基于粗糙变精度的食品安全决策树研究

鄂 旭[1,2],任骏原[1],毕嘉娜[1],沈德海[1]
(1. 渤海大学 信息科学与技术学院,辽宁 锦州 121001;
2. 中国产业安全研究中心,北京 100084)

摘 要:食品安全决策是食品安全问题研究的一项重要内容。为了对食品安全状况进行分析,基于粗糙集变精度模型,提出了一种包含规则置信度的构造决策树新方法。这种新方法针对传统加权决策树生成算法进行了改进,新算法以加权平均变精度粗糙度作为属性选择标准构造决策树,用变精度近似精度来代替近似精度,可以在数据库中消除噪声冗余数据,并且能够忽略部分矛盾数据,保证决策树构建过程中能够兼容部分存在冲突的决策规则。该算法可以在生成决策树的过程中,简化其生成过程,提高其应用范围,并且有助于诠释其生成规则。验证结果表明该算法是有效可行的。

关键词:决策树;粗糙集;变精度;置信度
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2014)01-0242-04
doi:10.3969/j.issn.1673-629X.2014.01.062

Research on Decision Tree for Food Safety Based on Variable Precision Rough Sets

E Xu[1,2], REN Jun-yuan[1], BI Jia-na[1], SHEN De-hai[1]
(1. College of Information Science & Technology, Bohai University, Jinzhou 121001, China;
2. China Center for Industrial Security Research, Beijing 100084, China)

Abstract: Food safety decision is an important content of food safety research. Based on variable precision rough sets model, a method of building decision tree with rules that have definite confidence is proposed for food safety analysis. It is an improvement for decision tree inducing approach presented in traditional methods. Present a new algorithm for constructing decision tree with variable precision weighted mean roughness as the criteria for selecting attribute. The new algorithm used variable precision approximate accuracy instead the approximate accuracy. Noisy data of training sets are considered enough. Limited inconsistency is allowed to existed examples of the positive regions. So the decision tree is simplified and its extensive ability is improved and more comprehensible. Experiments show that the algorithm is feasible and effective.

Key words: decision tree; rough sets; variable precision; confidence

0 引 言

食品安全是世界各国人民共同关注的一个重要问题,而食品安全决策又是食品安全问题中的一项重要课题[1-2]。决策树是决策分析的利器,是数据挖掘中一类重要的数据分析方法,因其高效的性能、直观的知识表示和容易理解的规则等优势而得到广泛的应用,文中将其应用于食品安全状况分析中。在决策树的构造过程中,为了能够得到一棵简洁明了的树,需要根据某种标准选择合适的属性作为划分节点,这是决策树问题的一个重点和难点。当前,国内外学者已经研究出许多决策树的构造方法,具有代表性的方法有基于信息熵的 ID3, C4.5 等算法,还有 SLIQ、CART、SPRINT、CHAID 等决策树构造方法[3-4]。

粗糙集理论是于 1982 年由波兰华沙理工大学的 Pawlak 教授提出来的,它是一种应用集合理论来处理内涵模糊、外延不确定的知识的数据处理新工具[5-7]。

收稿日期:2013-06-01 **修回日期**:2013-09-12 **网络出版时间**:2013-11-12
基金项目:中国博士后基金项目(2012M520158);辽宁省百千万人才基金择优资助项目(2012921058);辽宁省教育科研项目(L2012397, L2012396, L2012400)
作者简介:鄂 旭(1971-),男,博士后,教授,硕士生导师,辽宁省“百千万人才”、中国教育部学位与研究生管理中心专家,中国科技部国际科技合作计划评价专家,中国科技奖励评审专家,中国人工智能学会智能 CAD 与数字艺术专业委员会委员,研究方向为物联网智能计算与食品安全信息化。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20131112.1627.006.html

粗糙集理论具有能够直接从给定问题的描述集合出发找出研究问题内在本质的特点,而不需要预先给定研究问题的特征和属性描述,其解决问题基本思想更加接近实际情况[8-10]。目前,学术界已经有了许多应用粗糙集理论来构造决策树的新方法,并且其性能远优于传统的基于信息论的决策树构造方法。目前的基于粗糙集的决策树构建方法都是基于经典的 Pawlak 粗糙集模型,主要是基于明确区域、加权平均粗糙度、近似分类精度和近似分类质量等属性选择标准[11-12]。这些分类方法只能够处理精确的分类问题,缺少兼容某种近似数据的能力,但在实际项目的数据中,特别是在面向主题分析的数据仓库中需要处理大量含噪声的数据。这样,传统粗糙集模型在构造决策树的时候,受到噪声数据的影响很大,常常使得决策规则的结果变得过于细致,最终使得决策树包含过多的决策分支,结构过于冗余复杂。为此,研究人员基于不同视角对传统的粗糙集理论进行了改进和扩展,通过设定精度因子,弱化了对不可分辨关系的严格要求,提出了可变精度粗糙集,因此,可变精度粗糙集方法具有更强的泛化能力和抗噪声能力。

文中在变精度粗糙集理论的基础上,以加权平均变精度粗糙度作为属性选择标准,提出具有确切置信度规则的决策树新方法。

1　选择属性标准

定义 1:加权平均粗糙度。

$$\gamma_R(i) = 1 - \left(\sum_{j=1}^{m} \omega_j \mu_R(X_j)\right) \tag{1}$$

其中,$\mu_R(X_j) = \mathrm{card}(\underline{R(X_j)})/\mathrm{card}(\overline{R(X_j)})$;$i$ 是属性集合中的第 i 个条件属性;j 是属性集合中的第 j 个决策等价类别;m 是在决策属性中等价划分个数;X_j 是第 j 个等价类的决策属性集合。

从上述公式中可以看出,等价关系 R 下论域中少数元素会对集合 X 的近似精度产生较大的影响。如:$U/R = \{\{x_1,x_2,\cdots,x_{10}\},\{x_{11},x_{12},\cdots,x_{18}\},\{x_{19},\cdots,x_{30}\}\}$,$X = \{x_{12},x_{13},x_{14},x_{15},x_{16},x_{17},x_{18},x_{19}\}$,计算等价关系 R 下集合 X 的近似精度 $\alpha_R(X) = \mathrm{card}(\underline{R(X)})/\mathrm{card}(\overline{R(X)}) = 0/20 = 0$。如划分类别的误差设定为 $\beta = 0.2$,则根据等价关系 R 进行计算,$\mu_R^{\beta}(X) = \mathrm{card}(\underline{R^{\beta}(X)})/\mathrm{card}(\overline{R^{\beta}(X)}) = 8/8 = 1$ 为有关研究对象集 X 的近似精度值。

定义 2:加权平均变精度粗糙度。

$$\gamma_R^{\beta}(i) = 1 - \left(\sum_{j=1}^{m} w_j \mu_R^{\beta}(X_j)\right) \tag{2}$$

其中,$\mu_R^{\beta}(X) = \mathrm{card}(\underline{R^{\beta}(X)})/\mathrm{card}(\overline{R^{\beta}(X)})$;$\omega_j = \mathrm{card}(X_j)/\mathrm{card}(U)$;$i$ 是属性集合中的第 i 个条件属性;j 是属性集合中的第 j 个决策等价类别;m 是在决策属性中等价划分个数;X_j 是第 j 个等价类的决策属性集合。

$\gamma_R^{\beta}(i)$ 的值域是[0,1],$\gamma_R^{\beta}(i)$ 反映第 i 个属性包含的确定性,值越小则其包含的确定性因素越大。为此,在决策树的构造过程中,每次划分节点只选择值最小的属性,这样既避免了决策树的过分细化分类,又大大提高了决策树的普适泛化能力。

2　生成具有确切置信度的决策规则

可变精度粗糙集在分类问题的处理上不同于传统粗糙集,在对数据实例进行分类时,它允许存在置信度为 $1-\beta$ 的误差。目前以变精度粗糙集模型为基础来构造决策树的许多方法中,多数条件下只是设其置信度小于 $1-\beta$,而对于获取到的规则来说,却不能够确定每条规则的精确置信度。为此,给出具有确切置信度决策规则的主要生成思想:

设 $R \subseteq C$,$U/R = \{X_1,\cdots,X_i,\cdots,X_p\}$,$U/D = \{Z_1,\cdots,Z_j,\cdots,Z_q\}$。对于 X_i,如果存在 Z_j 使得 $\Pr(Z_j/X_i) \geqslant 1-\beta$,则把 X_i 归属于 Z_j 所在的同一个类,作为叶子节点不再进行细分,并将这条从根节点到叶节点所得的规则置信度设置为 $\Pr(Z_j/X_i)$。即使决策树的众多节点中包含有一定容量的冲突数据类别,但由于设定了其错误分类率小于或等于 β,从而可以很好地保证决策树分类的精度。

3　决策树算法描述

决策树生成算法如下:

输入:训练实例集 U,条件属性集 C,决策属性集 D,给定阈值 β($0 \leqslant \beta < 0.5$);

输出:具有确切置信度规则的决策树。

(1)初始化根节点,设置阈值参数 β;

(2)若 C 为空,则转入(10);

(3)若 U 中所有样本都已经被分类到相应的带标记的叶节点,则转入(10);

(4)若 U 中所有样本不存在没有标记的节点,则转入(10);

(5)若 U 中存在没有标记的节点,则可任意选择一个没有标记的节点,计算 $U'/D = \{Z_1',\cdots,Z_t',\cdots,Z_q'\}$,其中 U' 是所有这个节点上的样本集合;

(6)所选节点 U' 中如果 $\Pr(Z_j/X_i) \geqslant 1-\beta$,即在错误分类率 β 下所有样本都可归属于同一个类别,则可以标记这个节点为叶节点,并把 $\Pr(Z_t'/U')$ 视作为根节点到该叶节点的这条分类规则的置信度,然后转入(5),否则转入(7);

(7)依据加权平均变精度粗糙度定义,对于所选节点中的各个条件属性,分别计算它的加权平均变精度粗糙度;

(8)选取具有最小值的属性作为关键分支节点;

(9)根据关键分支节点对 U_j 进行等价划分,构建决策树分支,进而得到各分支相应的 U',转入(2);

(10)输出最终决策树。

4 实例分析

表 1 为一个食品安全信息表,文中将利用新算法对其进行详细分析,其中 $U=\{1,2,\cdots,26\}$ 是对象集合,$\{A,B,C,D\}$ 是条件属性集,$\{E\}$ 是决策属性集。其各个属性的等价类如下:

表 1 信息表

U	A	B	C	D	E
1	E	G	2	0	L
2	E	M	4	3	L
3	E	E	2	3	R
4	E	P	2	3	R
5	G	G	2	0	L
6	G	G	2	3	B
7	G	M	2	1	L
8	G	M	5	0	L
9	G	E	2	2	R
10	G	G	3	0	L
11	G	G	2	0	L
12	M	E	2	2	R
13	M	E	3	0	L
14	M	E	3	1	B
15	M	E	3	1	B
16	M	E	3	2	R
17	M	E	3	3	L
18	M	E	2	0	R
19	P	M	4	0	L
20	P	G	4	0	R
21	G	M	2	1	L
22	G	E	2	2	R
23	G	E	2	3	R
24	G	M	5	0	L
25	G	P	5	1	L
26	G	P	3	1	L

$U/A=\{\{1,2,3,4\},\{5,6,7,8,9,10,11,21,22,23,24,25,26\},\{12,13,14,15,16,17,18\},\{19,20\}\}$

$U/B=\{\{3,9,12,13,14,15,16,17,18,22,23\},\{1,5,6,10,11,20\},\{2,7,8,19,21,24\},\{4,25,26\}\}$

$U/C=\{\{1,3,4,5,6,7,9,11,12,18,21,22,23\},\{10,13,14,15,16,17,26\},\{2,19,20\},\{8,24,25\}\}$

$U/D=\{\{1,5,8,10,11,13,18,19,20,24\},\{7,14,15,21,25,26\},\{9,12,16,22\},\{2,3,4,6,17,23\}\}$

$U/E=\{\{1,2,5,7,8,10,11,13,17,19,21,24,25,26\},\{3,4,9,12,16,18,20,22,23\},\{6,14,15\}\}$

设定 $\beta=0.25$,依据加权平均变精度粗糙度定义,对于所选节点中的每一个条件属性,依次计算其对应的加权平均变精度粗糙度,结果:$\gamma_A^\beta=1$,$\gamma_B^\beta=0.7846$,$\gamma_C^\beta=0.9379$,$\gamma_D^\beta=0.6168$。

因为 D 的加权平均变精度粗糙度最小,所以首先把它作为决策树的根节点,并根据它把整个数据集划分成 4 个数据子集,得到 4 个决策子表 $S_1\sim S_4$,见表 2~表 5。

表 2 决策子表 S_1

U	A	B	C	E
1	1	2	1	L
5	2	2	1	L
8	2	3	4	L
10	2	2	2	L
11	2	2	1	L
13	3	1	2	L
18	3	1	1	R
19	4	3	3	L
20	4	2	3	R
24	2	3	4	L

表 3 决策子表 S_2

U	A	B	C	E
7	2	3	1	L
14	3	1	2	B
15	3	1	2	B
21	2	3	1	L
25	2	4	4	L
26	2	4	2	L

表 4 决策子表 S_3

U	A	B	C	E
9	2	1	1	R
12	3	1	1	R
16	3	1	2	R
22	2	1	1	R

表 5 决策子表 S_4

U	A	B	C	E
2	1	3	3	L
3	1	1	1	R
4	1	4	1	R
6	2	2	1	B
17	3	1	2	L
23	2	1	1	R

对于子表 S_1,存在决策属性的等价类 $Z_1=\{1,2,5,7,8,10,11,13,17,19,21,24,25,26\}$,使得 $\Pr(Z_1/X_1)=8/10\geqslant 1-\beta$,因此认为 X_1 属于 Z_1 所在的同一个类,把此样本集作为叶子节点,并且把这条规则的置信度设置为 80%。对于子表 S_4,不存在决策属性的等价

类使得 $\Pr(Z_j/X_2) \geqslant 1-\beta$,计算该节点的各个条件属性的加权平均变精度粗糙度,结果为: $\gamma_A^\beta = 0.75$, $\gamma_B^\beta = 0.583\,3$, $\gamma_C^\beta = 0.166\,6$。从结果中可以得出 C 的粗糙度值最小,所以选择 C 作为子表 S_4 的根节点。子表 S_3 的操作与 S_1 类似, S_2 的操作与 S_4 类似,依此类推,直到表中所有研究对象都被分类。应用该算法构造的决策树如图 1 所示。

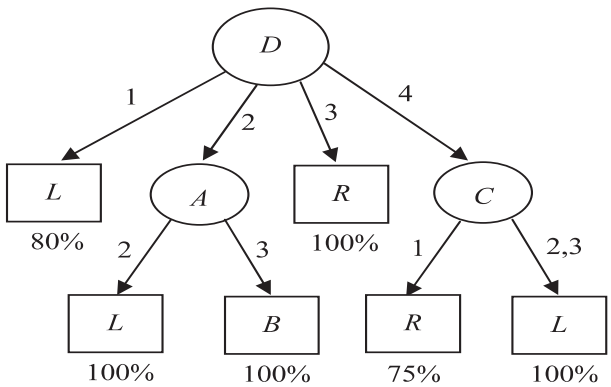

图 1　基于加权平均变精度粗糙度的决策树

传统粗糙度方法构造的决策树如图 2 所示。

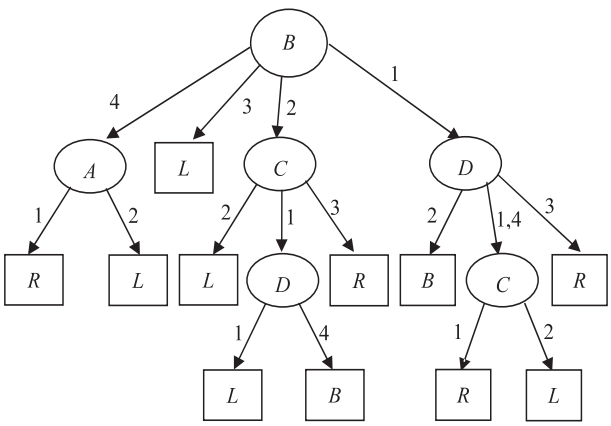

图 2　基于加权平均粗糙度的决策树

图 1 与图 2 相比较,具有如下显著特点:

(1)以该算法构造决策树可以大幅降低决策树结构的复杂性。应用该算法,得到的 9 个节点中共有 6 个终端节点,有 6 条决策规则;传统方法得到的节点中共有 11 个终端节点,得到 11 条决策规则;

(2)该算法建立的决策树能够兼容孤立数据,如属性 D 为 1 的 10 个数据中,绝大部分数据被分配到 L 类,只有 2 个数据被单独分类了。而基于加权平均粗糙度的决策树虽然对所有研究对象都进行了分类划分,但它并没有较好地考虑系统中存在干扰数据的情况,分类过程过于细致,非常容易促使数据过度拟合,导致分类规则使用面过窄;

(3)该方法构造的决策树具有确切的置信度规则,便于用户更好地理解和使用这些规则。

5　结束语

该算法基于变精度粗糙度概念提出了构造决策树的新思想。该方法计算各属性的加权平均变精度粗糙度值,选择值最小的属性作为相应节点,并构建具有精确置信度的决策规则。

通过实例验证,该算法构造的决策树具有很好的应用范围,能够兼容系统中的不相容数据,改善决策树的优良分类性能。

参考文献:

[1] 郭旭强,王大建,王秀霞,等. 影响水产食品安全因素的分析[J]. 齐鲁渔业,2009,26(12):49-51.

[2] 鄂　旭,韩　芳,侯　建,等. 面向食品安全评价的属性约简方法研究[J]. 吉林大学学报,2013,31(3):1-6.

[3] 武　森,高学东,Bastian M. 数据仓库与数据挖掘[M]. 北京:冶金工业出版社,2003.

[4] Pawlak Z. Rough set[J]. International journal of computer and information science,1982(1):341-356.

[5] 王国胤. Rough 集理论与知识获取[M]. 西安:西安交通大学出版社,2003.

[6] 张文修,吴伟志,梁吉业,等. 粗糙集理论与方法[M]. 北京:科学出版社,2001.

[7] Hu X H, Cercone N. Learning in relational databases: A rough set approach[J]. Computational intelligence,1995,11(2):323-337.

[8] Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for neural networks[J]. Computational intelligence,1995,11(2):339-347.

[9] 苗夺谦,胡桂荣. 知识约简的一种启发式算法[J]. 计算机研究与发展,1999,36(6):681-684.

[10] 韩素青,赵　岷. Reduct 理论[M]. 北京:清华大学出版社,2010.

[11] 常犁云,王国胤,吴　渝. 一种基于 Rough Set 理论的属性约简及规则提取方法[J]. 软件学报,1999,10(11):1206-1211.

[12] E Xu, Yang Yuqiang, Ren Yongchang. A new method of attribute reduction based on information quantity in an incomplete system[J]. Journal of software,2012,7(8):1881-1888.

基于粗糙变精度的食品安全决策树研究

作者： 鄂旭， 任骏原， 毕嘉娜， 沈德海， E Xu， REN Jun-yuan， BI Jia-na， SHEN De-hai

作者单位： 鄂旭,E Xu(渤海大学 信息科学与技术学院，辽宁 锦州 121001; 中国产业安全研究中心，北京 100084)， 任骏原,毕嘉娜,沈德海,REN Jun-yuan,BI Jia-na,SHEN De-hai(渤海大学 信息科学与技术学院,辽宁 锦州,121001)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(1)