

粗糙集理论的分层递阶约简算法及其信息理论基础

乔斌¹, 李玉榕², 蒋静坪¹

(1. 浙江大学 电气工程学院, 浙江 杭州 310027; 2. 福州大学 电气系, 福建 福州 350002)

摘要: 本文模拟人类认知的分层递阶原则, 提出一种粗糙集理论的分层递阶约简算法. 该算法首先将信息系统或决策系统的知识在由部分属性所构成的多种层次和多种粒度上表示出来, 然后分别对各个属性层次进行递阶约简. 因此, 该算法具有较强的实用性和较好的动态特性, 并且能并行运算. 同时, 本文从信息理论的角度证明了分层递阶约简的理论基础. 文章的最后将该算法应用于某水泥窑炉控制决策的获取中, 证实了其有效性.

关键词: 粗糙集理论; 信息理论; 信息系统; 决策系统; 分层递阶约简

中图分类号: TP18 **文献标识码:** A

Hierarchical reduction approach of rough sets theory and its basis on the information theory

QIAO Bin¹, LI Yu-rong², JIANG Jing-ping¹

(1. College of Electric Engineering, Zhejiang University, Zhejiang Hangzhou 310027, China;

2. Department of Electric Engineering, Fuzhou University, Fujian Fuzhou 350002, China)

Abstract: Simulating the hierarchical principle of human cognizance process, a hierarchical reduction algorithm of rough set theory is proposed in this paper. In the algorithm, the attributes are firstly partitioned into different parts allocated at several layers. Then the knowledge in the information systems or decision systems is presented hierarchically with multiple granularities at multiple layers, and the reduction is hierarchically applied to parts of attributes allocated at each layer. The algorithm has a strong applicability and a good dynamic property, and it can be implemented in parallel. At the same time, the foundation of the algorithm is justified on the basis of information theory. Finally, the hierarchical reduction algorithm is used to acquire the control decision of a cement kiln, which verifies the effectiveness of the algorithm.

Key words: rough sets theory; information theory; information system; decision system; hierarchical reduction

1 引言 (Introduction)

1982 年, 波兰数学家 Z. Pawlak 提出粗糙集理论^[1]. 知识约简是粗糙集理论处理信息系统的重要手段, 可以在保持信息系统分类能力不变的前提下, 导出问题的决策或分类规则^[2].

本文从工程应用角度出发, 提出一种粗糙集理论的分层递阶约简算法. 该算法根据属性的实际物理意义对属性进行分类, 使信息系统中所有属性在单层次和单粒度上的知识表示变为部分属性所构成的知识在多种层次和多种粒度上的表示; 所得的分层递阶约简也相应地使单层次和单粒度上的问题求解转为多种层次和多种粒度上的问题求解, 适合应用于实际问题.

本文首先介绍分层递阶约简算法; 然后根据信息理论分别证明信息系统和决策系统中属性分层递

阶与知识粒度和知识层次之间的关系, 并且论述分层递阶约简算法的性质; 最后, 分层递阶约简算法在某水泥窑炉控制决策获取中的应用证实了其有效性.

2 分层递阶约简算法 (Hierarchical reduction approach)

实际信息系统中, 知识是有层次和有粒度的, 并且不同层次的知识具有不同的粒度^[3]. 人类智能的显著特点就是可以在极不相同的层次上观察和分析同一问题, 并且遵循分层递阶的原则, 有步骤、有层次地利用已有知识缩小问题求解的范围, 直至逼近或得到最终结果.

本文模仿人类智能, 根据分层递阶的原则对信息系统和决策系统进行约简. 以决策系统为例, 分层递阶约简算法步骤如下:

1) 数据准备. 在原始数据库中, 确定条件属性

和决策属性集合,选择各属性的值域,并将原始数据离散化,得到决策表形式的决策系统,记作 $(U, C \cup \{d\})$.

2) 属性分层.根据实际问题中的具体情况,例如,数据采集难易程度、成本代价和实时性等,将属性分为 $C_1, C_2, \dots, C_m$ 共 m 类, m 一般取 $2 \sim 4$.并且 $C_1 \cup C_2 \cup \dots \cup C_m = C, C_i \cap C_j = \emptyset, i \neq j$.

同时,在 $C_1, C_2, \dots, C_m$ 中,数据采集难度、成本和实时性要求等逐层提高或加大.

3) 首层决策系统.以条件属性 C_1 和决策属性 d 构成首层决策系统,对决策表中的不相容情况,可以将决策属性以 $\{d\}$ 的子集表示,记作 $(U_1, C_1 \cup \{d_1\}), U_1 = U$.

对首层决策系统进行约简,得到其简约,共包含 n 组规则.

4) 次层决策系统.利用首层简约将 U 分为 $U_{21}, U_{22}, \dots, U_{2n}$ 共 n 个论域,并且

$$U_{21} \cup U_{22} \cup \dots \cup U_{2n} = U_1,$$

$$U_{2i} \cap U_{2j} = \emptyset, i \neq j.$$

在每个论域中,以条件属性 C_2 和 d_1 的子集构成次层决策系统,分别记作 $(U_{21}, C_2 \cup \{d_1\}), (U_{22}, C_2 \cup \{d_1\}), \dots, (U_{2n}, C_2 \cup \{d_1\})$.当某个首层决策系统中的决策属性唯一时,将不再构造其次层决策系统;对其余次层决策系统分别进行约简,并得到相应简约.

5) 对次层以后的各层决策系统重复步骤4).

3 分层递阶约简算法的理论基础和性质 (Basis and properties of the hierarchical reduction approach)

根据文献[4],设首层条件属性在 U 上的划分为 $X; X = \{X_1, X_2, \dots, X_n\}$;次层条件属性在 U_{2i} 上的划分为 $Y_i; Y = \{Y_{i1}, Y_{i2}, \dots, Y_{im}\}$;决策属性在 U 上的划分为 $D; D = \{D_1, D_2, \dots, D_r\}$;则 X, Y_i 和 D 在 U 和 U_{2i} 的子集组成的 σ -代数上定义的概率分布及 X 与 Y_i 和 D 的联合概率分布分别为

$$\begin{bmatrix} X \\ p \end{bmatrix} = \begin{bmatrix} X_i \\ p(X_i) \end{bmatrix}, p(X_i) = \frac{|X_i|}{|U|},$$

同理定义 $\begin{bmatrix} Y \\ p \end{bmatrix}$ 和 $\begin{bmatrix} D \\ p \end{bmatrix}$.

$$\begin{bmatrix} Y_{ij} \\ p \end{bmatrix} = \begin{bmatrix} X_i \cap Y_j \\ p(Y_{ij}) \end{bmatrix}, p(Y_{ij}) = \frac{|X_i \cap Y_j|}{|U_i|},$$

同理定义 $\begin{bmatrix} Y_{ikj} \\ p \end{bmatrix}$.

其中

$$\sum_i^n p(X_i) = 1, \sum_j^m p(Y_{ij}) = 1,$$

$i = 1, 2, \dots, n, j = 1, 2, \dots, m, k = 1, 2, \dots, r, |\cdot|$ 表示取集合中元素的个数.

文献[4]给出知识之间的粒度关系的定义.本文中,定义知识之间的层次关系:如果 $P < Q$,则称知识 P 比知识 Q 的层次深;反之,则称知识 P 比知识 Q 的层次浅;如果 $P = Q$,则称知识 P 与知识 Q 具有相同的层次.

在一个固定的论域中,引起知识粒度减小,知识层次加深的情况有两种:

1) 至少有一个属性的离散区间数目在离散化过程中增加,且该属性并未因此成为冗余属性.基于这种情况,刘发升等提出一种“基于粗糙集的多层次、逐步求精的发掘算法”,可以有效利用粗糙集,从不同的概念层次出发进行数据挖掘,快速提取数据中感兴趣的规则^[5].

2) 描述对象的属性集中增加了至少一个非冗余属性.本文中的分层递阶约简算法就是针对这种情况的解决方案.

苗夺谦利用信息熵的递增性对情况1)中不同粒度的知识之间的关系进行了论证^[4];本文则将利用信息熵的强可加性和可加性对情况2)中不同层次的知识之间的关系进行论证.

3.1 信息系统的属性分层递阶(Hierarchy of attributes in information systems)

由文献[4]可知下面的定理1,2和3,本文将证明定理4和定理5.

定理1 设 U 为论域, $K_p = (U, P)$ 和 $K_q = (U, Q)$ 是关于 U 的两个知识库,当 $P < Q$ 时, $H(P) > H(Q)$;当 $P = Q$ 时, $H(P) = H(Q)$;当 $P > Q$ 时, $H(P) < H(Q)$.

定理2 设知识 Q 相对于知识 P 的条件熵为 $H(Q/P)$, $H(Q/P) = 0$ 的充要条件为 $P < Q$ 或 $P = Q$.

定理3 设 U 为一个论域, $K_p = (U, P)$ 和 $K_q = (U, Q)$ 是关于 U 的两个知识库,且 $H(Q/P) = 0$;当 $H(P) > H(Q)$ 时, $P < Q$;当 $H(P) < H(Q)$ 时, $P > Q$;当 $H(P) = H(Q)$ 时, $P = Q$.

定理4 信息系统中,属性分层递阶并不改变整个系统的信息熵;当信息系统中的知识粒度减小时,知识层次加深时,信息系统的信息熵单调增加;反之亦然.

证 以分为两层的信息系统为例,在每一论域 U_{2i} 上的信息熵:

$$\begin{aligned}
 H(X_i Y) &= \\
 &= - \sum_j p(X_i Y_j) \log p(X_i Y_j) = \\
 &= - \sum_j p(X_i) p(Y_j/X_i) \log p(X_i) p(Y_j/X_i) = \\
 &= - \sum_j p(X_i) p(Y_{ij}) \log p(X_i) p(Y_{ij}) = \\
 &= - \sum_j p(X_i) p(Y_{ij}) \log p(X_i) - \sum_j p(X_i) p(Y_{ij}) \log p(Y_{ij}) = \\
 &= - p(X_i) \log p(X_i) - p(X_i) \sum_j p(Y_{ij}) \log p(Y_{ij}) = \\
 &= H(X_i) + p(X_i) H(Y_i). \quad (1)
 \end{aligned}$$

其中

$$\begin{aligned}
 H(Y_i) &= - \sum_j p(Y_{ij}) \log p(Y_{ij}), \\
 i &= 1, \dots, n, j = 1, 2, \dots, m. \quad (2)
 \end{aligned}$$

同理可得

$$H(XY) = \sum_i [H(X_i) + p(X_i) H(Y_i)], \quad (3)$$

$$\begin{aligned}
 H(XY) &= \sum_i H(X_i) + \sum_i p(X_i) H(Y_i) = \\
 &= H(X) + \sum_i p(X_i) H(Y_i). \quad (4)
 \end{aligned}$$

1) 首先,由式(4)可知,属性分层递阶并不改变整个系统的信息熵。

2) 其次,根据定理1可得,当 $(XY) < X$ 时, $H(XY) \geq H(X)$, 即当信息系统中的知识粒度减小时,信息系统的信息熵单调增加。

根据定理2可知,当 $(XY) < X$ 时, $H(X/XY) = 0$; 又根据定理3得,当 $H(XY) \geq H(X)$ 时, $(XY) < X$, 即当信息系统的信息熵单调增加时,信息系统中的知识粒度减小。

所以,当信息系统中的知识粒度减小时,系统信息熵单调增加;反之亦然。

3) 最后,由概率及熵函数的非负性^[6]可知,式(2)中的 $H(Y_i) \geq 0$; 则由式(4)可得,当信息系统中知识层次加深时,系统信息熵单调增加。

再来证明,当信息系统中的信息熵增加时,知识粒度单调减小,知识层次单调加深。

假设,分层递阶中,信息系统的信息熵单调增加时,知识的粒度不变或者增加,知识的层次不变或者减小。

则由式(4)可知,当知识粒度和知识层次不变

时,信息系统的信息熵也不变;当知识粒度增加,知识层次减小时,系统的信息熵减小;均与假设相矛盾.从而证明,当信息系统中的信息熵增加时,知识粒度单调减小,知识层次单调加深. 证毕。

定理5 分层递阶约简算法中,信息系统的信息熵保持不变,信息系统的平均知识粒度和平均知识层次保持不变。

证 1) 由定理4可知:信息系统的信息熵与整个信息系统的知识粒度和知识层次之间具有相互的单调关系;所以,信息系统的信息熵可以表示整个信息系统的平均粒度和平均层次。

2) 同时,由定理4可知:属性分层递阶的信息系统中,属性分层递阶并不改变整个系统的信息熵;另由文献[4]可知,知识约简过程中没有信息损失,即知识约简不会改变系统的信息熵.所以,属性分层递阶后再进行约简并不会改变信息系统的信息熵,即不会改变整个信息系统的平均粒度和平均层次。

证毕。

3.2 决策系统的属性分层递阶 (Hierarchy of attributes in decision systems)

由文献[4]亦可知下面的定理6和7,本文将证明定理8和定理9。

定理6 设 U 为一个论域, $K_p = (U, P)$ 和 $K_q = (U, Q)$ 是关于 U 的两个知识库, D 是 U 上的决策. 当 $P < Q$ 时, $I(P; D) > I(Q; D)$; 当 $P = Q$ 时, $I(P; D) = I(Q; D)$; 当 $P > Q$ 时, $I(P; D) < I(Q; D)$ 。

定理7 设 U 为一个论域, $K_p = (U, P)$ 和 $K_q = (U, Q)$ 是关于 U 的两个知识库, 且 $H(Q/P) = 0$. 当 $I(P; D) < I(Q; D)$, $P > Q$; 当 $I(P; D) > I(Q; D)$ 时, $P < Q$; 当 $I(P; D) = I(Q; D)$ 时, $P = Q$ 。

定理8 决策系统中,属性分层递阶并不改变整个系统的互信息;当决策系统中的知识粒度减小时,知识层次加深时,系统的互信息单调增加;反之亦然。

证

$$\begin{aligned}
 I(D_k; X_i Y) - I(D_k; X_i) &= \\
 &= [H(X_i Y) + H(D_k) - H(X_i D_k Y)] - \\
 &= [H(X_i) + H(D_k) - H(X_i D_k)] = \\
 &= [H(X_i Y) - H(X_i)] - [H(X_i D_k Y) - H(X_i D_k)]. \quad (5)
 \end{aligned}$$

由式(1)可得

$$\begin{aligned}
 [H(X_i Y) - H(X_i)] &= p(X_i) H(Y_i), \quad (6) \\
 [H(X_i D_k Y) - H(X_i D_k)] &=
 \end{aligned}$$

$$p(X_i D_k) H(Y_{ik}) = p(X_i) p(D_k/X_i) H(Y_{ik}). \quad (7)$$

将式(6)和式(7)代入式(5)得

$$\begin{aligned} I(D_k; X_i Y) = \\ I(D_k; X_i) + p(X_i) [H(Y_i) - p(D_k/X_i) H(Y_{ik})]. \end{aligned} \quad (8)$$

其中, $H(Y_i)$ 如式(2)所示.

$$H(Y_{ik}) = - \sum_j p(Y_{ikj}) \log p(Y_{ikj}),$$

$$i = 1, \dots, n, k = 1, \dots, r, j = 1, 2, \dots, m.$$

根据分层递阶原理易知

$$H(Y_i) - p(D_k/X_i) H(Y_{ik}) \geq 0. \quad (9)$$

同理可得

$$\begin{aligned} I(D; XY) = \\ I(D; X) + \sum_i p(X_i) [H(Y_i) - \sum_k p(D_k/X_i) H(Y_{ik})]. \end{aligned} \quad (10)$$

其中

$$H(Y_i) - \sum_k p(D_k/X_i) H(Y_{ik}) \geq 0.$$

其余证明与定理4同理.

定理9 分层递阶约简算法中, 决策系统的互信息保持不变, 决策系统的平均知识粒度和平均知识层次保持不变.

证 与定理5同理可证, 略.

3.3 分层递阶约简算法的性质 (Properties of the hierarchical reduction approach)

分层递阶约简算法中, 信息系统的信息熵和决策系统的互信息保持不变, 即分层递阶约简算法不会引起信息系统和决策系统的信息损失. 其性质如下:

1) 较强的实用性. 分层递阶约简所得到的简约避免应用那些获取方式复杂, 采集成本高, 实时性要求高的属性. 同时, 这些属性也是导致信息系统不完备性的重要原因; 所以, 分层递阶约简可以用于处理信息系统不完备的情况, 所得分层递阶简约也更适合应用于实际的不完备信息系统^[7].

2) 较好的动态特性. 当论域中有新的对象加入时, 可以逐层将对象与分层递阶简约进行匹配; 只要对由于该对象的加入而引起变化的相应层次以下的各论域进行约简, 即可更新分层递阶简约.

3) 算法速度快, 并且可以并行计算. 分层递阶约简算法极大地减少了粗糙集理论的约简计算量, 提高了约简速度. 并且, 各层约简可以实行并行计算.

粗糙集理论求得全部简约的约简计算量为 $O(2^{|A|} |A| |U|^2)$ ^[8]. 分层递阶约简算法的约简计

算量为

$$Q = 2^{|A_1|} |A_1| |U|^2 + \sum_i 2^{|A_2|} |A_2| |U_i|^2.$$

其中, $|A_1|$ 和 $|A_2|$ 分别为首层 A_1 和次层 A_2 的属性数量, $A_1 \cup A_2 = A$, $A_1 \cap A_2 = \emptyset$, $|A_1| + |A_2| = |A|$; $|U_i|$ 为次层中的对象数量. 可以证明, $Q < O(2^{|A|} |A| |U|^2)$ (证明略), 即分层递阶约简算法减少了粗糙集的理论约简计算量, 提高了约简速度.

4 实例分析 (Case study)

本节应用分层递阶约简算法, 重新分析文献[2]中司炉工人在某水泥窑炉运行中的操作记录. 该操作记录共52条; 其中, 燃烧区温度 BZT, 燃烧区颜色 BZC, 渣块颗粒 CG 和炉内颜色 KIC 分别记做条件属性 a, b, c 和 d ; 窑炉转速 KR 和给煤量 CWR 分别记为结论属性 e 和 f .

文献[2]中的约简结果表明属性 b 冗余. 且其两个简约 (I) 和 (II) 分别为

$$\begin{cases} a_3(d_1 \vee d_2) \rightarrow e_2 f_4, \\ a_2 \rightarrow e_1 f_4, \\ c_2 d_3 \rightarrow e_2 f_3, \\ c_3 \rightarrow e_2 f_2, \end{cases} \quad (\text{I})$$

$$\begin{cases} a_3(d_1 \vee d_2) \rightarrow e_2 f_4, \\ a_2 \rightarrow e_1 f_4, \\ a_3 d_3 \vee a_4 c_2 \rightarrow e_2 f_3, \\ c_3 \rightarrow e_2 f_2. \end{cases} \quad (\text{II})$$

4.1 分层递阶约简 (Hierarchical reduction)

对水泥窑炉实际生产情况的调查表明: 在4个条件属性中, 燃烧区温度 BZT 是由温度传感器采集的, 采集方式最为简单; 而燃烧区颜色 BZC, 渣块颗粒 CG 和炉内颜色 KIC 都需要通过特殊摄像机进行观察, 采集成本都比较高. 这样, 我们可以利用分层递阶约简将条件属性分为 C_1 和 C_2 两层. 其中首层 C_1 包括条件属性 a , 即燃烧区温度 BZT; 次层包括条件属性 b, c 和 d , 即燃烧区颜色 BZC, 渣块颗粒 CG 和炉内颜色 KIC.

通过对首层决策系统进行约简, 得到

$$\begin{cases} a_2 \rightarrow e_1 f_4, \\ a_3 \rightarrow e_2 f_4 \vee e_2 f_3, \\ a_4 \rightarrow e_2 f_3 \vee e_2 f_2. \end{cases}$$

分别对 $a = 3$ 和 $a = 4$ 构成的次层决策系统进行约简得到

$$a = 3 \text{ 时, } \begin{cases} d_3 \rightarrow e_2 f_3, \\ d_1 \vee d_2 \rightarrow e_2 f_4; \end{cases}$$

$$a = 4 \text{ 时, } \begin{cases} c_2 \rightarrow e_2 f_3, \\ c_3 \rightarrow e_2 f_2. \end{cases}$$

最终,分层递阶简约为

$$\begin{cases} a_2 \rightarrow e_1 f_4, \\ a_3 \rightarrow \begin{cases} d_3 \rightarrow e_2 f_3, \\ d_1 \vee d_2 \rightarrow e_2 f_4, \end{cases} \\ a_4 \rightarrow \begin{cases} c_2 \rightarrow e_2 f_3, \\ c_3 \rightarrow e_2 f_2. \end{cases} \end{cases} \quad (\text{III})$$

4.2 分层递阶简约的优点 (Virtue of the hierarchical reduct)

分层递阶简约(III)比上文中简约(I)和(II)更适用于水泥窑炉的生产实际.这是因为,应用简约(I)和(II)时,需要同时采集 a 、 c 和 d 三个条件属性,然后才能做出控制决策.而如图1所示,应用分层递阶简约(III)时,在采集得到 a 后就可以做出控制决策或者至少可以判断应该采集 c 还是 d 才能得出控制决策.这样,通过充分利用采集方式简单的条件属性 a ,可以减少对采集成本较高的条件属性 c 和 d 的采集,从而达到节省决策时间和决策成本的目的.

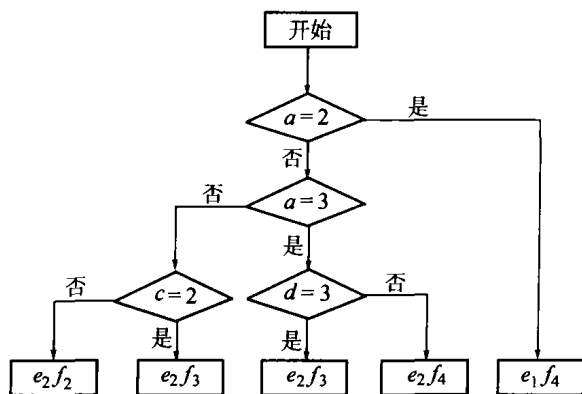


图1 分层递阶简约的应用

Fig. 1 Application of the hierarchical reduct

5 结论 (Conclusion)

为适应数据库的动态数据变化,本文将研究粗糙集理论的动态分层递阶约简算法.同时,也将研究粗糙集理论与模糊逻辑、神经网络和遗传算法等人工智能理论的融合.

参考文献 (References):

- [1] PAWLAK Z. Rough sets [J]. *Communications of ACM*, 1995, 38 (11): 89-95.
- [2] PAWLAK Z. *Rough Set: Theoretical Aspects of Reasoning about Data* [M]. Dordrecht: Kluwer Academic Publishers, 1991.
- [3] 张钹,张铃.问题求解理论及应用[M].北京:清华大学出版社,北京:1990.
(ZHANG Bo, ZHANG Ling. *Theory and Applications of Problem Solving* [M]. Beijing: Tsinghua University Press, 1990.)
- [4] 苗夺谦. Rough Set 理论及其在机器学习中的应用研究[D].北京:中国科学院自动化研究所,1997.
(MIAO Duoqian. *The Study of Application of RST in Machine Learning* [D]. Beijing: Institute of Automation Chinese Academy of Science, 1997.)
- [5] 刘发升,杨炳儒.一种基于粗糙集的多层次、逐步求精的发掘算法[J]. *计算机工程与应用*, 1999, 25(5): 11-12, 17.
(LIU Fasheng, YANG Binru. A stepwise data mining method based on rough sets [J]. *Computer Engineering and Applications*, 1999, 25 (5): 11-12, 17.)
- [6] 姜丹,钱玉美.信息理论与编码[M].北京:中国科学技术出版社,1992.
(JIANG Dan. *Information Theory and Coding* [M]. Hefei: University of Science and Technology of China Press, 1992.)
- [7] 乔斌,郑洪涛,郭智疆,等.针对信息系统不完备性的粗糙集分层递阶约简[J]. *电路与系统学报*, 2001, 6(2): 78-82.
(QIAO Bin, ZHENG Hongtao, GUO Zhijiang, et al. The hierarchical reduction approach to incomplete information system [J]. *J of Circuits and Systems*, 2001, 6(2): 78-82.)
- [8] GUAN J W, BELL D A. Rough computational methods for information systems [J]. *Artificial Intelligence*, 1998, 105(1): 77-103.

作者简介:

乔斌 (1975—),男,现为浙江大学电气工程学院博士研究生.研究方向为粗糙集理论,数据挖掘和信息融合. E-mail: peakqiao@126.com;

李玉榕 (1973—),女.2001年在浙江大学获博士学位,福州大学讲师.主要研究方向为信息融合,进化计算和粗糙集理论. E-mail: lyrxj@126.com;

蒋静坪 (1935—),男.浙江大学电气工程学院教授,博士生导师.主要研究方向为智能控制和计算机控制. E-mail: eejiang@zju.edu.cn.