formation. Artificial Intelligence, 1989, **40**(**1**～**3**): 11～61.

3 Lebowitz M. Categorizing numeric information for generalization. Cognitive Science, 1985, **9**: 285～309.

4 Fisher D. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 1987, **2**: 305～336.

5 Martin J D, Billman D O. Acquiring and combining overlapping concepts. Machine Learning, 1994, **16**: 121～155.

THE REPRESENTATION AND CLUSTERING OF NUMERIC ATTRIBUTES IN CONCEPT FORMATION

ZHANG Huajie QIANG Fangzuo YUAN Guobin

(*Department of Computer Science China University of Geosciences Wuhan* 430074)

Abstract Through the research on COBWEB and CLASSIT, this paper presents a new representation method of numeric attributes. This method represents numeric attributes based on its value distribution. The relative problems and properties of clustering in this representation are discussed in detail. This paper also proposes a clustering evaluation function which can deal with both symbolic attributes and numeric attributes uniformly. A concept formation system CFLS(concept formation learning system) has been implemented and applied to geological classification area. CFLS is also introduced in this paper.

Key words Artificial intelligence, knowledge acquisition, machine learning, learning systems, concept formation.

Class number TP18