文章编号:1000-582X(2008)01-0057-04

# 数值离散化中粒度熵与分类精度的相关性

王立宏,孙立民,孟佳娜
(烟台大学 计算机学院,烟台 264005)

摘　要:研究离散化方案中断点数、粒度熵与分类精度之间的关系,证明了粒度熵随着断点数的增加而下降。设计了一种混合型的数值离散化算法来提供多种相容离散决策表。实验发现:粒度熵和分类精度之间的相关程度有时高于断点数和分类精度之间的相关程度。

关键词:粒度熵; 离散化; 断点; 分类精度;粗集
中图分类号:TP18　　文献标志码:A

# Correlation Between Granular Entropy and Classification Accuracy in Discretization

*WANG Li-hong*, *SUN Li-min*, *MENG Jia-na*
(School of Computer Science and Technology, Yantai University, Yantai, Shundong 264005, P. R. China)

**Abstract**: This paper discusses the correlation between the number of cut points, granular entropy and classification accuracy in discretization. It is proven that granular entropy decreases if the number of cut points increases. A hybrid discretization algorithm is proposed to provide discretization schemes for studying these measures. The simulation experiments show that the absolute value of the correlation coefficient between number of cut points and classification accuracy is quite large, as it for granular entropy and classification accuracy. Sometimes, the correlation between the granular entropy and classification accuracy is smaller than that between the cut points and classification accuracy.

**Key words**: granular entropy; discretization; cut point; classification accuracy; rough set

数值离散化是机器学习和数据挖掘等领域的重要研究课题,研究焦点是如何设计数值离散化算法,尽可能达到离散化的目标。离散化的目标可能是:1)使学习集合或测试集合有更高的分类精度[1];2)使离散化所需的断点数最少[2-6],同时要求保持数值属性之间的依赖性不变[2-3],或者保持决策表的相容性[4-6],或者保持决策表的泛化决策值不变[6]等。分类精度、断点数、条件信息熵等都可以作为度量离散化方案的指标。

文中将粒度熵作为离散化方案的度量指标。为了能提供大量的数值离散化方案供测试使用,文中设计并实现了一种混合型的离散化算法,可以随机生成断点数近似最少的离散化方案。实验证实粒度熵与分类精度之间的相关系数和断点数与分类精度之间的相关系数相当,有时前者更高一些。

## 1　相关定义

粗集理论中,决策表表示为$(U, C \cup \{d\})$,$U$ 是一个论域,$C = \{C_1, C_2, \cdots, C_k\}$ 是全体条件属性的集合,文中设定条件属性取连续数据。$d$ 是决策属性,取离散值。设 $P$、$Q$ 为 $U$ 上的两个等价关系(即知识),$P$、$Q$ 在 $U$ 上导出的划分为 $U/P = \{X_1, X_2, \cdots, X_n\}$ 和 $U/Q = \{Y_1, Y_2, \cdots, Y_m\}$,相应的概率分布为

收稿日期:2007-09-12
基金项目:国家自然科学基金资助项目(60473115);山东省自然科学基金资助项目(Y2006G22)
作者简介:王立宏(1970-),女,博士,烟台大学教授,主要从事数据挖掘与知识发现方面的研究。(Tel)13589870907;(Email)wanglh@ytu.edu.cn 或 wanglh_000@163.com。

$\{p(X_1), p(X_2), \cdots, p(X_n)\}$, $\{p(Y_1), p(Y_2), \cdots, p(Y_m)\}$,其中 $p(X_i) = |X_i|/|U|$, $p(Y_j) = |Y_j|/|U|$。

知识 $Q$ 相对于知识 $P$ 的条件熵 $H(Q|P)$[7]为

$$H(Q \mid P) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j \mid X_i) \log p(Y_j \mid X_i), \tag{1}$$

其中,$p(Y_j|X_i) = |Y_j \cap X_i|/|X_i|$。

如果把属性集合 $C$ 上的不可区分关系 IND($C$)当作 $P$,把决策属性 $d$ 上的等价关系当作 $Q$,则 $H(Q|P)$变成条件信息熵 $H_c(\{d\}|C)$。

定义 1 条件信息熵 $H_c(\{d\}|C)$

$$H_c(\{d\} \mid C) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j \mid X_i) \log p(Y_j \mid X_i), \tag{2}$$

其中,$U/\mathrm{IND}(\{d\}) = \{Y_1, Y_2, \cdots, Y_m\}$, $Y_j$ 称为决策类,$U/\mathrm{IND}(C) = \{X_1, X_2, \cdots, X_n\}$。条件信息熵能有效地度量条件向量的决策值分布情况,文献[4]采用条件信息熵的改变量作为评价断点的依据。

如果把 $d$ 上的等价关系当作 $P$,把属性集合 $C$ 上的不可区分关系 IND($C$)当作 $Q$,则 $H(Q|P)$变成粒度熵 $H_g(C|\{d\})$。

定义 2 粒度熵 $H_g(C|\{d\})$

$$H_g(C \mid \{d\}) = -\sum_{j=1}^{m} p(Y_j) \sum_{i=1}^{n} p(X_i \mid Y_j) \log p(X_i \mid Y_j), \tag{3}$$

其中 $p(X_i|Y_j) = |Y_j \cap X_i|/|Y_j|$。

粒度熵 $H_g(C|\{d\})$在文献[8]中称为相对粒度熵,用于度量知识的粒度。该文认为随着粒度熵的增加,能得到更精细的分类知识,但并没有给出严格的证明。

从粒度熵的定义可知,

1)如果 $U$ 中每条记录离散化后的条件向量都彼此不同,则 $|X_i| = 1, i = 1, 2, \cdots, n, n = |U|$,此时决策类 $Y_j$ 中有 $|Y_j|$ 个 $X_i \cap Y_j \neq \Phi$,而且 $|X_i \cap Y_j| = 1$,即 $P(X_i|Y_j) = 1/|Y_j|$,粒度熵为 $\sum_{j=1}^{m} p(Y_j) \log |Y_j|$,只与决策类 $Y_j$ 的分布有关。

2)如果 $U$ 中每条记录离散化后的条件向量都完全相同,则 $X_i = U$,此时 IND($C$)只有一个等价类,$n = 1$, $P(X_i|Y_j) = 1$,粒度熵达到最小值 0。

定义 3 相容决策表

设决策表 $U$ 中 $U/\mathrm{IND}(C) = \{X_1, X_2, \cdots, X_n\}$, $U/\mathrm{IND}(\{d\}) = \{Y_1, Y_2, \cdots, Y_m\}$,如果对任何 $X_i$,存在决策类 $Y_j$,满足 $X_i \subseteq Y_j$,则称决策表相容。

## 2 数值离散化的度量指标

对离散化方案有多种度量方法,如:断点数、分类精度、条件信息熵等。

### 2.1 断点数

“断点数最小”是很多算法的最优化目标。例如,文[4]研究了基于信息熵的数值离散化方法。算法选择条件信息熵改变量最大的断点加入断点集合,直到条件信息熵为 0,得出使决策表相容的最小(极小)断点集合。文[5]先将数值决策表采用等间隔算法进行离散化,在保持决策表相容性的前提下,相对均衡地删除断点,以求得尽可能少的断点。

### 2.2 分类精度

训练样本得出的规则如果能较准确地反映出样本中包含的知识,测试样本的正确分类率就会较高。因此,预测精度也是常用的度量指标,文献[1]以提高预测精度作为算法的有效性证明。由于训练样本几乎不可能包含样本数据库中的全部知识,因此预测精度一般不会达到 100%。

### 2.3 条件信息熵

从条件信息熵的定义中不难看出,决策表相容部分的条件熵为 0,信息熵由不相容部分产生。随着断点的增加,条件属性集合 $C$ 上的可区分性逐渐加强。决策表一旦变成相容决策表,条件信息熵一直为 0,不再对离散化方案有度量作用。

### 2.4 粒度熵

通常认为决策表中的每条记录是一条决策规则。实际上,机器学习数据库中的记录只是一些样本数据,并没有广泛的决策意义。如果通过数值离散化能将各个样本之间的共同点找出来,就会得出一些有意义的决策规则。一个决策类 $Y_j$ 可能包含若干个关于属性集合 $C$ 的等价类,等价类的数目越多,表明得出决策值 $d_j$ 的规则越多,这些规则之间的共性越少。如果能减少每个决策类的规则数,相应地就增加了每条规则的支持数,得到了更为强壮的规则。

定理 1 加入断点,决策表的粒度熵增加。

证明 假设决策表已经选取了若干个断点进行离散化,$E = \{X_1, X_2, \cdots, X_n\}$ 是按照离散化后的属性 $C$ 划分得到的等价类。在属性 $c$ 的取值区间加入新的断点 $P_c$ 后,每个等价类 $X_i$ 都可能分成两个等价类 $X_i^l$ 和 $X_i^r$。

$$X_i^l = \{x \mid x \in X_i \wedge c(x) < P_c\},$$

$$X_i^r = \{x \mid x \in X_i \wedge c(x) > P_c\}。$$

此时

$$H_g(C \mid \{d\}) = -\sum_{j=1}^{m} p(Y_j) \sum_{i=1}^{n} (P(X_i^l \mid Y_j) \log P(X_i^l \mid Y_j) + P(X_i^r \mid Y_j) \log P(X_i^r \mid Y_j)),$$

由于

$$P(X_i^l \mid Y_j)\log P(X_i^l \mid Y_j) \leqslant P(X_i^l \mid Y_j)\log(P(X_i^l \mid Y_j) + P(X_i^r \mid Y_j)),$$

$$P(X_i^r \mid Y_j)\log P(X_i^r \mid Y_j) \leqslant P(X_i^r \mid Y_j)\log(P(X_i^l \mid Y_j) + P(X_i^r \mid Y_j)),$$

所以

$$(P(X_i^l \mid Y_j)\log P(X_i^l \mid Y_j) + P(X_i^r \mid Y_j)\log P(X_i^r \mid Y_j)) \leqslant (P(X_i^l \mid Y_j) + P(X_i^r \mid Y_j))\log(P(X_i^l \mid Y_j) + P(X_i^r \mid Y_j)) = P(X_i \mid Y_j))\log(P(X_i \mid Y_j)。$$

因此,加入断点后决策表的粒度熵上升。显然,减少断点会导致决策表的粒度熵下降。

## 3　粒度熵与分类精度的相关性

文献[4]选择条件信息熵改变量最大的断点加入到属性的取值区间,使不相容的决策表成为相容的决策表,条件信息熵降为 0,这可以看作是临界点。进一步增加断点,条件信息熵不再发生变化。表 1 给出了几种度量指标的变化趋势。

粒度熵在增加断点的过程中一直在上升,而预测精度的变化并不直观可见。一般情况下,断点数较少时分类精度较高,但也不是绝对的。预测精度的大小涉及到训练样本内包含的知识多少、提取规则的准确程度和后续测试样本的分布情况,很难给出一种精确的变化规律。从下面的实验中也能看出,断点个数相同时预测精度也能有很大差异。为此,文中给出一种实验度量方法,考察断点数与分类精度之间的相关性、粒度熵与分类精度之间的相关性。

表 1　度量指标随着断点数增加的变化趋势

| 断点数 | 条件信息熵 | 粒度熵 | 预测精度 |
|---|---|---|---|
| 少 | 下降（大于 0） | 单调上升 | 待定 |
| 临界点 | 等于 0 | | |
| 多 | 等于 0 | | |

### 3.1　离散化算法设计

测试相关性需要生成多种离散化方案。为此,文中设计了一种混合型算法:先随机指定每个数值属性的断点位置,然后判断该决策表是不是相容的,如果不相容,则采用文献[4]中的算法增加断点;如果相容,就在保持决策表相容的前提下随机减少断点,直到不能再减少断点为止。这时的离散化方案是一个断点极少的方案。

条件属性集合为 $C = \{C_1, C_2, \cdots, C_k\}$,对属性 $c \in C$,论域 $U$ 中有限个属性值经过排序后得到序列 $v_1^c < v_2^c < \cdots < v_{nc}^c$,其中 $n_c$ 是属性 $c$ 在 $U$ 中的现有取值个数,$\{v_1^c, v_2^c, \ldots, v_{n_c}^c\} = \{c(x): x \in U\}$。

离散化的候选断点集合定义为

$$CP = \{p_i^c \mid p_i^c = \frac{v_i^c + v_{i+1}^c}{2}, 1 \leqslant i \leqslant nc - 1, c \in C\}$$

设原始决策表是相容的,因此把所有候选断点都加入后决策表是相容的。设离散化方案已经采用的断点集合为 $CP_1$,而未采用的断点集合为 $CP_0 = CP - CP_1$。断点需要记录所属属性、断点值。

训练和测试算法的形式化描述如下,训练过程对训练集合进行 0)~9),测试过程对测试集合进行 10)~11)。

0)初始化:$CP_1$ 为空集,$CP_0 = CP$。

1)在 $CP$ 中随机选取若干个断点加入 $CP_1$,按照 $CP_1$ 将决策表进行离散化。

2)检查决策表是否相容(即:相同条件向量对应的决策值是否相同),如不相容,转 3);相容,转 5)。

3)对 $CP_0$ 中的断点 $u$ 逐个进行如下计算:

3.1)按照 $CP_1 + \{u\}$ 重新离散化。

3.2)按(2)式计算离散决策表的条件信息熵,记为 $H(u)$。

4)选择最小的 $H(u)$ 对应的 $u$,$CP_1 := CP_1 + \{u\}$,$CP_0 = CP_0 - \{\mu\}$。如 $H(u) = 0$(决策表相容),转 8);否则转 3)。

5)$CP_1$ 中所有断点标号为 $N$。

6)$CP_1$ 中是否还有断点标号为 $N$? 如没有,转 8);如有,从 $CP_1$ 中随机选择一个标号为 $N$ 的断点 $u$,用 $CP_1 - \{u\}$ 来离散化决策表。

7)检查决策表的相容性,如相容,$CP_1 = CP_1 - \{u\}$;否则将 $u$ 标号为 $Y$,$CP_1$ 不变。无论是否相容都转 6)。

8)采用属性重要性方法进行属性约简。

9)逐条记录进行属性值约简,去掉冗余规则。

10)对测试集合进行规则匹配,记录正确识别、误识、拒识情况。

11)按照(3)式计算离散决策表的粒度熵。

从算法步骤中可以看出,加入断点和删除断点是完全不同的两个处理分支,因此算法不会出现震荡现象。在计算条件信息熵和粒度熵时,统计条件向量 $X_i$ 的个数来计算 $p(X_i)$,统计决策类 $Y_j$ 中 $X_i$ 的个数来计算 $p(X_i \mid Y_j)$,统计条件向量 $X_i$ 对应决策类 $Y_j$ 的次数来计算 $p(Y_j \mid X_i)$。

属性约简采用属性重要性方法,思想和文[6]中选择断点的想法相同。一个属性能区分的属于不同类的“对象对数”越多,该属性越重要。采用贪婪算法,逐个选择重要的属性,直到属于不同类的对象都能被区分开为止。属性值约简方法类似于属性约

简方法,采用贪婪算法为每个规则找到一个最小属性值约简,然后去除冗余的规则,就得到样本学习的最终规则了。

### 3.2 实验与结果分析

文中采用机器学习数据库中的3个数据集进行测试,数据集的情况见表2。

表2 几个数据集的参数

| 数据集名称 | 行数 | 列数 | 类数 |
|---|---|---|---|
| Glass | 214 | 9 | 6 |
| Heart | 270 | 13 | 2 |
| Wine | 178 | 13 | 3 |

实验中随机选取数据集的50%作为学习样本,另外50%作为测试数据。每个数据集按照文中算法随机生成100种离散化方案,从学习样本集合中抽取规则,利用这些规则对测试集合进行分类。每个数据集都记录了100组断点数、粒度熵、测试数据正确识别率、误识率、拒识率。两个向量 $x$, $y$ 的相关系数按(4)式计算,结果如表3、表4。

$$r = \frac{\sum_{i=1}^{100}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{100}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{100}(y_i - \bar{y})^2}}。 \quad (4)$$

从表3中可以看出,粒度熵和正确识别率之间是负相关关系,而且相关系数的绝对值较大,这表明粒度熵越小,正确识别率越高,这和前面的分析是一致的:粒度熵越小,表明规则越强壮,泛化能力也就越强。

表3 粒度熵与识别率的相关系数

| 数据集 | 粒度熵与正确识别率 | 粒度熵与误识率 | 粒度熵与拒识率 |
|---|---|---|---|
| Glass | -0.380 47 | -0.042 5 | 0.438 58 |
| Heart | -0.484 50 | 0.224 8 | 0.421 10 |
| Wine | -0.610 20 | 0.401 7 | 0.533 00 |

表4 断点数与识别率的相关系数

| 数据集 | 断点数与正确识别率 | 断点数与误识率 | 断点数与拒识率 |
|---|---|---|---|
| Glass | -0.222 5 | -0.084 36 | 0.319 01 |
| Heart | -0.630 1 | 0.283 70 | 0.555 40 |
| Wine | -0.586 0 | 0.365 40 | 0.535 30 |

表4中断点数和正确识别率之间也是负相关关系,而且相关系数的绝对值也较大,这表明断点数越小,正确识别率越高,从实验方面证实了"断点数最少"可以作为离散化方案的衡量标准。

但是横向对比这两种相关系数的绝对值大小就会发现:粒度熵和正确识别率之间的关系有时会更紧密一些,因此有理由相信粒度熵可以作为一个较好的离散化度量标准。

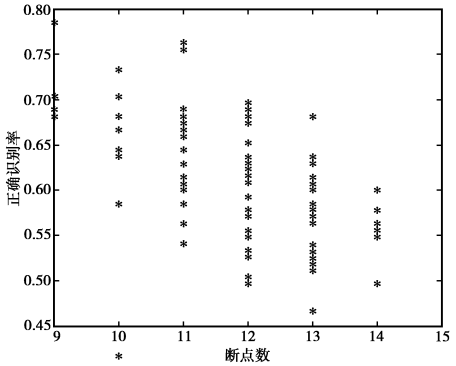


图2 断点数与正确识别率之间的关系

图2、图3所示为heart数据集的实验结果数据。对该数据集进行了100种离散化,得出的断点数与预测精度之间的相关系数绝对值略高于粒度熵和预测精度之间的相关系数绝对值。图2是断点数和预测精度之间的数据点分布图,共100个点,大体上随着断点数增加,正确识别率在下降。图3是粒度熵和预测精度之间的数据点分布图,也是100个点,正确识别率低的点粒度熵相对都比较大。其他两个数据集也有类似的分布图,篇幅所限,不再赘述。

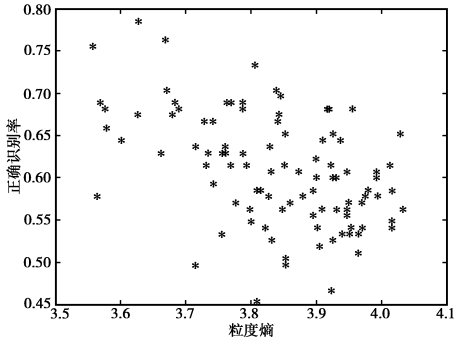


图3 粒度熵与正确识别率之间的关系

## 4 结 论

文中以粒度熵作为离散化方案的度量指标。设计并实现了一种混合型的数值离散化算法,该算法可以提供大量的相容离散决策表供实验使用。实验表明:断点数和预测精度之间的相关系数以及粒度熵和预测精度之间的相关系数绝对值都很大,但断点数和预测精度之间的相关系数有时要低于粒度熵和预测精度之间的相关系数,粒度熵可以作为一个较好的离散化度量标准。

(下转第66页)

计算条件有限的情况下,对流动现象和传热现象采用机理建模方法,对在阳极内发生的电化学反应现象采用格子 Boltzmann 方法,同时运用一个全局类型的数据库来耦合不同尺度的模型。

参考文献:

[1] 罗批,司光亚,胡晓峰. 基于 Agent 的复杂系统建模仿真方法研究进展[J]. 装备指挥技术学院学报,2003,14(1):78-82.
LUO PI, SI GUANG-YA, HU XIAO-FENG, et al. Review of agent-based modeling and simulation in complex system [J]. Journal of the Academy of Equipment Command & Technology, 2003,14(1):78-82.

[2] 李静海,葛蔚. 过程工业中的多尺度效应及离散化单元模拟[J]. 化工进展,1999(5):11-13.
LI JING-HAI, GE WEI. Multi-scale effect and discrete element simulation in process industries [J]. Chemical Engineering Progress, 1999(5):11-13.

[3] 王夔. 突破层次、尺度和时间跨越,向复杂系统逼近——今后化学发展的趋势之一[J]. 自然科学进展,2000,10(8):693-697.
WANG KUI. Breakthrough levels, scale and time lea Pto the approximation of complex systems——one of the chemistry development trend [J]. Natural Science Progress, 2000,10(8):693-697.

[4] JIN HAI LI, JIAY UAN ZHANG, WEI GE, et al. Multi-scale methodology for complex systems [J]. Chemical Engineering Science, 2004(59):1687-1700.

[5] JERRY BIESZCZARD. A Framework for the Language and Logical Computer - Aided Phenomena - Based Process Modeling [C] // The paper for doctor's degree. B. S. Chemical Engineering University of Connecticut, Storrs, 2000.

[6] A A LINNINGER, S CHOWDHRY, V BAHL, et al. A system approach to mathematical modeling of industrial processes [J]. Computers and Chemical Engineering, 2000(24): 591-598.

[7] ANDREAS A. Linninger Recent Advances in process System Engineering [M]. Budapest, Hungary: Technology Conference, 2001.

[8] 廖守亿,戴金海. 复杂适应系统及基于 Agent 的建模与仿真方法[J]. 系统仿真学报,2004,16(1):113-117.
LIAO SHOU-YI, DAI JIN-HAI. Study on Complex Adaptive System and Agent - Based Modeling & Simulation [J]. Journal of System Simulation, 2004,16(1):113-117.

[9] S CHEN, S P DAWSON, G D DOOLEN, et al. Lattice methods and their applications to reacting systems [J]. Computers chem Engng, 1995,19(6): 617-646.

[10] 郭照立, 郑楚光, 李青, 等. 流体动力学的格子 Boltzmann 方法[M]. 武汉: 湖北科学技术出版社, 2002.

(编辑 陈移峰)

(上接第60页)

参考文献:

[1] 李刚, 童频. 基于混合概率模型的无监督离散化算法[J]. 计算机学报,2002,25(2): 158-164.
LI G, TONG F. An unsupervised discretization algorithm based on mixture probabilistic model [J]. Chinese Journal of Computers, 2002, 25(2): 158-164.

[2] STEPHEN D B. Detecting group differences: mining contrast sets [J]. Data Mining and Knowledge Discovery, 2001, (5): 213-246.

[3] STEPHEN D B. Multivariant discretization for set mining [J]. Knowledge and Information Systems, 2001, (3): 491-512.

[4] 谢宏, 程浩忠, 牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报, 2005, 28(9): 1570-1574.
XIE H, CHENG H Z, NIU D X. Discretization of continuous attributes in rough set theory based on information entropy [J]. Chinese Journal of Computers, 2005, 28(9): 1570-1574.

[5] 王立宏, 吴彦, 吴耿锋. 离散格的一种启发式搜索算法[J]. 计算机应用, 2004, 24(8): 41-43.
WANG L H, WU Y, WU G F. Heuristic algorithm for discretization lattice searching [J]. Computer Applications, 2004, 24(8): 41-43.

[6] NGUYEN H S, SKOWRON A. Quantization of real value attributes: Second Annual Joint Conference on Information Sciences (JCIS'95) [C], Wrightsville Beach, North Carolina, USA, 1995:34-37.

[7] 苗夺谦, 王珏, 粗糙集理论中概念与运算的信息表示[J]. 软件学报, 1999, 10(2): 113-116.
MIAO D Q, WANG J. An information representation of the concepts and operations in rough set theory [J]. Journal of Software, 1999, 10(2): 113-116.

[8] 耿志强, 朱群雄, 李芳. 知识粗糙性的粒度原理及其约简[J]. 系统工程与电子技术, 2004, 26(8): 1112-1116.
GENG Z Q, ZHU Q X, LI F. Principle of granularity of knowledge roughness and reduct computing [J]. Systems Engineering and Electronics, 2004, 26(8): 1112-1116.

(编辑 吕建斌)