基于信息熵的粗糙集属性离散化方法及应用

沈永红, 王发兴
SHEN Yong-hong, WANG Fa-xing

1.天水师范学院 数理学院, 甘肃 天水 741000
2.复旦大学 数学科学学院, 上海 200433
1.College of Mathematics and Physics, Tianshui Normal University, Tianshui, Gansu 741000, China
2.School of Mathematical Sciences, Fudan University, Shanghai 200433, China
E-mail: shyh_2004@163.com

SHEN Yong-hong, WANG Fa-xing.Method of attributes discretization in rough sets using information entropy and its application.Computer Engineering and Applications, 2008, 44(5): 221-224.

Abstract: This paper analyzes particularity of the rough set theory in processing questions, gives a new method of continuous attributes discretization on the basis of the existing findings, and the method will be applied to fault diagnosis.The experimental results show the decision rules have better classification effect in fault diagnosis, which are constructed by this algorithm.
Key words: rough set; information entropy; discretization; reduce; fault diagnosis

摘 要: 首先分析了粗糙集理论处理问题的特殊性, 在现有研究结果的基础之上给出了一种新的连续属性离散化方法, 并将其应用于故障诊断中, 通过实验结果表明依据该算法构建的决策规则具有较好的故障诊断分类效果。
关键词: 粗糙集; 信息熵; 离散化; 约简; 故障诊断
文章编号: 1002-8331(2008)05-0221-04 **文献标识码**: A **中图分类号**: TP391

当前, 随着设备结构的日趋复杂, 其故障类别越来越多, 反映故障的状态、特征也相应增加。在实际诊断过程中, 为了使诊断结果足够准确、可靠, 总要采集尽可能多的数据样本, 以获得足够多的信息。但是, 数据样本的分类边界往往是不确定的, 并且故障与征兆之间的关系往往也是不确定的[1]。因此对故障诊断问题的分析处理必然要涉及到不确定、不完整知识等概念。而粗糙集理论(Rough Set Theory)是由波兰理工大学的Z·Pawlak教授于上世纪80年代提出的一种新的处理模糊和不确定性知识的数学工具, 具有无需提供除问题所需处理的数据集合之外的任何先验信息, 仅根据观测数据删除冗余信息, 比较不完整知识的程度—粗糙度、属性间的依赖性和重要性来抽取分类规则等的能力。其主要思想就是在保持系统分类能力不变的前提下, 通过知识约简, 导出问题的决策或分类规则。目前, 粗糙集理论已被广泛地应用于机器学习、决策分析、模式识别与数据挖掘、故障诊断等领域。

虽然在众多领域粗糙集都得到了较好的应用, 但是在大量的决策问题中, 决策信息系统中的属性值往往是连续的, 或者是一个真实的数据。而粗糙集方法只能处理离散属性值, 因此连续属性的离散化是制约粗糙集理论应用的一个关键问题。文献[2]给出了目前常用的离散化方法, 根据离散过程中是否考虑信息系统具体的属性值主要分为两大类: 无监督离散化方法和监督离散化方法。而无监督离散化方法主要包括等宽度离散化方法和等频率离散化方法, 监督离散化方法主要包括单规则离散化方法、信息熵离散化方法、超平面离散化方法和超曲面离散化方法等。文献[3]也提到了目前常用的一些具体的属性离散化方法并提出了一种基于信息熵的属性离散法算法。文献[4]也给出了一种基于信息熵的全局连续属性离散化方法。本文主要是在综合文献[2~4]所给出方法的基础上, 提出了一种新的基于信息熵的属性离散化方法, 并在歼击机故障诊断中加以应用, 取得了有效的诊断效果。

1 粗糙集和信息熵

1.1 粗糙集理论[1,2,5~7]

(1) 知识表达系统和决策系统

定义1 在粗糙集理论中, 知识表达系统被定义为如下一个四元组 $S=(U, A, V, f)$。其中 $U=\{x_1, x_2, \cdots, x_n\}$ 为对象的非空有限集合, 也称为论域; $A=\{a_1, a_2, \cdots, a_m\}$ 为属性的非空有限集合; V 为属性值域, $V=\bigcup_{a\in A} V_a$; $f: U\times A\to V$ 为一信息函数, 表示对每一 $a\in A$, $x\in U$, $f(x, a)\in V_a$。当上述知识表达系统中属性 $A=C\cup D$, $C\cap D=\Phi$, 其中 C 为条件属性集, D 为决策属性集时, 知识表达系统也称为决策系统。一般地, 决策系统通常采用决策表来表达。

(2) 不可分辨关系

不可分辨关系是粗糙集理论中的一个重要概念, 在决策表

作者简介: 沈永红(1982-), 男, 助教, 主要从事粗糙集、小波分析和神经网络方面的研究; 王发兴(1981-), 男, 硕士研究生。
收稿日期: 2007-06-11 修回日期: 2007-08-20

中,描述对象的属性是一种不精确信息,这种不精确信息造成了对象之间是不可分辨的或是不分明的,观察这种不可分辨关系的对象正是粗糙集理论研究的出发点。

定义 2 S为知识表达系统,若 $P\subseteq A$,则定义属性集 P 的不可区分关系 $ind(P)$ 为:

$$ind(P)=\{(x,y)\in U\times U,\ \forall a\in P,f(x,a)=f(y,a)\} \quad (1)$$

如果 $(x,y)\in ind(P)$,则称 x 和 y 是 P 不可分辨的。不可分辨关系实际上是一种等价关系,具有不可分辨关系的对象是属性值完全相同的对象。符号 U/P 表示不可分辨关系 $ind(P)$ 在 U 上导出的划分,$ind(P)$ 中的等价类称为 P 基本类。

(3) 粗糙集的下近似、上近似及正域

定义 3 令 $X\subseteq U$,R 是 U 上的一个等价关系。当 X 为 R 的某些等价类的并时,称 X 是 R 可定义的,否则称 X 是 R 不可定义的。R 可定义集称为 R 精确集,R 不可定义集称为 R 粗糙集。粗糙集可以用两个精确集,即粗糙集的下近似和上近似来描述。

定义 4 给定知识表达系统 $S=(U,A,V,f)$,$X\subseteq U$,R 是其上一等价关系,则其对应的上、下近似可分别定义为如下两个集合:

$$\underline{R}(X)=\bigcup\{Y\subseteq U/R|Y\cap X\neq\Phi\} \quad (2)$$

$$\overline{R}(X)=\bigcup\{Y\subseteq U/R|Y\subseteq X\} \quad (3)$$

有了上、下近似以后可以由此给出边界域与正域的定义,$BN_R(X)=\overline{R}(X)-\underline{R}(X)$ 称为 X 的 R 边界域,$POS_R(X)=\underline{R}(X)$ 称为 X 的 R 正域。

若 P,Q 为 U 中的等价关系,Q 的 P 正域记为 $POS_P(Q)=\bigcup_{X\in U/Q}\underline{P}(X)$,反映的是 U 中所有根据分类 U/P 的信息可以准确地划分到关系 Q 的等价类中去的对象集合。

(4) 属性约简

在决策表中,不同的条件属性具有不同的重要程度,一些属性提供了丰富的信息,对产生决策起至关重要的作用,而其它一些属性却似乎是可有可无的。因此,可以在保证决策表具有正确分类能力的同时,对条件属性进行约简,去掉不必要的冗余信息。因此这就涉及到对属性约简的问题,其定义如下:

定义 5 对于一给定的知识表达系统 $S=(U,A,V,f)$,条件属性 C 的约简是 C 的一个非空子集 P。它满足:①$\forall a\in P$,a 都是 D 不可省略的;②$POS_P(D)=POS_C(D)$,则称 P 是 C 的一个约简。

(5) 决策规则

定义 6 对于每个 $x\in U$ 及每个 $a\in C\cup D$,称函数

$$d_x:A\to V,d_x(a)=f(x,a)$$

为对应于给定决策表中的决策规则,x 是决策规则 d_x 的标示,即决策表中集合 U 的元素不表示任何实际的事物,只是决策规则的标示符。

1.2 信息熵

信息熵表征了信源整体的统计特征,是总体的平均不确定性的量度。对于某一特定的信息源,其信息熵就只有一个,不同的信息源,因统计特性不同,其熵也不同。Shannon 定义自信息的数学期望为信息熵,即信息源的平均信息量:

$$H(X)=E[-\log P(x_i)]=-\sum_{i=1}^{N}P(x_i)\log P(x_i) \quad (4)$$

式中 $P(x_i)$ 表示事件 x_i 发生的先验概率。

给定知识表达系统 S,U 为论域,P 为 U 上的等价关系,令 $U/P=\{X_1,X_2,\cdots,X_n\}$,记 $P(X_i)=\frac{|X_i|}{|U|}$,则依据式(4),知识 P 的熵可定义为[8]:

$$H(P)=-\sum_{i=1}^{n}P(X_i)\log P(X_i) \quad (5)$$

2 基于信息熵的粗糙集属性离散化方法

连续属性的离散化实质上就是在特定的连续属性的值域范围内设定若干个离散化划分点,将属性的值域范围划分成一些离散化区间。连续属性的离散化方法很多,不同的离散化方法会产生不同的离散化结果,但任何一种离散化方法都应尽可能满足以下两点:(1)属性离散化后的空间维数应尽量少,也就是经过离散化后的每一个属性都应包含尽量少的属性值的种类;(2)属性值被离散化后丢失的信息尽量少。信息熵是信息系统中属性不确定性的一种量度,因此这里在文献[2~4]的基础上给出了基于信息熵的属性离散化方法。

对于决策表 $S=(U,C\cup D,V,f)$,对每一个连续型条件属性 $a\in C$,论域中其有限个属性值经过排序后为:

$$l_a=v_0^a<v_1^a<\cdots<v_{n_a}^a=r_a$$

于是候选断点可取为:$c_i^a=(v_{i-1}^a+v_i^a)/2\ (i=1,2,\cdots,n_a)$。

设 $X\subseteq U$ 为子集,其实例个数为 $|X|$,其中决策属性为 $j\ (j=1,2,\cdots,t(d))$ 的实例个数为 k_j,于是依据式(5)可给出子集 X 的信息熵为:

$$H(X)=-\sum_{j=1}^{t(d)}p_j\log p_j,\ p_j=\frac{k_j}{|X|} \quad (6)$$

一般地 $H(X)\geq 0$,信息熵 H(X) 越小,说明集合 X 中个别决策属性值占主导地位,因此混乱程度越小,特别有当且仅当 X 中实例的决策属性值都相同时 $H(X)=0$,这一性质保证了以下离散化算法不改变决策表的相容度。

对于断点 c_i^a,决策属性值为 $j\ (j=1,2,\cdots,t(d))$ 的实例中,属于集合 X 且属于 a 的值又小于断点值 c_i^a 的实例的个数记为 $l_j^X(c_i^a)$,大于断点 c_i^a 的实例的个数记为 $r_j^X(c_i^a)$,令

$$l^X(c_i^a)=\sum_{j=1}^{t(d)}l_j^X(c_i^a) \quad (7)$$

$$r^X(c_i^a)=\sum_{j=1}^{t(d)}r_j^X(c_i^a) \quad (8)$$

因此断点 c_i^a 可以将集合 X 分成两个子集 X_l 和 X_r,且有

$$H(X_l)=-\sum_{j=1}^{t(d)}p_j\mathrm{lb}\,p_j,\ p_j=\frac{l_j^X(c_i^a)}{l^X(c_i^a)} \quad (9)$$

$$H(X_r)=-\sum_{j=1}^{t(d)}q_j\mathrm{lb}\,q_j,\ q_j=\frac{r_j^X(c_i^a)}{r^X(c_i^a)} \quad (10)$$

因此定义断点 c_i^a 针对集合 X 的信息熵为:

$$H(c_i^a)=\frac{|X_l|}{|U|}H(X_l)+\frac{|X_r|}{|U|}H(X_r) \quad (11)$$

综合以上所述,可以给出如下基于信息熵的属性离散化算法,为此首先引进记号:记 P 为已选取的断点集合,B 为候选断

点的集合, H为决策表信息熵, V_a为属性a的值域, 初值由式(6)取为H=H(X), 其算法步骤如下:

(1) P=Φ, H=H(X);

(2) 计算对每一个断点c∈B针对集合X的信息熵, 记为H(c, X);

(3) 若H≤min{H(c, X)}或者min{H(c, X)}=0, 则结束转(10); 否则转(4);

(4) 选择使H(c, X)最小的断点c_{min}加到P中, H=min{H(c, X)}, B=B-{c_{min}};

(5) 由步骤(4), 断点c_{min}将集合X划分成X_1和X_2两类, 依据步骤(2)针对X_1和X_2分别计算使得H(c, X_1)和H(c, X_2)取得最小的断点, 分别记为c_{1min}和c_{2min};

(6) 若min{H(c, X_1)}<min{H(c, X_2)}, 则转(7);

若min{H(c, X_1)}>min{H(c, X_2)}, 则转(8); 否则转(9);

(7) 令X=X_2, H=H(X_2), 转(2);

(8) 令X=X_1, H=H(X_1), 转(2);

(9) 选取X_1和X_2中断点数目较少的集合记为X(r=1 or 2), 并令X=X_r, H=H(X_r), 转(2);

(10) 对任一属性a, 若存在断点c^a∈P, 而c^a=min{V_a}或c^a=max{V_a}, 则依据离散化时区间的选择对得到的断点集P进行检查, 从而决定对断点c^a进行取舍。

3 实验结果

选取文献[2]中某歼击机结构故障的部分数据来对所述方法进行可行与有效性分析。其构造故障诊断决策表如表1所示, 其中A={a_1, a_2, …, a_7}表示条件属性, 表示歼击机的状态, 分别为攻角、侧滑角、俯仰角速度、滚转角速度、偏航角速度和法向及侧向过载, d为决策属性, 有6种决策值, 分别是无故障、右平尾损伤故障、左平尾故障、右副翼损伤故障、左副翼损伤故障和方向舵损伤故障。

依据所述方法对表1进行离散化处理, 可得到表2。

依据文献[2]中给出的基于差别矩阵的属性约简算法对表2进行属性约简, 得到如下两个约简集合: {a_1, a_3, a_4}和{a_1, a_4, a_6}。依据该约简集可以提取出如下28条决策规则:

a_3([15.907, *]) AND a_6([29.525, *]) AND a_4([-0.286 9, 0.081 9]) ⇒d(1 0)

表1 故障诊断决策表

序号	a_1	a_2	a_3	a_4	a_5	a_6	a_7	d
1	15.907 6	0.000 0	29.611 4	0.000 0	0.000 0	1.951 4	0.000 0	0
2	15.906 5	0.000 0	29.562 1	-0.163 9	0.002 2	1.933 2	-0.000 1	1
3	15.904 9	0.000 1	29.488 2	-0.409 9	0.005 6	1.959 8	-0.000 3	1
4	15.902 2	0.000 1	29.364 9	-0.819 8	0.011 1	2.004 2	-0.000 6	1
5	15.906 5	0.000 0	29.562 1	0.163 9	-0.002 2	1.933 2	0.000 1	2
6	15.904 9	-0.000 1	29.488 2	0.409 9	-0.005 6	1.959 8	0.000 3	2
7	15.902 2	-0.000 1	29.364 9	0.819 8	-0.011 1	2.004 2	0.000 6	2
8	-0.000 3	-0.790 0	-0.009 6	90.210 0	-1.252 5	0.004 5	-0.010 2	3
9	-0.000 7	-0.790 0	-0.024 0	90.059 9	-1.240 4	0.011 2	-0.010 4	3
10	-0.001 3	-0.790 0	-0.048 0	89.799 0	-1.236 9	0.022 5	-0.010 8	3
11	0.000 3	-0.790 0	0.009 6	90.210 0	-1.242 5	-0.004 5	-0.010 2	4
12	0.000 7	-0.790 0	0.024 0	90.059 9	-1.240 4	-0.011 2	-0.010 4	4
13	0.001 3	-0.790 0	0.048 0	89.799 0	-1.236 9	-0.022 5	-0.010 8	4
14	0.000 0	10.194 4	0.000 0	-20.988 6	16.695 6	0.000 0	0.101 7	5
15	0.000 0	10.192 0	0.000 0	-20.401 7	16.572 5	0.000 0	0.113 8	5
16	0.000 0	10.188 0	0.000 0	-22.090 0	16.367 2	0.000 0	0.133 9	5

a_3([7.951 6, 15.907]) AND a_6([29.525, *]) AND a_4([-0.286 9, 0.081 9]) ⇒d(1 1)

a_3([7.951 6, 15.907]) AND a_6([0.036, 29.525]) AND a_4([-10.610 9, -0.286 9]) ⇒d(1 1)

a_3([7.951 6, 15.907]) AND a_6([29.525, *]) AND a_4([0.081 9, 0.614 9]) ⇒d(1 2)

a_3([7.951 6, 15.907]) AND a_6([0.036, 29.525]) AND a_4([0.081 9, 0.614 9]) ⇒d(1 2)

a_3([7.951 6, 15.907, *]) AND a_6([0.036, 29.525]) AND a_4([0.614 9, 45.304 9]) ⇒d(1 2)

a_3((*, 7.951 6]) AND a_6([*, -0.004 8]) AND a_4([89.927 5, *]) ⇒d(1 3)

a_3((*, 7.951 6]) AND a_6([*, -0.004 8]) AND a_4([45.304 9, 89.927 5]) ⇒d(1 3)

a_3((*, 7.951 6]) AND a_6([0.004 8, 0.016 8]) AND a_4([89.927 5, *]) ⇒d(1 4)

a_3((*, 7.951 6]) AND a_6([0.016 8, 0.036]) AND a_4([89.927 5, *]) ⇒d(1 4)

a_3((*, 7.951 6]) AND a_6([0.036, 29.525]) AND a_4([45.304 9, 89.927 5]) ⇒d(1 4)

表2 离散化后的故障诊断决策表

序号	a_1	a_2	a_3	a_4	a_5	a_6	a_7	d
1	[15.907, *)	[-0.395 1, 5.094 1)	[29.525, *)	[-0.286 9, 0.081 9)	[-0.003 9, 16.474 5)	[0.968 9, 1.982)	[-0.000 45, 0.000 5)	0
2	[7.951 6, 15.907)	[-0.395 1, 5.094 1)	[29.525, *)	[-0.286 9, 0.081 9)	[-0.003 9, 16.474 5)	[0.968 9, 1.982)	[-0.000 45, 0.000 5)	1
3	[7.951 6, 15.907)	[-0.395 1, 5.094 1)	[0.036, 29.525)	[-10.610 9, -0.286 9)	[-0.003 9, 16.474 5)	[0.968 9, 1.982)	[-0.000 45, 0.000 5)	1
4	[7.951 6, 15.907)	[-0.395 1, 5.094 1)	[0.036, 29.525)	[-10.610 9, -0.286 9)	[-0.003 9, 16.474 5)	[1.982, *)	[-0.005 4, -0.000 45)	1
5	[7.951 6, 15.907)	[-0.395 1, 5.094 1)	[29.525, *)	[0.081 9, 0.614 9)	[-0.003 9, 16.474 5)	[0.968 9, 1.982)	[-0.000 45, 0.000 5)	2
6	[7.951 6, 15.907)	[-0.395 1, 5.094 1)	[0.036, 29.525)	[0.081 9, 0.614 9)	[-0.008 3, -0.003 9)	[0.968 9, 1.982)	[-0.000 45, 0.000 5)	2
7	[7.951 6, 15.907)	[-0.395 1, 5.094 1)	[0.036, 29.525)	[0.614 9, 45.304 9)	[-0.624, -0.008 3)	[1.982, *)	[0.000 5, 0.051 2)	2
8	(*, 7.951 6)	(*, -0.395 1)	(*, -0.004 8)	[89.927 5, *)	(*, -1.241 4)	[-0.002 2, 0.007 8)	[-0.010 6, -0.005 4)	3
9	(*, 7.951 6)	(*, -0.395 1)	(*, -0.004 8)	[89.927 5, *)	[-1.241 4, -1.236 9)	[0.007 8, 0.968 9)	[-0.010 6, -0.005 4)	3
10	(*, 7.951 6)	(*, -0.395 1)	(*, -0.004 8)	[45.304 9, 89.927 5)	[-1.236 9, -0.624)	[0.007 8, 0.968 9)	(*, -0.010 6)	3
11	(*, 7.951 6)	(*, -0.395 1)	[0.004 8, 0.016 8)	[89.927 5, *)	(*, -1.241 4)	[-0.007 8, -0.002 2)	[-0.010 6, -0.005 4)	4
12	(*, 7.951 6)	(*, -0.395 1)	[0.016 8, 0.036)	[89.927 5, *)	[-1.241 4, -1.236 9)	(*, -0.007 8)	[-0.010 6, -0.005 4)	4
13	(*, 7.951 6)	(*, -0.395 1)	[0.036, 29.525)	[45.304 9, 89.927 5)	[-1.236 9, -0.624)	(*, -0.007 8)	(*, -0.010 6)	4
14	(*, 7.951 6)	[5.094 1, *)	[-0.004 8, 0.004 8)	(*, -20.695 5)	[16.474 5, *)	[-0.002 2, 0.007 8)	[0.051 2, 0.123 8)	5
15	(*, 7.951 6)	[5.094 1, *)	[-0.004 8, 0.004 8)	[-20.695 5, -10.610 9)	[16.474 5, *)	[-0.002 2, 0.007 8)	[0.051 2, 0.123 8)	5
16	(*, 7.951 6)	[5.094 1, *)	[-0.004 8, 0.004 8)	(*, -20.695 5)	[-0.003 9, 16.474 5)	[-0.002 2, 0.007 8)	(0.123 8, *)	5

$a_1((*, 7.951\ 6])$ AND $a_3([-0.004\ 8, 0.004\ 8])$ AND $a_4([*, -20.695\ 5]) \Rightarrow d(5)$

$a_1((*, 7.951\ 6])$ AND $a_3([-0.004\ 8, 0.004\ 8])$ AND $a_4([-20.695\ 5, -10.610\ 9]) \Rightarrow d(5)$

$a_1([15.907, *])$ AND $a_2([-0.286\ 9, 0.081\ 9])$ AND $a_6([0.968\ 9, 1.982]) \Rightarrow d(0)$

$a_1([7.951\ 6, 15.907])$ AND $a_2([-0.286\ 9, 0.081\ 9])$ AND $a_6([0.968\ 9, 1.982]) \Rightarrow d(1)$

$a_1([7.951\ 6, 15.907])$ AND $a_2([-10.610\ 9, -0.286\ 9])$ AND $a_6([0.968\ 9, 1.982]) \Rightarrow d(1)$

$a_1([7.951\ 6, 15.907])$ AND $a_2([-10.610\ 9, 0.286\ 9])$ AND $a_6([1.982, *]) \Rightarrow d(1)$

$a_1([7.951\ 6, 15.907])$ AND $a_2([0.081\ 9, 0.614\ 9])$ AND $a_6([0.968\ 9, 1.982]) \Rightarrow d(2)$

$a_1([7.951\ 6, 15.907])$ AND $a_2([0.614\ 9, 45.304\ 9])$ AND $a_6([1.982, *]) \Rightarrow d(2)$

$a_1((*, 7.951\ 6])$ AND $a_2([89.927\ 5, *])$ AND $a_6([-0.002\ 2, 0.007\ 8]) \Rightarrow d(3)$

$a_1((*, 7.951\ 6])$ AND $a_2([89.927\ 5, *])$ AND $a_6([0.007\ 8, 0.968\ 9]) \Rightarrow d(3)$

$a_1((*, 7.951\ 6])$ AND $a_2([45.304\ 9, 89.927\ 5])$ AND $a_6([0.007\ 8, 0.968\ 9]) \Rightarrow d(3)$

$a_1((*, 7.951\ 6])$ AND $a_2([89.927\ 5, *])$ AND $a_6([-0.007\ 8, -0.002\ 2]) \Rightarrow d(4)$

$a_1((*, 7.951\ 6])$ AND $a_2([89.927\ 5, *])$ AND $a_6([*, -0.007\ 8]) \Rightarrow d(4)$

$a_1((*, 7.951\ 6])$ AND $a_2([45.304\ 9, 89.927\ 5])$ AND $a_6([*, 0.007\ 8]) \Rightarrow d(4)$

$a_1((*, 7.951\ 6])$ AND $a_2([*, -20.695\ 5])$ AND $a_6([-0.002\ 2, 0.007\ 8]) \Rightarrow d(5)$

$a_1((*, 7.951\ 6])$ AND $a_2([-20.695\ 5, -10.610\ 9])$ AND $a_6([-0.002\ 2, 0.007\ 8]) \Rightarrow d(5)$

$a_1((*, 7.951\ 6])$ AND $a_2([*, -20.695\ 5])$ AND $a_6([-0.002\ 2, 0.007\ 8]) \Rightarrow d(5)$

以上决策规则便形成一个分类器，依据此分类器便可对采集的相应数据进行识别分类。本文采集了 3 种状态时的数据，依据上述分类器进行分类，得到分类结果如表 3。

表 3 故障诊断及诊断结果（最后一列为检测结果）

项目	a_1	a_2	a_3	a_4	a_5	a_6	a_7	d
右平尾损伤故障	15.904 4	0.000 1	29.463 5	-0.491 9	0.006 7	1.968 7	-0.000 4	1
左副翼损伤故障	0.000 9	-0.790 0	0.033 6	89.953 1	-1.239 0	-0.015 7	-0.010 6	4
方向舵损伤故障	0.000 0	10.1896	0.000 0	-21.814 9	16.449 3	0.000 0	0.125 9	5

由表 3 诊断结果来看，利用所提出的方法得到的分类器能正确对故障类型加以判断识别，而且形成的决策规则数并不多，从而表明所提出方法具有一定的可行性与有效性。

4 结论

在实际故障诊断中采集到的数据往往是一个真实的数据，而且这些数据样本的分类边界是不确定的，故障与征兆之间的关系往往也是不确定的。粗糙集理论不需要任何附加信息，依据隐藏在数据中的真实特性做出决策。但因其只能处理量化数据使其应用受到很大限制，信息熵是系统属性不确定性的一种量度，本文在前人研究结果的基础之上提出了一种新的属性离散化方法，并在故障诊断中加以应用。依据实验结果来分析，本文所述方法很有效，对 3 类故障都进行了准确识别，因此该方法将为更为准确和有效地进行故障诊断提供了更多可靠性基础，另外也为粗糙集理论在其它诸如识别、分类和决策等领域更为广泛的应用提供了可能。

参考文献：

[1] **郭小荟，马小平.基于粗糙集的故障诊断特征提取**[J].**计算机工程与应用**, 2007, 43(1): 221-224.

[2] **胡寿松，何亚群.粗糙决策理论与应用**[M].**北京：北京航空航天大学出版社**, 2006.

[3] **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.

[4] Chmielewski M R, Grzymala-Busse J W.Global discretization of continuous attributes as preprocessing for machine Learning[J].International Journal of Approximate Reasoning 1996, 15: 319-331.

[5] Pawlak Z.Rough Set[J].International Journal of Computer and Information Science, 1982, 11: 341-356.

[6] **曾黄麟.粗集理论及其应用**[M].**重庆：重庆大学出版社**, 1998.

[7] **王国胤.Rough 集理论与知识获取**[M].**西安：西安交通大学出版社**, 2001.

[8] **梁吉业，孟晓伟.信息熵在粗糙集理论中的应用**[J].**山西大学学报：自然科学版**, 2002, 25(3): 281-284.

（**上接 130 页**）

[3] Hu N N, Steenkiste P.Evaluation and characterization of available bandwidth probing techniques[J].IEEE Journal on Selected Areas in Communications, 2003, 21(6): 879-894.

[4] Papagiannaki K, Moon S, Fraleigh C.Measurement and analysis of single-hop delay on all IP backbone network[J].IEEE Journal on Selected Areas in Communications, 2003, 21(6): 908-921.

[5] Gruber F, Karrenberg D.Providing active measurements as a regular service for ISP's[C]//Proc of the Passive and Active Measurement Workshop 2001(PAM 2001).Amsterdam: RIPE NCC, 2001: 51-62.

[6] Aashtiani H Z, Magnanti T L.Equilibria on a congested transportation networks[J].SIAM Journal on Algebraic and Discrete Methods, 1981(2): 213-226.