

# 重权关联因子可信度联想的疾病确诊方法研究

谭文学<sup>1</sup>, 席金菊<sup>1</sup>, 荣秋生<sup>1</sup>, 王京仁<sup>2</sup>

TAN Wen-xue<sup>1</sup>, XI Jin-ju<sup>1</sup>, RONG Qiu-sheng<sup>1</sup>, WANG Jing-ren<sup>2</sup>

1.湖南文理学院 计算机科学与技术学院, 湖南 常德 415000

2.湖南文理学院 生命科学学院, 湖南 常德 415000

1.School of Computer Science & Technology of Hunan University of Arts and Science, Changde, Hunan 415000, China

2.School of Life Science of Hunan University of Arts and Science, Changde, Hunan 415000, China

E-mail: twxecgi@163.com

**TAN Wen-xue, XI Jin-ju, RONG Qiu-sheng, et al. Research on bigger-weighted-related factors' forecasting certainty factor computer disease diagnostic method. Computer Engineering and Applications, 2008, 44(33): 205-207.**

**Abstract:** Constructing and using intelligent computer systems to assist human experts to solve problems of disease diagnosis about people and livestock is one of hot spots in the field of medical artificial intelligence research. Traditional diagnosis systems taking no consideration of timing and associability between clinical characteristics are always with the high rate of misdiagnosis, and lack of practicality. At the aim to improve the efficiency of machinery-diagnosis, in case of goats, the clinical relevance of the disease is analyzed, and mechanisms of associated factor into the method of Certainty-Factor-knowledge-representation is introduced; and the CF-value of associate-factors with big weights is predicted through the channel of forecast by lowest-value, highest-value, weighted-value. Experimental results show that this method can significantly the improve inference rule activation rate and reduce the rate of misdiagnosis.

**Key words:** bigger-weighted-related; Certainty Factor forecast; weighted uncertainty reason; disease diagnosis; Certainty Factor Integration

**摘要:**设计利用智能计算机系统协助人类专家求解人畜疾病诊断问题是医学人工智能的研究热点之一。传统诊断系统没有处理临床之间的时序相关性,漏诊率高,实用性低。为改进机器诊疗效率,以山羊为例,分析了疾病临床的关联性,在可信度知识规则表示方法中引入关联因子机制;用低值联想,高值联想,加权联想等算法修正重正关联因子可信度。实验表明:重权关联因子可信度联想显著地改进了推理规则激活率,大大降低了漏诊率。

**关键词:**重权关联因子;可信度联想;加权不确定推理;疾病诊断;可信度合成

**DOI:**10.3778/j.issn.1002-8331.2008.33.062 **文章编号:**1002-8331(2008)33-0205-03 **文献标识码:**A **中图分类号:**TP183

自20世纪以来,计算机技术和人工智能理论在农业应用领域不断取得突破,疾病诊断专家系统在畜牧养殖领域得到广泛应用,它汇集许多经过规范化的专家知识,辅助人类专家进行远程虚拟会诊,方便了疾病防治,成为养殖畜牧业非常实用的辅助性生产工具。如文献[1]中的鸡病诊断专家系统。不确定推理算法在其中普遍使用,然而,疾病常表现出动态生长发展特性,临床之间常存在衍生关系,大部分系统却忽略了这层关系,或没任何处理,对已观察到的临床赋予某可信度,而将其衍生的若干临床,由于当时没有发生,而可信度直接取0值,而它们往往佐证力量很大,或带有很大权值,因此,致使综合可信度很小,达不到阈值,一些知识规则无法应用,常常得出结果或造成漏诊,大大降低了实用性与辅助性<sup>[2]</sup>。论文以常德地区黑山羊为例,围绕机器疾病诊断的衍生临床可信度联想方法进行了

研究和探索。

## 1 知识表示基础:可信度方法

可信度方法由 E.H.Shortliffe 等在 Theory of Confirmation 的基础上,结合概率论提出的一种不确定性的知识表示方法,后来发展成为带阈值的确定性方法,加权不确定性方法和前提带可信度因子不确定性方法等变体,参考[3],其中加权不确定性方法在疾病确诊系统中广泛应用。

**定义 1**(加权不确定规则) 某疾病  $D$  对应临床集  $L$  的  $j$  号确诊规则(不同专家对同一组临床会做出不雷同的诊断)形如(1),使用确定性因子  $CF_j(D, L)$  来描述强度,  $CF_j(D, L)$  越大则前提  $E$  支持结论  $H$  的程度越大,结论  $H$  的真概率增量越大;  $CF_j(D, L)=0$ , 则前提  $E$  与结论  $H$  无关即相互独立,对  $H$  的真概

**基金项目:**湖南省科技计划项目(No.2008FJ3050);湖南省教育厅基金项目(No.08C610);湖南省“十一五”重点建设学科项目(动物学);湖南文理学院教学改革研究课题重点项目(No.JGZD0608);2007年常德市科技特派员专项资助。

**作者简介:**谭文学(1973-),男,高级工程师,系统分析师;研究方向:人工智能与专家系统;席金菊(1973-),女,副教授,研究方向:人工智能与模式识别;荣秋生(1973-),男,副教授,研究方向:数据挖掘;王京仁(1963-),男,教授,研究方向:草食及名优特种动物饲养与保健。

**收稿日期:**2007-12-10 **修回日期:**2008-03-18

率没有贡献;  $CF_j(D, L) < 0$ , 越小则增加了  $\sim H$  的真概率, 其值在知识工程过程中, 专家根据经验给出<sup>[4]</sup>。  $0 \leq w_i \leq 1 (i=1, 2, \dots, n)$ ,  $w_i$  是加权因子,  $\sum w_i = 1$ ,  $\lambda$  是阈值, 值由专家给出。

if  $L_{j_1}(w_{j_1})$  and  $L_{j_2}(w_{j_2}) \dots$  and  $L_{j_m}(w_{j_m})$   
then  $D(CF_j(D, L), \lambda_j), 1 \leq j \leq m$  (1)

## 2 症状基础数据的相关性分析和疾病对象建模

临床和疾病是领域知识的基本概念和对象, 称基础数据, 专家知识表示了临床和疾病之间复杂的因果关系, 这种复杂性体现在关系的概率性、模糊性、不确定性<sup>[5]</sup>。知识表示是建立在临床疾病等基础数据的表示和组织基础之上的, 基于过程的结构化方法不能有效地对疾病等领域知识对象进行信息封装与隐藏和重用。羊病对象类采用面向对象的方法来定义, 某具体羊病对象示例如下。

```
Class<Disease > /* 羊肠毒血症 */
{public:
Name; /* 病名 */ ntype; /* 类别 */, Diagnosis; /* 症状 */: :
Prevent-Cure /* 防治方法 */;
```

“上午打一针阿莫西林, 下午或晚上打一针水阳酸钠, 同时配合中药治疗。

中药配方: 梔子、黄柏、柴胡、地骨皮各 25 克, 射干 50 克, 黄连 100 克, 加水 10 公斤, 文火

煎至 3.5 公斤, 3-5 层纱布过滤 2 次, 每日饮两次, 连饮三天。

另, 药渣子砸碎, 和黄豆面一并喂羊吃掉。...

```
Clinic /* 临床表现及(代码, 致病可信度, [权值]*/: {
01 体温升高至 41℃ [0.6]
02 精神沉郁、不爱吃草, 临死前状态不稳 [0.7]
03 呼吸困难、磨牙、口鼻流血 [0.5]
04 有时四肢抽搐痉挛 [0.75]
05 结膜及黏膜充血, 轻度发炎 [0.6]
06 腹部膨大、腹痛 [0.85]
07 结膜潮红, 苍白, 四肢、耳尖发凉 [0.9]
08 肠系膜淋巴结切面有黄白色病灶 [0.7]
}
```

在这个案例中, 确诊羊肠毒血症可以观察 8 个症状, 它们不是完全独立的, 有些症状存在一定的衍生关系, 如“腹部膨大、腹痛”症状发生后, 如果不采取措施则“结膜潮红, 苍白, 四肢、耳尖发凉”后面的症状很可能发生, 还有 8 号症状也类似, 这些都是确诊羊肠毒血症的关键特征, 说明临床之间表现出一定的相关性。

## 3 重权关联因子不确定性知识规则的 O-O 表示

规则是从领域知识提炼得到的知识精品, 是专家“知识”的表示形式, 可以从不同的角度表达这种复杂的联系, 如基于概率论的主观贝叶斯方法, CF 理论的加权有确定方法等<sup>[5]</sup>。从不同角度表达不确定性, 得到不同类型的规则。某临床  $L_i, L_j$ , 如果  $L_i$  以某可信度值被观测到, 则未来某个时间  $L_j$  也会发生, 则称  $L_j$  为  $L_i$  的衍生临床或后继临床,  $L_j$  可能有多个兄弟, 也会有衍生临床,  $\langle L_i, L_j, L_k, \dots \rangle$  称为衍生集。在这里只强调其同源性, 而不考虑时序性, 而且通常被指定不会低于源临床的权值, 它们称为称重权关联因子, 由专家指定<sup>[6]</sup>。规则直接源于专家, 间接源于专家实践环境和经历, 环境和经历对专家提炼经验规则影响很大, 要求不同专家就某目标疾病形成统一的确诊规则是

不现实的, 不同专家专业素质水准也有差别, 规则引入作者属性便于跟踪专家规则的激活和应用, 也便于系统对专家进行评判。综合分析, 采用 OO 的建模分析方法将规则建模 C++ 伪码如下。

```
Class <DeriveClinicList> /* 临床衍生集类 */
public
Virtual Class<CRule> /* 规则的虚基类 */
{public: /* */
Author; /* 作者(专家署名) */
nCode; /* 编号 */
nType; /* 类型码(主观贝叶斯规则/加权不确定规则/...) */
nObject; /* 目标疾病码 */
}
Class<WeightedUncertainty>: public CRule /* 加权不确定规则 */
{CF_HE; /* 知识强度 */
Threshold; /* 阈值 */
Weighted-Clinic-List: { /* 临床权值列表 */
< nClinicCode /* 临床码 */ ,
Weight; /* 临床权值 */ > ...}
DeriveClinicLists {临床衍生集列表
<nClinicCode /* 临床码 */ , ...>
<nClinicCode /* 临床码 */ , ...>
...}
}
```

## 4 重权关联因子的不确定性联想和不确定性传递

定义 2(重权关联因子的不确定性联想) 在有重权关联因子  $E_i, E_{i+1}$  等的规则库中, 如果只观察到  $E_1$  而当前没有观察的其衍生因子而 CF 都取 0 值, 显然会得到很低的综合证据 CF 值, 也不符现实情况, 并导致达不到阈值规则无法激活而误诊, 漏诊。针对这种情况, 要准确地运用知识则必须对重权因子的 CF 值猜测式修正, 确定  $CF(E_{i+1})$  的过程称为不确定性联想<sup>[7]</sup>。有多种方法可以对其联想。

重权因子 CF 值取 CF 向量的最低分量, 如式(2), 称低值联想法:

$$CF = \text{Min}\{CF(L_i) | CF(L_i) > 0\} \quad (2)$$

重权因子 CF 值取 CF 向量最高分量, 称高值联想法, 如式(3):

$$CF = \text{Max}\{CF(L_i) | CF(L_i) > 0\} \quad (3)$$

CF 向量中有非零分量, j 号规则的这些分量给出相应权值, 用权值加权归一后的 CF 值为 j 号规则的重权因子的 CF 预测值, 式(3), 称加权联想法。

$$CF_j = \frac{(\sum_{CF(L_j) > 0} CF(L_j) \times w_{j_i})}{(\sum_{CF(L_j) > 0} w_{j_i})} \quad (4)$$

不确定性联想确诊结论的可信度综合方法: 没有可信度联想, 重权因子规则激活量很小或没有, 联想之后, 则会出现相反的情况, 针对某组临床观测集, 常会出现关于某个目标疾病的多个结论, 区别表现在权值向量与阈值上, 一方面体现了诊断系统中的精品知识来自专家集体, 或者来各个专家成熟时期, 汇集了较完备的领域专家知识, 体现诊断民主性、全面性、准确性, 成为真正意义上的专家系统而不是“专人系统”<sup>[8]</sup>。但是, 系统必须给出一个综合性结论表达机器专家的确诊, 这个过程称结论可信度综合, 参考文献[9]中提到有限和法, 极大值法, 加权

求和法, 递推计算法等。由于现实世界的复杂多样性, 很难设计一个通用算法, 需要结合具体问题进行实验比较。

疾病  $D$  有  $m$  条规则激活得一个  $CF$  值集,  $CF_i(D)$  表示子项, 采用求并方法合成  $CF$ , 求并是非递减运算, 当结论超过 1 时, 用较小值函数归一化处理(式(5)), 称为集合并合成法;

$$CF(D) = \min \left( \sum_{i=1}^{c_1} CF_i(D) - \sum_{1 < j \leq m} CF_j(D) \times CF_j(D) + \sum_{i < k \leq m} CF_i(D) \times CF_j(D) \times CF_k(D) - \dots + (-1)^{m-1} \prod_{i=1}^m CF_i(D), 1 \right) \quad (5)$$

规则的知识强度和其阈值关系很大, 而两者都由专家给出, 人为主观因素影响很大, 将强度和阈值乘积作为系数, 再归一化处理(式(6)), 称强度阈值系数法, 这样种方法消除了主观因素影响。

$$CF(D) = \frac{\sum_{i=1}^m CF_i(D) \times CF_i(D, L) \times \lambda_i}{\sum_{i=1}^m CF_i(D, L) \times \lambda_i} \quad (6)$$

### 5 实验数据及效果分析

为验证方法有效性, 研究组收集了一组常见多发的羊肠毒血症病例数据集, 其临床及编码表 1。其中 4, 5, 6, 7 号临床有重权关联关系<sup>[9]</sup>。实验中, 由将临床观察数据输入诊断系统原型, 并将临床报告单打印 5 份分别递送给 5 个独立的羊病专家, 进行确诊, 结果用模糊语言: [完全不可能(0), 不可能(0.35), 可能(0.65), 很可能(0.8), 肯定(1.0)] 给出隶属度描述, 限于篇幅, 只给出专家对目标为羊肠毒血症的数据, 见表 2。其均值=0.68, 落在可能和很可能的区间[0.65, 0.8]之中。专家系统给出 6 条确诊规则, 编码和权值向量及用户观测的  $CF$  向量, 见表 3。重权联想推理方法的数据对比如表 4。可看出, 没有联想, 只有 2 号规则运用, 结论值 0.499 2, 实用价值很低; 低值法的结论区间: [0.527 0, 0.687 0], 高值法的结论区间: [0.642 6, 0.778 9]; 加权法的结论区间: [0.608 3, 0.761 1]; 规则激活数量大增, 结论区

表 1 羊肠毒血症的确诊临床

临床	编码
温度升高至 41℃~42℃, 消瘦	0
精神沉郁、不爱吃草, 临死前状态不稳	1
呼吸困难、磨牙、口鼻流血	2
有时四肢抽搐痉挛	3
羊快疫以腹部膨大、腹痛	4
结膜潮红、苍白, 四肢、耳尖发凉	5
脱毛	6
肠系膜淋巴结切面有黄白色病灶	7

表 2 专家给出确诊数据

专家	0	0.35	0.65	0.8	1.0
A			√		
B		√			
C					√
D				√	
E			√		

间与人类专家的判决情况较接近。用不同的  $CF$  综合方法合成  $CF$  预测推理结论(表 5); 有限和、递推法及集合并法结果偏高, 近似为 1。强度阈值系数法和加权法得较准确的结果, 其中加权联想推理结果用强阈法综合结论为 0.645 5 和专家结论 0.68 相当接近。

表 3 重权因子联想不确定推理的采集数据

规则号	1	2	3	4	5	6
临床码	$CF$	$w_i$	$w_i$	$w_i$	$w_i$	$w_i$
0	1	1	2	3	2	4
1	0.8	2	2	4	2	2
2	0.9	12	17	3	2	2
3	0.9	14	17	4	4	4
*4	0.6	6	7	8	8	12
*5		8	6	11	8	7
*6		8	5	8	8	4
*7		4	4	9	4	4
$\lambda$	0.60	0.50	0.55	0.75	0.60	0.60
$CF(D, L)$	0.90	0.78	0.80	0.85	0.75	0.90

表 4 不同推理方法的数据对比

规则号	1	2	3	4	5	6
无联想推理						
归一化	0.538 2	0.640 0	0.346 0	0.363 2	0.394 6	0.505 6
结论 $CF$	0.000 0	0.499 2	0.000 0	0.000 0	0.000 0	0.000 0
低值联想推理						
归一化	0.756 4	0.790 0	0.682 0	0.678 9	0.702 7	0.705 6
结论 $CF$	0.680 7	0.616 2	0.545 6	0.000 0	0.527 0	0.635 0
高值联想推理						
归一化	0.865 5	0.865 0	0.850 0	0.836 8	0.856 8	0.805 6
结论 $CF$	0.778 9	0.674 7	0.680 0	0.711 3	0.642 6	0.725 0
加权联想推理						
$CF$ 联想	0.845 7	0.853 3	0.786 4	0.766 7	0.811 1	0.758 3
归一化	0.845 7	0.853 3	0.786 4	0.766 7	0.811 1	0.758 3
结论 $CF$	0.761 1	0.665 6	0.629 1	0.651 7	0.608 3	0.682 5

表 5 结论  $CF$  不同综合方法对比

	有限和	加权法	最大值	递推法	集合并	强阈法
低值	1	0.501 3	0.680 7	0.990 3	0.990 3	0.476 4
高值	1	0.704 8	0.778 9	0.999 3	1.000 0	0.706 2
加权	1	0.647 0	0.761 1	0.994 6	0.994 6	0.645 5

### 6 结束语

带权不确定性知识表示及推理方法在医学诊疗系统广泛使用, 但很多疾病诊断系统忽略了临床特征的相关性, 规则激活率低, 导致漏诊。项目以山羊疾病诊断为例, 分析动物疾病诊断临床因子的关联性, 设计支持重权因子  $CF$  知识规则的表示方法; 提出证据可信度联想的思想, 提出并使用低值法、高值法、加权法对重权因子可信度因子进行联想; 针对由此而引起的确诊结论“多元化”的情况, 提出  $CF$  集合并综合算法和强度阈值系数综合方法整合机器诊断结论。实验数据表明: 这一系列方法能有效改善规则激活率和减少误诊率, 增强了诊疗系统实用性, 下一步将更大范围, 更大规模地验证其有效性, 完善联想方法, 在条件允许下, 在专家指引下, 研究作用于人类疾病诊疗的可能性和效果。