

一种基于改进粗糙集模型的归纳学习方法

洪菁, 陈强, 刘惠彬

(上海工程技术大学 计算中心, 上海 201620)

摘 要:对传统的粗糙集理论进行了扩展,提出了一种改进的粗糙集归纳学习方法。一方面,针对连续属性离散化,利用模糊集理论对连续属性进行模糊化,再根据模糊贴适度构造模糊相似矩阵,并用 $k-w$ 方法粗略评估各连续属性的重要度,建立基于模糊相似关系的划分,最终生成相容的决策表。另一方面,针对解决最优属性的选择问题,提出一种加权求和的属性重要度定义。基于以上模型开发了一个原型系统,并以一个工程实例验证了此方法的有效性。

关键词:离散化;粗糙集;模糊相似关系;属性重要度;归纳学习

中图分类号:TP18

文献标识码:A

文章编号:1673-629X(2006)10-0032-03

An Inductive Learning Approach Based on Modified Rough Set

HONG Jing, CHEN Qiang, LIU Hui-bin

(Computer Center, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: In this paper, an inductive learning approach based on modified rough set is proposed. Firstly, the continuous attributes in the decision table are fuzzified with the proper fuzzy membership functions, and the fuzzy similar matrix of the attributes is constructed with the fuzzy degree of nearness, then the $k-w$ method is applied to evaluate the relative importance of every continuous attribute. The continuous decision table is discretized into a compatible table based on the fuzzy similarity relation. Secondly, an improved definition of the attribute significance based on the weighed sum is proposed. A prototype system based on the proposed approach is developed. Finally, an engineering example proves the effectiveness and feasibility of the proposed method.

Key words: discretization; rough set; fuzzy similarity relation; attribute significance; inductive learning

粗糙集作为一种处理不确定、不完全信息的有效工具,是知识获取的重要手段之一^[1]。但其在应用领域还存在诸多不足之处,首先,粗糙集只能从离散属性组成的决策表中挖掘有用的知识、辅助决策,而现实决策表中往往含有连续属性,常规的连续属性人为区间划分方法会造成明显的边界问题,目前粗糙集缺乏有效的连续属性离散化方法。其次,复杂决策表的约简一般并不是唯一的,求取最小约简是NP问题^[2]。

针对上述问题,文中首先提出一种连续属性的离散化方法。该方法先对连续属性的模糊化,在保证属性离散后决策表是相容的前提下,建立基于相似关系的论域划分。其次,对决策表最小约简问题,文中给出了基于加权求和的属性重要度定义以解决约简的最优属性选择问题。最后,基于以上思想,提出了一种改进的粗糙集归纳学习算法,开发了原型系统并以一个工程实例验证此方法的有效性。

1 连续属性离散化

1.1 连续属性重要性度量

在粗糙集理论中,已有一些概念可以评价属性的重要度,但它们只适用定性属性,还存在连续属性的离散化问题。文中选择模式识别中的 $k-w$ 方法粗略评估系统属性的重要度^[3]。 $k-w$ 方法评价某个属性 A_i 对决策属性的重要度的具体步骤如下:

1) 按一定顺序(如从小到大)列举对应此属性 A_i 的取值,并给每个取值编号,如最小编号为1,较大者为2,依此类推。

2) 对属性 A_i ,计算每类对象编号的平均值 P_1, P_2, \dots, P_n 。

3) 对属性 A_i ,计算其统计量 HP_i (属性对应编号的组间离差):

$$HP_i = \frac{12}{N(N-1)} \sum_{i=1}^n N_i (P_i - \frac{N+1}{2})^2 \quad (1)$$

其中 N 表示决策表包含对象个数, N_i 表示第 i 类包含对象个数。

对应其他属性,依次按照上述步骤计算统计量。比较各属性的 H 值,当 H 越大时,该属性相对于决策属性越重要,分类能力越强。

收稿日期:2005-12-28

基金项目:国家“八六三”计划资助项目(2002AA134020);上海市高校青年教师科研专项基金(05XPYQ45)

作者简介:洪菁(1973-),女,江苏人,助教,硕士,主要从事智能信息处理、数据挖掘方面的研究。

1.2 基于模糊相似关系的论域划分

传统粗糙集中的论域划分是基于等价关系的,在实际应用中,可以放松集合关系中的约束条件,即去除传递性,把等价关系推广为模糊相似关系。

定义 1 模糊关系 $\tilde{R} \in F(U \times U)$ 为 U 上的模糊相似关系,如果 \tilde{R} 满足如下条件:

- * 自反性: $\mu_{\tilde{R}}(u, u) = 1, \forall u \in U$
- * 对称性: $\mu_{\tilde{R}}(u, v) = \mu_{\tilde{R}}(v, u), \forall u, v \in U$

定义 2 $U = \{u_1, u_2, \dots, u_n\}$ 为有限论域, U 上的模糊相似关系 $\tilde{R} \in R_{n \times n}$ 称为模糊相似矩阵,即 \tilde{R} 是个模糊矩阵,且满足如下条件:

- * 自反性: $r_{ii} = 1, i = 1, 2, \dots, n$
- * 对称性: $r_{ij} = r_{ji}, i, j = 1, 2, \dots, n$

为了建立模糊相似关系,必须引入模糊相似关系的度量,即计算相似系数 r_{ij} 的方法,通常有最大最小法、几何平均最小法、夹角余弦法等。文中采用基于明可夫斯基距离的贴进度计算模糊相似系数。

定义 3 已知 $\tilde{R} \in R_{n \times n}$ 称为模糊相似矩阵,引入一置信水平 λ ,经过如下的操作,得到模糊矩阵 $\tilde{R}_\lambda, \tilde{R}_\lambda$ 成为置信水平 λ 下的普通相似关系矩阵。

$$\begin{cases} r_{ij} = 1, & r_{ij} \geq \lambda \\ r_{ij} = 0, & r_{ij} < \lambda \end{cases} \quad i, j = 1, 2, \dots, n$$

文中提出的连续属性离散化的指导原则是在保证离散后决策表相容的前提下,采用尽可能少的离散值。其基本思想是:先用 $k-w$ 方法初步评价各连续属性的重要度;然后用隶属度函数对连续属性模糊化,求得各属性的模糊相似矩阵;再引入置信水平 λ ,求得各属性的普通相似关系矩阵;最后根据编网法原理^[4]得到各属性的模糊相似关系对整个论域的划分。若决策表相容,离散化过程结束,否则先增大重要属性的置信水平 λ ,其次再考虑增大其他次要属性的置信水平 λ 。如此重复,直至决策表相容为止。此离散化方法与其它方法相比^[5],具有以下优点:

1)不同的条件属性相对决策属性的重要性不同,在离散过程中不应同等对待。通常,重要的属性对决策影响较大,分类区间应多一些。

2)离散后得到的决策表是相容的。

3)论域划分基于模糊相似关系,避免了传统粗糙集论域划分过细的缺点。

2 基于加权粗糙集属性重要度定义

目前,属性重要度的标准主要有两种^[6]:基于依赖度的属性重要度和基于信息熵的属性重要度。但是,基于依赖度的属性重要度只考虑了论域中确定性元素集合,而忽略了边界域中的元素概率分布的信息;基于信息熵的属性重要度刻画了边界域中不确定元素集合提供的信息,而忽略了知识的粒度。文献[7]论证了单独这两种属性重要度定义的不完备性。因此综合考虑这两方面,提出一种基于加权粗糙集属性重要度标准。

在完整信息系统中,即决策表 $T = \langle U, C \cup D, V, f \rangle$ 中不含有不确定的元素,即 D 的 $R \subseteq C$ 上近似为 U ,记为 $\bar{R}(D) = U$ 。由粗糙集理论得,

$$BN_R(D) = U - \bar{R}(D) \quad (2)$$

定义决策表 T 中属性 R 下的确定性元素比例和不确定性元素比例分别为:

$$\omega_1(R, D) = \text{card}(\text{POS}_R(D)) / \text{card}(U) \quad (3)$$

$$\omega_2(R, D) = \text{card}(BN_R(D)) / \text{card}(U) \quad (4)$$

由式(3),(4)可简化为:

$$\omega_2(R, D) = 1 - \omega_1(R, D) \quad (5)$$

文中使用加权和方法对两种常用标准进行综合考虑。由于 $0 \leq \gamma(R, D) \leq 1$, 而 $0 \leq H(D|R) \leq \log_2 n$, 其中 $n = \text{card}(U)$, 并且 $\gamma(R, D)$ 追求最大化, 而 $H(D|R)$ 追求最小化。因此对 $H(D|R)$ 进行如下的转换:

$$H^*(D|R) = 1 - \frac{H(D|R)}{\log_2 n} \quad (6)$$

然后,构造如下的加权和:

$$\text{STD}(D, R) = \omega_1(R, D) \cdot \gamma(R, D) + \omega_2(R, D) \cdot H^*(D|R) \quad (7)$$

由此,任意属性 $a \in C - R$ 的重要度可以定义如下:

$$\text{SIG}_1^C(a, R, D) = \omega_1(R \cup \{a\}, D) \cdot (\gamma(R \cup \{a\}, D) - \gamma(R, D)) + \omega_2(R \cup \{a\}, D) \cdot (H(D|R) - H(D|R \cup \{a\})) / \log_2 n \quad (8)$$

3 基于改进粗糙集模型的归纳学习算法

根据上面的模糊粗糙集模型和加权粗糙集的属性重要度定义,提出一种基于改进粗糙集模型的归纳学习算法 IR-ILA,其描述如下:

输入:数据库,隶属度函数,置信水平因子

输出:模糊决策模式

- (1)用 $k-w$ 方法评估各属性的重要度;
- (2)利用 Kohonen 网络自组织映射算法确定 k 个模糊划分的中心 m_i , 并采用三角隶属度函数对连续属性进行模糊化;
- (3)对每个属性采用明可夫斯基贴进度计算其模糊相似矩阵;
- (4)根据 $k-w$ 计算结果,确定每个属性的置信水平;
- (5)将每个属性的模糊相似矩阵转化为某置信水平下的普通相似关系矩阵;
- (6)用编网法计算模糊相似关系对整个论域的划分;
- (7)检查离散后的决策表是否相容,若决策表相容,转(9);
- (8)更改某些属性的置信水平,转(5);
- (9)求取条件属性核,并根据加权和属性重要度定义计算决策表的最小约简;
- (10)删除冗余属性,得到条件属性的最小简化,删除重复实例;
- (11)对每个实例求其属性值的核,删除多余的属性

值,得到其最小属性值简化;

(12)再次删除决策表中重复实例,归纳出决策规则。

4 算 例

根据此算法,采用 VC++ 6.0 开发了原型系统。为了验证此方法的有效性,选择表 1 所示的某煤矿瓦斯涌出量数据库为实例。此数据库有 7 个属性:煤层埋藏深度(A),煤层厚度(B),煤层瓦斯含量(C),煤层间距(D),日产量(E),日进度(F),绝对瓦斯涌出量(G)},其中,绝对瓦斯涌出量}为决策属性,其余为条件属性。

首先,对 6 个条件属性用 k-w 方法计算其 H,分别得:煤层埋藏深度:13.7647;煤层厚度:12.9559;煤层瓦斯含量:12.9559;煤层间距:7.43015;日产量:12.9559;日进度:10.5184。

再次,利用三角隶属度函数对 6 个连续属性进行模糊化,利用基于明可夫斯基距离的贴进度分别对 6 个属性建立模糊相似关系,引入置信水平,把模糊相似矩阵转为一般相似矩阵。并求出模糊相似关系对整个论域的划分。当 $\lambda a = 0.95, \lambda b = 0.9, \lambda c = 0.9, \lambda d = 0.7, \lambda e = 0.9, \lambda f = 0.8$ 时,得到属性离散化后的编码依据如下:

* 煤层埋藏深度(1:[408,432],2:[450,456],3:[516,531],4:[544,550],5:[563,563],6:[590,590],7:[604,607],8:[629,629]);

* 煤层厚度(9:[1.8,3],10:[5.9,6.4]);

* 煤层瓦斯含量(11:[1.92,2.58],12:[2.8,2.8],13:[3.16,3.22],14:[3.35,3.35],15:[3.61,3.68],16:[4.03,4.03],17:[4.21,4.21],18:[4.34,4.62]);

* 煤层间距(19:[11,14],20:[16,22]);

* 日产量(21:[1527,1527],22:[1751,1825],23:[1979,2104],24:[2207,2410],25:[3087,3456]);

* 日进度(26:[2.64,2.85],27:[3.28,4.67])。

表 1 所示为某煤矿统计数据来源。

表 1 某煤矿统计数据源

| No | A(m) | B(m) | C(m ³ /t) | D(m) | E(t) | F(%) | G(m ³ /t) |
|----|------|------|----------------------|------|------|------|----------------------|
| 1 | 408 | 2 | 1.92 | 20 | 1825 | 4.42 | 3.34 |
| 2 | 411 | 2 | 2.15 | 22 | 1527 | 4.16 | 2.97 |
| 3 | 420 | 1.8 | 2.14 | 19 | 1751 | 4.13 | 3.56 |
| 4 | 432 | 2.3 | 2.58 | 17 | 2.78 | 4.67 | 3.62 |
| 5 | 450 | 2.2 | 2.43 | 16 | 1996 | 4.32 | 4.06 |
| 6 | 456 | 2.2 | 2.4 | 20 | 2104 | 4.51 | 4.17 |
| 7 | 516 | 2.8 | 3.22 | 12 | 2242 | 3.45 | 4.6 |
| 8 | 527 | 2.5 | 2.8 | 11 | 1979 | 3.28 | 4.92 |
| 9 | 531 | 2.9 | 3.35 | 13 | 2288 | 3.68 | 4.78 |
| 10 | 544 | 2.7 | 3.16 | 13 | 2207 | 3.81 | 4.92 |
| 11 | 550 | 2.9 | 3.61 | 14 | 2325 | 4.02 | 5.23 |
| 12 | 563 | 3 | 3.68 | 12 | 2410 | 3.53 | 5.56 |
| 13 | 590 | 5.9 | 4.21 | 18 | 3139 | 2.85 | 7.24 |
| 14 | 604 | 6.2 | 4.03 | 16 | 3354 | 2.64 | 7.8 |
| 15 | 607 | 6.1 | 4.34 | 17 | 3087 | 2.77 | 7.68 |
| 16 | 629 | 6.4 | 4.62 | 19 | 3456 | 2.8 | 8.04 |

最后通过加权和属性重要的定义对属性进行约简,最终归纳学习得出如图 1 所示的归纳属性简化表。

| 煤层埋藏深度 a | 煤层瓦斯含量 c | 绝对瓦斯涌出量 |
|----------|----------|---------|
| 1 | * | 28 |
| 2 | * | 29 |
| 3 | * | 29 |
| 4 | 13 | 29 |
| 5 | 15 | 30 |
| 6 | * | 31 |
| 7 | * | 31 |
| 8 | * | 31 |

注: *表示该属性值可为任意值

图 1 归纳属性简化结果

根据此归纳属性简化表,得到如下的规则知识:

即:Rule1 IF 408≤煤层埋藏深度≤432 THEN 2.97≤绝对瓦斯涌出量≤3.62

Rule2 IF 450≤煤层埋藏深度≤456 THEN 4.06≤绝对瓦斯涌出量≤4.92

Rule3 IF 516≤煤层埋藏深度≤531 THEN 4.06≤绝对瓦斯涌出量≤4.92

Rule4 IF 煤层埋藏深度 = 590 THEN 7.24≤绝对瓦斯涌出量≤8.04

Rule5 IF 煤层埋藏深度 = 629 THEN 7.24≤绝对瓦斯涌出量≤8.04

Rule6 IF 604≤煤层埋藏深度≤607 THEN 7.24≤绝对瓦斯涌出量≤8.04

Rule7 IF 3.16≤煤层瓦斯含量≤3.22 THEN 4.06≤绝对瓦斯涌出量≤4.92

Rule8 IF 3.61≤煤层瓦斯含量≤3.68 THEN 5.23≤绝对瓦斯涌出量≤5.56

以上 8 条规则对表 1 所包含实例的覆盖率和正确率均为 95%,根据不同的精度要求,通过调整置信水平的大小,可以得到不同数目的规则。

5 结束语

文中首先将模糊概念应用于决策表连续属性的离散化,通过模糊相似矩阵的构造和 k-w 方法对连续属性重要度的粗略评估,在保证离散决策表相容的前提下,实现了基于模糊相似关系的论域划分,克服了传统离散方法的不足,然后提出了一种加权后的属性重要度定义。在这两点基础上构造了一种归纳学习算法,使得到的规则自然、简洁、便于理解。通过工程实例验证了该方法的有效性。

参考文献:

- [1] Pawlak Z. AI and intelligent industrial applications: the rough set perspective[J]. Cybernetics & Systems: An International Journal, 2000, 31(4): 227-252.
- [2] Wong S K M, Ziarko W. On optional decision rules in decision tables[J]. Bulletin of Polish Academy of Sciences, 1985, 33:

(下转第 36 页)

协议根据 IP 数据报中的目的 IP 地址来进行路由选择,一旦决定了如何为一个 IP 数据报选择路径,就将数据报发往所选择的路径中,不需要额外的包头,即不存在额外的封装。OSPF 可以在很短的时间里使路由选择表收敛。OSPF 还能够防止出现回路,这种能力对于网状网络或使用多个网桥连接的不同局域网是非常重要的。

最短路径优先 (SPF) 路由算法^[2]是 OSPF 的基础。当 SPF 路由器加点后,它就初始化路由协议数据结构,然后等待下层协议关于接口已可用的通知信息。当路由器确认接口已准备好,就用 OSPF Hello 协议^[3]来获取邻居信息,即具有在共同的网络上接口的路由器。路由器向邻居发送 Hello 包并接收它们的 Hello 包。除了帮助学习邻居外,Hello 包也有 keep-alive 的功能。

3 OSPF 和 EIGRP 的比较

EIGRP 协议和 OSPF 协议相比具有以下优点:

(1) 路由负载均衡能力^[4]。

EIGRP 可以根据优先级不同,自动匹配流量;而 OSPF 虽然能根据接口的速率、连接可靠性等信息,自动生成接口路由优先级,但通往同一目的的不同优先级路由,OSPF 只选择优先级较高的转发,不同优先级的路由,不能实现负载均衡。只有相同优先级的,才能达到负载均衡的目的。

(2) 配置复杂度。

由于网络区域划分和网络属性的复杂性,需要网络分析员有较高的网络知识水平才能配置和管理 OSPF 网络^[5];而使用 EIGRP 协议组建网络,路由器配置非常简单,它没有复杂的区域设置,也无需针对不同网络接口类型实施不同的配置方法。使用 EIGRP 协议只需使用 router eigrp 命令在路由器上启动 EIGRP 路由进程,然后再使用 network 命令使能网络范围内的接口即可。

(3) 占用带宽。

路由的发送使用增量发送方法,当路径信息改变以后,DUAL 只发送那条路由信息改变了的更新,而不是发送整个路由表。发送的路由更新报文采用可靠传输,如没有收到确认信息则重新发送,直至确认。EIGRP 还可以对发送的 EIGRP 报文进行控制,减少 EIGRP 报文对接口带宽的占用率,从而避免连续大量发送路由报文而影响正常数据业务的事情发生。

(4) 收敛速度。

EIGRP 协议由于使用了 Diffusing Update (DUAL) 算法,EIGRP 在路由计算时,只会对发生变化的路由进行重新计算。路由器使用 EIGRP 来存储所有到达目的地的备份路由,以便进行快速切换。如果没有合适的或备份路由在本地路由表中的话,路由器向它的邻居进行查询来选择一条备份路由,使得路由计算的收敛时间也有好的保证。

当然 EIGRP 协议也具有一定的不足:

a. OSPF 协议是开放的协议,是 IETF 组织公布的标准,而 EIGRP^[6]是 Cisco 公司的私有协议。在一个大型网络中,假如不是所有的设备都是 Cisco 的,EIGRP 明显就不行,因为它是私有的,故只能使用 OSPF 协议或者路由 redistribution (路由协议之间的翻译服务)。

b. OSPF 在大规模网络的情况下,可以通过划分区域来规划和限制网络规模;而 EIGRP 没有区域 (area) 的概念,所以 EIGRP 适用于网络规模相对较小的网络。

4 结 论

OSPF 协议和 EIGRP 协议都是收敛速度较快并且不会形成环路的算法,网络带宽占用较小,使用灵活,安全性较好。但是从以上分析可以看出,EIGRP 协议在路由负载均衡能力、配置复杂度、占用带宽、收敛速度等方面优于 OSPF 协议,而在协议开放性和适用网络规模方面 OSPF 协议更好一些。

参考文献:

- [1] Moy J. OSPF Version 2[S]. Internet Draft, 1992.
- [2] Bertsekas D, Gallager R. Data Networks[M]. Second Edition [s. l.]: Prentice-Hall, Inc, 1992.
- [3] Zaumen W T, Garcia-Luna-Aceves J J. Dynamics of Link-State and Loop-Free Distance-Vector Routing Algorithms [J]. Journal of Internetworking, 1992, 3: 161-188.
- [4] Garcia-Luna-Aceves J J, Zaumen W T. Extensions to the Diffusing Update Algorithms for Area Routing in Computer Networks and Internetworks, Invention Disclosure[Z]. SRI International, Menlo Park, CA: [s. n.], 1993.
- [5] Apostolopoulos G, Williams D, Guerin R, et al. QoS routing mechanisms and OSPF extensions[S]. Internet Request for Comment. RFC 2676, 1999.
- [6] Hill B. Cisco 完全手册[M]. 北京: 电子工业出版社, 2002.

(上接第 34 页)

693-696.

- [3] 曾 谦, 曾黄麟. 系统参数重要性评价方法[J]. 四川轻工业学院学报, 1999, 12(2): 10-13.
- [4] 赵汝怀. 弗晰聚类的编网法[J]. 西安交通大学学报, 1980 (4): 43-47.
- [5] 王 珏, 王 任, 苗夺谦. 基于 Rough Set 理论的数据浓缩 [J]. 计算机学报, 1998, 21(5): 393-399.

- [6] Wong S K M, Ziarko W, Li Y R. Comparison of rough-set and statistical methods in inductive learning[J]. Int J of Man-Machine Studies, 1986, 24: 53-73.
- [7] Shi F, Lou Z L, Zhang Y Q. An improved strategy for attribute reduction in rough set[A]. the Sixth International Conference for Young Computer Scientists[C]. [s. l.]: [s. n.], 2001. 41-44.