文章编号: 1001-0920(2008)01-0051-05

# 基于离散度的决策树构造方法

亓常松[1,2], 孙吉贵[1], 于海鸿[1]
(1. 吉林大学 a. 计算机科学与技术学院, b. 符号计算与知识工程教育部重点实验室, 长春 130012; 2. 通化师范学院 软件研究所, 吉林 通化 134002)

**摘 要**: 在构造决策树的过程中,属性选择将影响到决策树的分类精度. 对此,讨论了基于信息熵方法和 WMR 方法的局限性,提出了信息系统中条件属性集的离散度的概念. 利用该概念在决策树构造过程中选择划分属性,设计了基于离散度的决策树构造算法 DSD. DSD 算法可以解决 WMR 方法在实际应用中的局限性. 在 UCI 数据集上的实验表明,该方法构造的决策树精度与基于信息熵的方法相近,而时间复杂度则优于基于信息熵的方法.

**关键词**: 决策树; 离散度; 属性选择
**中图分类号**: TP301.6 **文献标识码**: A

# Approach for constructing decision trees based on dispersion degrees

*QI Chang-song*[1,2], *SUN Ji-gui*[1], *YU Hai-hong*[1]
(1a. College of Computer Science and Technology, 1b. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China; 2. Software Institute, Tonghua Normal College, Tonghua 134002, China. Correspondent: QI Chang-song, E-mail: changsongqi@email.jlu.edu.cn)

**Abstract:** In the process of constructing a decision tree, the criteria of selecting partitional attributes will influence the classification accuracy of the tree. Therefore, the limitations of information entropy based approach and weighted mean roughness (WMR) approach are discussed, and a concept of conditonal attributes dispersion degrees in information systems is proposed. This concept is used to choose partitional attributes in the algorithm of decision trees construction. This approach can overcome the limitations of WMR approach. The results of experiments on the UCI datasets show that the precison retios of the decision trees constructed by using dispersion degree (DSD) are approximate to that of the ones constructed by using entropy-based approach. However, the time complexity of DSD approach is lower.

**Key words:** Decision tree; Dispersion degree; Attribute selection

## 1 引 言

分类是数据挖掘中的一个重要操作. 数据分类方法有很多种,如决策树、神经网络、贝叶斯方法、支持向量机方法以及统计学方法等. 决策树方法有很多优点,如分类速度快、精度高以及易于理解等,因此成为数据挖掘中广泛使用的一种分类方法.

在构造决策树的过程中,划分属性的选择算法非常重要. 本文分析了基于信息熵的方法和基于粗糙集理论方法的不足,提出了基于信息系统中条件属性集离散度的概念,并将其应用于决策树构造过程中的划分属性的选择,描述了基于离散度的决策树构造算法. 例子和实验结果表明了本文方法的有效性.

## 2 划分属性选择方法概述

### 2.1 基于信息熵的方法

基于信息熵的方法[1-3]是构造决策树最重要的方法,著名的 ID3[1] 和 C4.5[1,2] 就是采用了这种方法.

设 $S$ 是 $s$ 个数据样本的集合,分别属于 $m$ 个不同的类别 $C_i(i = 1,2,\cdots,m)$. 令 $s_i = \mathrm{card}(C_i)$,则对一个给定的样本分类所需的期望信息为

$$I(s_1, s_2, \cdots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i), \qquad (1)$$

其中 $p_i = s_i/s$ 是任意样本属于 $C_i$ 的概率.

**收稿日期**: 2006-10-07; **修回日期**: 2007-04-10.
**基金项目**: 国家自然科学基金项目(60473003).
**作者简介**: 亓常松(1972—),男,山东潍坊人,讲师,博士生,从事决策支持系统、数据挖掘的研究; 孙吉贵(1962—),男,辽宁庄河人,教授,博士生导师,从事人工智能、约束程序设计等研究.

设属性 $A$ 有 $v$ 个不同值$(a_1, a_2, \cdots, a_v)$,则用 $A$ 可将 $S$ 划分为 $v$ 个等价类$(S_1, S_2, \cdots, S_v)$. 根据 $A$ 的这种划分的期望信息称作 $A$ 的熵,用下式计算:

$$E(A) = \sum_{j}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j} + \cdots + s_{mj}). \quad (2)$$

$A$ 上该划分获得的信息增益为

$$\mathrm{Gain}(A) = I(s_1, s_2, \cdots, s_m) - E(A). \quad (3)$$

基于信息熵的属性选择方法即计算 $S$ 中样本的每个属性的信息增益,取具有最高信息增益的属性作为划分属性,对划分的各个子集(分支)执行以上操作,最后得到期望的决策树.

### 2.2 基于粗糙集的方法

粗糙集理论[4-6]是一种处理模糊和不确定知识的数学工具,目前已在人工智能、知识与数据发现、模式识别与分类、故障检测等方面得到了较为成功的应用. 粗糙集的基本定义请参见文献[4-6].

Jiang Yun[7]等提出了如下基于粗糙集理论的 WMR(Weighted Mean Roughness)算法:

**定义 1** 对于信息系统 $S = (U, A, V, f)$, $A = C \cup D$,且 $C = \{c_1, c_2, \cdots, c_n\}$, $U/\mathrm{IND}(D) = \{X_1, X_2, \cdots, X_m\}$. 则对于 $X \subset U$, $B \subseteq A$, $X$ 在 $B$ 上的 WMR 定义为

$$\beta_B(c_i) = 1 - \left(\sum_{j}^{m} \omega_j \mu_B(X_j)\right). \quad (4)$$

其中

$$\omega_j = \mathrm{card}(X_j)/\mathrm{card}(U),$$

$$\mu_B(X_j) = \mathrm{card}(\underline{B(X_j)})/\mathrm{card}(\overline{B(X_j)}).$$

$\beta_B(c_i)$ 是条件属性 $c_i$ 对 $U$ 的划分分类确定程度的一个度量:$\beta_B(c_i)$ 越小,其确定程度越高;若 $\beta_B(c_i) = 0$,则 $c_i$ 对 $U$ 的划分分类是完全确定的;否则,若 $\beta_B(c_i) = 1$,则划分的分类是完全不确定的.

WMR 方法的基本过程是计算所有条件属性的 WMR,选择具有最大 WMR 值的属性,根据该属性的取值将样本空间划分为若干子集,对这些子集再使用 WMR 方法继续划分,直到子集中的所有对象都属于同一个决策类.

## 3 基于离散度的属性选择方法

### 3.1 离散度提出

在决策树构造过程中,划分属性的选择依据是条件属性对样本空间的分类能力,信息熵增益和 WMR 便是该分类能力的两种不同的度量方式.

基于信息熵的方法最主要的缺点是决策树中的一棵子树可能重复,并且在树的某些分支中,一个属性可能被重复使用,因此降低了分类效率[2,8];其次是该方法需要大量的对数运算,从而增加了算法的时间复杂度.

基于粗糙集理论的划分属性选择方法的局限性在于: Pawlak 粗糙集是完全按等价类来分类的,因此粗糙集上下近似的定义非常严格[9]. WMR 方法中,为了能比较各条件属性对样本空间的分类能力,要求至少有一个条件属性的至少一个取值的决策分类是完全确定的,否则,所有条件属性的 WMR 值都将为 1,因此无法选择"最好"的划分属性. 然而,在有大量样本的真实信息系统中,这种限制显然是不合理的.

本文提出的离散度(DSD)概念是条件属性对样本空间分类能力的一种新的度量方式,其基本思想是:一个条件属性的分类能力大小也就是该条件属性对样本空间的分类一致性的程度. 若一个条件属性对样本空间划分出的子集中样本都属于同一个分类,则称该条件属性对样本空间的分类是一致的;若一个条件属性划分出的子集中样本在决策类上均匀分布,则称该属性对样本空间的分类是完全不一致的. 构造过程中每一步所选择的划分属性都应使其划分出的子集中样本尽可能地属于同一个分类,也就是选择对样本空间分类一致性程度最高的条件属性,才有可能构造出比较小且精度高的决策树. 下面将给出 DSD 的定义.

### 3.2 离散度的定义

为便于定义,对于信息系统 $S = (U, A, V, f)$, $A = C \cup D$,作如下约定:

不失一般性,令 $D = \{d\}$,即只有一个决策属性. 记 $\mathrm{Vd} = \{d_1, d_2, \cdots, d_n\}$. 条件属性集 $B \subseteq A$,记 $V_B = \{v_1^B, v_2^B, \cdots, v_m^B\}$ 为 $B$ 的所有可能取值. 并记

$$U/\mathrm{IND}(B) = \{Y_1, Y_2, \cdots, Y_m\}, \quad (5)$$

其中 $Y_i = \{y \mid y \in U \wedge f(y, B) = v_i^B\}$;

$$U/\mathrm{IND}(D) = \{X_1, X_2, \cdots, X_N\}, \quad (6)$$

其中 $X_i = \{x \mid x \in U \wedge f(x, D) = d_i\}$;

$$U/\mathrm{IND}(\mathrm{BD}) = \{Z_{11}, Z_{12}, \cdots, Z_{ij}, \cdots, Z_{mn}\}, \quad (7)$$

其中 $Z_{ij} = \{z \mid z \in U \wedge f(z, B) = v_i^B \wedge f(z, D) = d_j\}$.

需要特别说明的是,下文中用到的 $m$ 和 $n$ 分别表示条件属性集和决策属性的不同取值个数.

基于以上约定,首先给出分类系数的概念.

**定义 2**(分类系数) 属性集 $B$ 的一组取值 $v_i^B$ 关于决策属性集 $D$ 的分类系数为

$$\sigma(v_i^B) = \sum_{j}^{n} \left(\frac{\mathrm{card}(Z_{ij})}{\mathrm{card}(Y_i)}\right)^2. \quad (8)$$

分类系数是一个属性集对样本空间分类一致性的度量:若 $B$ 的一组取值 $v_i^B$ 关于决策属性集 $D$ 的分类是一致的,即 $\forall x, y \in Y_i$: $f(x, D) = f(y, D)$,也就是所有在属性集 $B$ 上取 $v_i^B$ 值的对象决策类别

相同,设为 $d_l$,则 $Z_{il} = Y_i$,且 $Z_{ij} = 0 (j \neq l)$,从而 $\sigma(v_i^B) = 1$;而若 $B$ 的一组取值 $v_i^B$ 关于决策属性集 $D$ 的分类是完全不一致的,即 $\mathrm{card}(Z_{ij}) = \mathrm{card}(Y_i)/n$,也就是说,在属性集 $B$ 上取 $v_i^B$ 值的对象在所有决策类别上是均匀分布的,则 $\sigma(v_i^B) = 1/n$.

虽然分类系数可以作为一个属性集对样本空间分类一致性的度量,但属性集的每一组取值在样本空间上一般不是均匀分布的,即不同取值的分类系数在样本空间上的"重要程度"不同. 一组取值在样本空间上的"重要程度"描述了该组取值对属性集在样本空间上分类能力的"贡献"大小:取值对应的样本数越多,则"贡献"越大;反之则越小. 因此,一个属性集在样本空间上的分类能力是由其每一组取值的分类系数和该取值对应的"贡献"综合而成的.

基于以上讨论,定义离散度概念如下:

**定义 3** (离散度)　属性集 $B$ 关于决策属性集 $D$ 的离散度为

$$\mathrm{DSD}_B = 1 - \sum_{i}^{m} \sigma(v_i^B) \times \frac{\mathrm{card}(Y_i)}{\mathrm{card}(U)}. \qquad (9)$$

若 $B$ 的所有取值关于决策属性集 $D$ 的分类是一致的,即 $\sigma(v_i^B) = 1 (i = 1,2,\cdots,m)$,则 $\mathrm{DSD}_B = 0$;若 $B$ 的所有取值关于决策属性集 $D$ 的分类都是完全不一致的,则 $\sigma(v_i^B) = \frac{1}{n} (i = 1,2,\cdots,m)$,$\mathrm{DSD}_B = \frac{n-1}{n}$. $\mathrm{DSD}_B$ 越小,属性集 $B$ 关于决策属性集 $D$ 的分类越确定.

一个条件属性的离散度是其对样本空间对象分类能力的度量. 离散度越小,则其分类能力越强. 将该概念应用于构造决策树的过程,并总是选择离散度最小的条件属性,于是,依据该属性的所有取值对样本空间划分而成的子集,比离散度更大的条件属性划分出的子集中的对象更集中于某一个分类,由此可能构造出复杂度更小的决策树.

### 3.3　基于离散度的决策树构造算法

基于离散度的决策树构造算法 DSD 的基本思想是:计算所有条件属性的离散度,并选择离散度最小的条件属性,根据该属性的取值将样本数据集划分为一些子集,由此构建决策树的一层. 对每个子集重复以上步骤,直到子集中某个属性的离散度为 0,即该属性的所有取值对样本的分类都是准确的,则为该属性的每个取值创建一个其所属分类的叶节点.

然而在实际应用中,将离散度等于 0 作为算法结束条件会导致决策树的过度生长,因此,DSD 算法中对离散度设置一个阈值来控制决策树分支的深度. 如阈值为 0.05,表示如果一个子集中某个属性的离散度小于 0.05,则该子集无需继续分支,或可近似认为该子集的样本所属的类别相同;否则,分支过程将继续进行. 阈值越小,则生成的决策树越复杂;反之则越简单.

**算法 1**　基于离散度的决策树构造算法 DSD$(Q, C, D, \mathrm{td})$.

输入:训练数据集 $Q$,条件属性集 $C$,决策属性集 $D$,离散度阈值 td.

输出:决策树.

1) 计算条件属性集 $C$ 中每一个条件属性的离散度.

2) 选择离散度最小的条件属性 $c$ 作为划分属性.

3) 若该最小离散度 $\leq$td,则为属性 $c$ 的每一个取值创建一个该取值所属分类的叶节点,转步 6).

4) 将 $c$ 作为决策树的根节点:

①Root $\leftarrow c$//Root 是决策树的根节点;

②对属性 $c$ 的每一个可能取值创建一个 Root 的分支节点 Branch$(v_i)$,该分支所表示的子集中的样本对象 $x$ 满足 $f(x,c) = v_i$.

5) 对每一个分支节点 Branch$(v_i)$:

①如果所有的样本属于同一决策类,则为该分支创建该决策类别的叶节点;

②否则,递归调用 DSD$(Q, C - \{c\}, D, \mathrm{td})$.

6) 结束.

## 4　算法举例

基于信息熵的方法,WMR 方法以及基于离散度的方法 DSD,这 3 者之间的主要区别在于属性选择的量度不同. 这里使用一个示例信息表(表 1) 对这 3 种方法进行比较.

**表 1　示例信息表**

| 样本对象 | 条件属性 $C$ | | | | 决策属性 $D$ |
|---|---|---|---|---|---|
| | $a_1$ | $a_2$ | $a_3$ | $a_4$ | |
| 1 | 2 | 1 | 1 | 2 | 2 |
| 2 | 2 | 2 | 1 | 3 | 1 |
| 3 | 2 | 2 | 1 | 1 | 1 |
| 4 | 2 | 1 | 2 | 2 | 1 |
| 5 | 3 | 3 | 2 | 2 | 2 |
| 6 | 3 | 1 | 1 | 2 | 1 |
| 7 | 2 | 1 | 1 | 3 | 2 |
| 8 | 3 | 2 | 2 | 3 | 2 |
| 9 | 2 | 2 | 1 | 2 | 1 |
| 10 | 1 | 3 | 1 | 1 | 1 |
| 11 | 1 | 3 | 2 | 2 | 2 |
| 12 | 1 | 2 | 1 | 2 | 1 |

### 4.1 利用 DSD 算法构造决策树

首先计算所有条件属性的离散度如下:

$$DSD_{a_1} = 1 - \left[\left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) \times \frac{3}{12} + \left(\left(\frac{5}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right) \times \frac{6}{12} + \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) \times \frac{3}{12}\right] = 0.361\ 11,$$

$$DSD_{a_2} = 1 - \left[\left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right) \times \frac{4}{12} + \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right) \times \frac{5}{12} + \left(\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right) \times \frac{3}{12}\right] = 0.369\ 45,$$

$$DSD_{a_3} = 1 - \left[\left(\left(\frac{7}{8}\right)^2 + \left(\frac{1}{8}\right)^2\right) \times \frac{8}{12} + \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right) \times \frac{4}{12}\right] = 0.270\ 83,$$

$$DSD_{a_4} = 1 - \left[\left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right) \times \frac{2}{12} + \left(\left(\frac{5}{7}\right)^2 + \left(\frac{2}{7}\right)^2\right) \times \frac{7}{12} + \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) \times \frac{3}{12}\right] = 0.432\ 54.$$

因属性 $a_3$ 的离散度最小,故选择 $a_3$ 作为决策树的根节点. $a_3$ 的可能取值 $(1,2)$ 将样本对象划分为两个子集: $a_3 = 1$: $(1,2,3,6,7,9,10,12)$; $a_3 = 2$: $(4,5,8,11)$. 递归执行以上操作,最后可得决策树如图 1 所示.

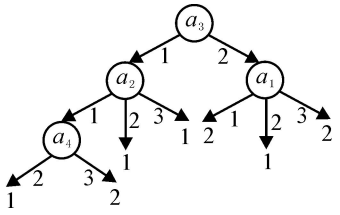

图 1 使用离散度方法构造的决策树

基于离散度的方法可以区分条件属性之间的微小差异,从而可以更好地选择划分属性.

### 4.2 利用 WMR 方法构造决策树

显然 $\beta(a_1) = \beta(a_2) = \beta(a_3) = \beta(a_4) = 1.0$,各条件属性之间的差异无法区分,从而无法选择最佳划分属性. 因此,按属性先后顺序选择 $a_1$ 作为根节点,最终构造出的决策树如图 2 所示.

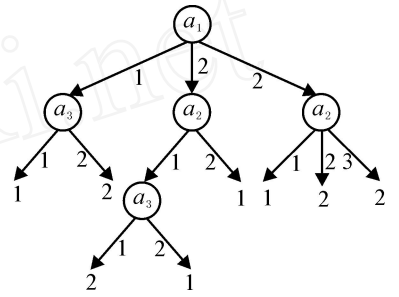

图 2 使用 WMR 方法构造的决策树

### 4.3 利用信息熵方法构造决策树

基于信息熵的方法构造的决策树,其结构与使用离散度方法构造的决策树相同,参见图 1.

由以上 3 种方法对本示例数据集构造决策树的过程与结果表明,DSD 方法可以在 WMR 方法无法完成的情况下选择划分属性. 对于该示例信息表,DSD 方法所构造的决策树与基于信息熵的方法相同,但基于信息熵的方法需要使用对数运算,而条件属性离散度的计算只是简单的四则运算,因此,计算复杂度要小于基于信息熵的方法.

## 5 实验研究

为评价本方法的分类精度和构造决策树的复杂性及算法的时间复杂度,本文实现了 DSD 算法,并使用 UCI 机器学习数据集[10]与基于信息熵的算法和 WMR 方法进行了比较.

为便于比较,DSD 算法和 WMR 算法的实现都是将 C4.5 算法中的属性选择部分分别替换为 DSD 方法和 WMR 方法. 实验结果如表 2 和表 3 所示.

表 2 实验结果(1)

| 数据集 | 样本数 | 条件属性 | 决策属性 | 基于信息熵方法 | | | WMR 方法 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | 准确度/ % | 节点数 | 时间/ ms | 准确度/ % | 节点数 | 时间/ ms |
| Monk1 | 124 | 6 | 1 | 83.9 | 18 | - | 84.1 | 17 | - |
| Monk2 | 169 | 6 | 1 | 76.3 | 31 | - | 75.9 | 35 | - |
| Monk3 | 122 | 6 | 1 | 93.4 | 8 | - | 94.1 | 8 | - |
| lymph | 112 | 18 | 1 | 69.6 | 46 | - | 71.3 | 41 | - |
| Soybean | 683 | 35 | 1 | 96.3 | 95 | 41.2 | 91.3 | 109 | 49.4 |
| connect-4 | 67 557 | 42 | 1 | 87.5 | 6 457 | 3 296 | 82.6 | 6 725 | 4 128 |

表 3　实验结果(2)

| 数据集 | 样本数 | 条件属性 | 决策属性 | DSD 阈值 = 0.01 | | | DSD 阈值 = 0.05 | | | DSD 阈值 = 0.10 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 准确度/ % | 节点数 | 时间/ ms | 准确度/ % | 节点数 | 时间/ ms | 准确度/ % | 节点数 | 时间/ ms |
| Monk1 | 124 | 6 | 1 | 80.5 | 21 | - | 83.9 | 18 | - | 73.2 | 13 | - |
| Monk2 | 169 | 6 | 1 | 73.8 | 37 | - | 77.4 | 27 | - | 68.1 | 22 | - |
| Monk3 | 122 | 6 | 1 | 91.1 | 10 | - | 93.4 | 8 | - | 80.9 | 6 | - |
| lymph | 112 | 18 | 1 | 75.3 | 52 | - | 68.9 | 48 | - | 60.4 | 43 | - |
| Soybean | 683 | 35 | 1 | 94.2 | 101 | 58.7 | 96.5 | 91 | 33.5 | 82.5 | 89 | 30.6 |
| connect-4 | 67 557 | 42 | 1 | 83.5 | 6 572 | 4 153 | 87.8 | 5 387 | 2 659 | 78.7 | 5 104 | 2 375 |

注:1) C4.5 算法为从 http://cerium.raunvis.hi.is/～tpr/courseware/ml/source/c4_5/ 下载的 Windows 版本;

2) 所有算法均在 Windows XP,Pentium1.6 GHz,256 M 环境中测试运行.

实验结果表明,在阈值设为 0.05 时,DSD 方法与基于信息熵的方法所构造的决策树的精度相似,但 DSD 方法有较小的时间复杂度.这主要是由于基于信息熵的方法需要对数运算,而条件属性离散度的计算只是简单的四则运算,并且在 DSD 方法中,随着决策树层次的增长,所需计算的条件属性个数减少,即一个分支中已选择的属性不再参与计算,这样可避免基于信息熵方法中一个分支中属性重复使用的问题,从而降低了算法的时间复杂度,并可得到较少节点数的决策树.

从实验结果还可看出,正是由于 Pawlak 粗糙集完全按等价类来分类,上下近似的定义非常严格,因此,尽管 WMR 方法所构造的决策树的精度在小样本空间上表现较好,但随着样本空间的复杂性提高和样本数量的增多而有较明显的下降.

表 3 的结果还说明,用 DSD 方法构造的决策树的尺寸和精确度受离散度阈值的设置影响较大.为对比阈值大小对 DSD 方法构造的尺寸和精度的影响,分别选取介于 0～0.1 之间的阈值进行了实验(部分结果如表 2,表 3 所示).实验结果表明,阈值在0.05左右时所构造的决策树较优.这是因为,阈值太小,会造成决策树的过度生长,算法的时间复杂度升高,决策树的精度反而会降低;而阈值设置太大,则决策树的生长不足,构造出的决策树的精度也会严重下降.

## 6　结　　语

信息系统中条件属性的离散度是该属性对样本空间中对象分类能力的度量.一个条件属性的离散度越小,则其分类能力越强.将该概念应用于构造决策树的过程,离散度小的属性比离散度大的属性对样本空间划分而成的子集中的对象更集中于某一个分类,因此可以构造出复杂度更小的决策树.并且,随着决策树的层次增长,所需计算的条件属性个数不断减少,即一个分支中已选择的属性将不再参与计算,从而避免了基于信息熵方法中一个分支属性的重复使用问题,降低了算法的时间复杂度.DSD 算法可以解决 WMR 方法在实际应用中的不足,且该方法所构造的决策树的精度与基于信息熵的方法相近,而在时间复杂度上则优于基于信息熵的方法.

DSD 方法的缺点是,所构造的决策树的精度和尺寸受所选择阈值的影响较大,而如何为一个特定的样本空间选择合适的阈值还有待于进一步研究.

## 参考文献( References)


[1] Quinlan J R. C4.5: Program for machine learning[M]. San Marteo: Morgan Kaufmann Publishers, 1993.

[2] Ruggieri S. Efficient C4.5 [J]. IEEE Trans on Knowledge and Data Engineering, 2002, 14(2): 438-444.

[3] Jiawei H, Kamber M. Data mining: Concepts and techniques [M]. Sam Francisco: Morgan Kaufmann Publishers, 2001.

[4] Pawlak Z W. Rough sets [J]. Int J of Information and Computer Science, 1982, 11(5): 341-356.

[5] Pawlak Z W. Rough sets and intelligent data analysis [J]. Information Sciences, 2002, 147(1/4): 1-12.

[6] Pawlak Z. Some issues on rough sets [C]. Trans on Rough Sets I, LNCS 3100. Berlin/Heidelberg: Springer-Verlag, 2004: 1-58.

[7] Jiang Yun, Li Zhan-huai, Zhang Yang, et al. A new approach for selecting attributes based on rough set theory [C]. IDEAL 2004, LNCS 3177. Berlin/Heidelberg: Springer-Verlag, 2004: 152-158.

[8] 苗夺谦,王珏. 其于粗糙集的多变量决策树构造方法[J]. 软件学报, 1997, 8(6): 425-430.
(Miao Duo-qian, Wang Jue. Rough sets based approach for multivariate decision tree construction [J]. J of Software, 1997, 8(6): 425-430.)

[9] Wojciech Ziarko. Variable precision rough set model [J]. J of Computer and System Sciences, 1993, 46(1): 39-59.

[10] Murphyhttp P, Aha W. UCI repository of machine learning databases [EB/OL]. http://www.ics.uci.edu/～mlearn/databases. 2006.