

解决证据推理中一类“0绝对化”问题的方法^{*}

Method to Solve a Kind of “0 Absolutization” Problems in Evidential Reasoning

徐从富 耿卫东 潘云鹤

(浙江大学人工智能研究所 杭州310027)

TP18

Abstract In fact, evidential reasoning is a kind of commonsensible reasoning, so the reasoning results must be consistent with the basic common senses of human beings. In this paper, regarding to a kind of “0 absolutization” problem in the evidential reasoning, two real examples are given, and the handling method is brought forward. Compared with traditional approaches, this method embodies much more man’s decision-making common senses. The “0 absolutization” problems in two examples are solved reasonably, and the results are discussed finally.

Keywords Dempster-Shafer evidential reasoning, 0 absolutization, Non-determinant reasoning

1 引言

Dempster-Shafer 证据理论^[1]应用于不确定性推理领域即为证据推理。证据推理已经成为不确定性推理理论中的一种重要方法,实质上它是一种常识推理。证据推理同其它的不确定性推理方法相比,有其独特的优点,它特别适合表达模糊的、不确定的和“不知道”的信息,这类信息保存在基本可信度分配 BPA(mass 函数)中,且在推理过程中保留了这些信息。在证据推理中,最关键的是 mass 函数的分配与合成。mass 函数是一种假设的度量函数,它是人们主观给出的,或者凭经验和感觉给出^[4]。文[3]指出领域专家很难恰当地将不确定性和主观信度表示成 mass 分配,比如在医疗诊断中,即使医术非常高明的医生也很难对某些症状或疾病给出一个确定的信度分配,因为在现实世界中,有些感觉和经验是“只可意会不可言传”的。因此,原始的基本可信度分配 BPA 不一定完全合理,人们对个别假设的 mass 分配很可能是偏颇的,而对有些假设很可能忽略了对其进行应有的 mass 分配,哪怕数值是很小的(在现实中,这意味着该事件发生的概率很小,但还是可能发生的)。有时会出现 mass 函数的“0分配”问题,即对假设空间(识别空间) Θ 中的某些假设(目标)的 BPA 赋0值,从而导致无法进行 Dempster 合成或合成结果很不符合人类的决策常识,本文将这两种情况统称为“0绝对化”问题。文中例举了两个简单的实例,并提出一种模仿人类“整理数据”的方法,即对检查

出有问题的原始 mass 分配进行适当的“修正”,从而得到合理的、符合人类常识的结论。文中给出了两个实例经该方法处理后的结果且对其进行了讨论。

2 问题的提出

下面举两个简单的实例来说明证据推理中的“0绝对化”问题。

例1 设假设空间 $\Theta = \{h_1, h_2, h_3\}$, h_1 表示“张三”, h_2 表示“李四”, h_3 表示“王五”,分别为三个犯罪嫌疑人,已知其中必有一人是罪犯,现有三个检察官各自给出基本可信度分配 BPA 如表1所示。经 Dempster 合成规则运算后,得

$$m(\{h_1\})=0, m(\{h_2\})=0.33, m(\{h_3\})=0.67$$

其它均为0。计算结果表明 h_3 , 即“王五”犯罪的可能性最大。表1中 $m(\{h_2\})=0$ 导致了“0绝对化”问题。

表1 基本可信度分配 BPA

$m \backslash h$	$\{h_1\}$	$\{h_2\}$	$\{h_3\}$	$\{h_1, h_2\}$	$\{h_1, h_3\}$	$\{h_2, h_3\}$	Θ
m_1	0.9	0.05				0.05	
m_2		0.2	0.8				
m_3	0.8	0.1	0.1				

注:表中空项的数值为0。

例2 设识别空间 $\Theta = \{h_1, h_2, h_3\}$, h_1 表示“驱逐舰”, h_2 表示“核潜艇”, h_3 表示“航空母舰”,现有八个传感器分别给出基本可信度分配 BPA 如表2所示。

* 本文的研究得到国家自然科学基金、国防科工委预研项目的资助。

表2 基本可信度分配 BPA

$m \backslash h$	$\{h_1\}$	$\{h_2\}$	$\{h_3\}$	$\{h_1, h_2\}$	$\{h_1, h_3\}$	$\{h_2, h_3\}$	Θ
m_1	0.6	0.1	0.2				0.1
m_2	0.1	0.1	0.7		0.05		0.05
m_3	0.8		0.2				
m_4	0.7	0.2	0.1				
m_5	0.6	0.2	0.2				
m_6		0.9	0.1				
m_7	0.1	0.2	0.7				
m_8	0.9	0.1					

注:表中空项的数值为0。

对表2的数据应用 Dempster 合成规则进行计算,出现归一化常数 $N=0$,说明对这批数据,DS 方法无法作出决策。事实上,只要对上述两表中的很少几个值做细微的调整,就可以得到很满意结果。这样的调整是有事实根据的,因为人们在对某个假设空间(识别框架) Θ 进行基本可信度分配时,不可避免地带有一定的主观随意性,只要对 BPA 的某些值调整的幅度很小,并不意味着改变人们对某些假设的原意,相反可以修正原始 BPA 的某些不合理性。

3 解决方法

3.1 基本思想

在常规证据理论中,已经体现出证据的“收集”和“合成”工作,本文在它们之间增加“数据处理”工作,故有必要对检查出有问题的原始 mass 分配进行适当的调整,从而得到更满意的结果。不合理的 mass 分配主要包括两类:一类是无法进行 Dempster 合成的数据;另一类是虽然能够得到某些结果,但是这样的结果同人类的常识是相背离的。设假设空间 $\Theta = \{h_1, h_2, \dots, h_m\}$ 有 m 个假设,在 Θ 上有 n 个原始 mass 分配 $m_1, m_2, \dots, m_n$,可用如表3所示的 mass 分配表来说明问题,设对应于 mass 函数 m_i 的集合为 $A_{ij}, 1 \leq i \leq n, 1 \leq j \leq 2^m - 1$,其中那些满足 $m(A_{ij}) > 0$ 的 A_{ij} 为焦点元素(focal elements)。不合理的 mass 分配主要有以下特征:

(1)从理论上说,若归一化常数 N 出现如下情形:

$$N = \sum_{A_{1k} \cap A_{2p} \cap \dots \cap A_{nq} \neq \emptyset} m_1(A_{1k}) \cdot m_2(A_{2p}) \cdot \dots \cdot m_n(A_{nq}) = 0$$

其中, $1 \leq k, p, \dots, q \leq 2^m - 1$, 则用 Dempster 合成规则将得不到任何结果。在证据推理中,认为这批证据是矛盾的,故无法作出决策。从算法实现的角度来看,随着假设空间 Θ 中的假设个数 $|\Theta|$ 以及参加合成的 mass 函数个数 n 的不断增多,要检查出得不到任何结果的情况,其计算复杂度为 $2^n \cdot |\Theta|$,故出现了指数爆炸的问

题。可以考虑利用贝叶斯近似法^[2]先将每一个 mass 函数近似变换成贝叶斯 mass 函数,然后,根据变换后的 mass 函数进行 Dempster 合成,若表3中从 $\{h_1\}$ 到 $\{h_m\}$ 的 m 列中,每一列都至少有一个 mass 函数值为0,则该批数据无法进行 Dempster 合成。

表3 mass 分配情况表

$m \backslash h$	$\{h_1\}$	$\{h_2\}$	$\dots$	$\{h_1, h_2, \dots, h_j\}$	$\dots$	Θ
m_1	$M_1(A_{11})$	$m_1(A_{12})$	$\dots$	$m_1(A_{1k})$	$\dots$	$m_{1,2^m-1}(\Theta)$
m_2	$M_2(A_{21})$	$m_2(A_{22})$	$\dots$	$m_2(A_{2k})$	$\dots$	$m_{2,2^m-1}(\Theta)$
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$
M_n	$M_n(A_{n1})$	$m_n(A_{n2})$	$\dots$	$m_n(A_{nk})$	$\dots$	$m_{n,2^m-1}(\Theta)$

(2)若归一化常数 $N > 0$,说明能进行 Dempster 合成。对表3中的前 m 列 mass 函数值分别求算术平均值,若发现算术平均值最大的那列有一个或多个 mass 函数值为0,或小于允许的最小值 MinMass (如0.01),则这批证据的推理结果很可能出现“0绝对化”的问题。

所以,对以上两种情况,都有必要进行适当的调整。在实际应用中,解决以上问题的基本思想是:将某些特定的 mass 函数值拆分成两部分:一部分与原值很接近,仍赋给原假设;另一部分则很小,赋给相应的 BPA 为0或小于 MinMass 的那些假设。

3.2 方法步骤

(1)验证原始 mass 分配的合理性

①先利用贝叶斯近似公式将每一个 mass 函数近似变换成贝叶斯 mass 函数。

②对近似计算后的 mass 函数分配表中的每一列分别求算术平均值。

③求归一化常数 N :

$$N = \sum_{A_{1k} \cap A_{2p} \cap \dots \cap A_{nq} \neq \emptyset} m_1(A_{1k}) \cdot m_2(A_{2p}) \cdot \dots \cdot m_n(A_{nq}) = 1 - \sum_{A_{1k} \cap A_{2p} \cap \dots \cap A_{nq} = \emptyset} m_1(A_{1k}) \cdot m_2(A_{2p}) \cdot \dots \cdot m_n(A_{nq})$$

若 $N=0$,或 $N>0$ 但 mass 分配表中算术平均值最大的那列有 BPA 值小于允许的最小值 MinMass ,则继续下一步;否则,直接转入(3)。

(2)适当调整某些 mass 函数的 BPA 判断 N 的值:

a)若 $N=0$,找到 mass 分配表中算术平均值最大的那列中 BPA 为0的行;

b)若 $N>0$,找到 mass 分配表中算术平均值最大的那列中 BPA 小于允许的最小值 MinMass 的行;

对这些行的 BPA 进行调整,将位于算术平均值最大的那列的 BPA 值改为 MinMass,相应地找到这行中数值最大的那个 BPA,并将其值减少 MinMass。

(3) 利用 Dempster 合成规则进行合成 n 个 mass 函数 $m_1, m_2, \dots, m_n$ 的 Dempster 合成公式为:

$$(m_1 \oplus m_2 \oplus \dots \oplus m_n)(A) = \frac{1}{N_{A_1 \cap A_2 \cap \dots \cap A_n = A}} \sum_{A_1 \cap A_2 \cap \dots \cap A_n = A} m_1(A_1) m_2(A_2) \dots m_n(A_n)$$

(4) 计算信度测度 $Bel(A)$ 和似然测度 $Pl(A)$ 因为 BPA 在(1)中已转换成贝叶斯 mass 函数,故有:

$$Bel(A) = Pl(A) = \sum_{B \subseteq A} m(B)$$

4 结果与讨论

4.1 结果

下面使用上述方法对2中的两个例子进行处理。

(1) 对例1的情况,经贝叶斯近似变换后得到的 mass 分配如表4所示。

表4 贝叶斯近似变换后表1的基本可信度分配 BPA

$m \backslash H$	$\{h_1\}$	$\{h_2\}$	$\{h_3\}$
m_1	0.8574	0.0950	0.0476
m_2	0	0.2	0.8
m_3	0.8	0.1	0.1

根据上述的解决方法,应调整 $m_2(\{h_1\})$ 和 $m_2(\{h_3\})$ 的值,现改为:

$$m_2(\{h_1\}) = 0.01 \text{ 和 } m_2(\{h_3\}) = 0.79$$

然后进行 Dempster 合成,得:

$$Bel(\{h_1\}) = Pl(\{h_1\}) = m(\{h_1\}) = 0.548$$

$$Bel(\{h_2\}) = Pl(\{h_2\}) = m(\{h_2\}) = 0.152$$

$$Bel(\{h_3\}) = Pl(\{h_3\}) = m(\{h_3\}) = 0.300$$

因此,张三的犯罪嫌疑最大。

(2) 对例2的情况,经贝叶斯近似变换后得到的 mass 分配如表5所示。

表5 贝叶斯近似变换后表2的基本可信度分配 BPA

$m \backslash h$	$\{h_1\}$	$\{h_2\}$	$\{h_3\}$
m_1	0.5833	0.1667	0.2500
m_2	0.1739	0.1304	0.6957
m_3	0.8	0	0.2
m_4	0.7	0.2	0.1
m_5	0.6	0.2	0.2
m_6	0	0.9	0.1
m_7	0.1	0.2	0.7
m_8	0.9	0.1	0

根据上述的解决方法,应调整 $m_8(\{h_1\})$ 和 $m_8(\{h_2\})$ 的值,现改为:

$$m_8(\{h_1\}) = 0.01 \text{ 和 } m_8(\{h_2\}) = 0.89$$

然后进行 Dempster 合成,得:

$$Bel(\{h_1\}) = Pl(\{h_1\}) = m(\{h_1\}) = 1$$

$$Bel(\{h_2\}) = Pl(\{h_2\}) = m(\{h_2\}) = 0$$

$$Bel(\{h_3\}) = Pl(\{h_3\}) = m(\{h_3\}) = 0$$

因此,最终推理结果为“驱逐舰”。

4.2 讨论

在4.1(1)中,将 $m_2(\{h_1\})$ 从0.001逐渐增加到0.01,而 $m_2(\{h_3\})$ 从0.799逐渐减少到0.790,信度 $Bel(\{h_1\})$ 同调整值之间的关系如表6所示。

从表6可看出,当 $m_2(\{h_1\})$ 大于0.008时, h_1 的最终信度就超过0.5。

在4.1(2)中,随着 $m_8(\{h_1\})$ 从0.001逐渐增加到0.01,但由于 $m_3(\{h_2\})$ 和 $m_8(\{h_3\})$ 始终为0,故 $Bel(\{h_1\})$ 也始终为1。相反,如果将 $m_3(\{h_2\})$ 或 $m_8(\{h_3\})$ 赋值为任何非零的正数,而 $m_8(\{h_1\}) = 0$,则 $Bel(\{h_2\})$ 或 $Bel(\{h_3\})$ 将变为1,这显然是不合理的。如果将 $m_3(\{h_2\})$ 和 $m_8(\{h_3\})$ 赋值为0.01,相应地 $m_3(\{h_1\})$ 和 $m_8(\{h_1\})$ 分别改为0.79和0.89,则 $m_8(\{h_1\})$ 从0.001逐渐增加到0.01,相应地 $m_8(\{h_2\})$ 从0.899逐渐减少到0.89, $Bel(\{h_1\})$ 的变化如表7所示。

表6 $Bel(\{h_1\})$ 随 $m_2(\{h_1\})$ 变化而变化的情况

$m_2(h_1)$	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010
$Bel(h_1)$	0.107	0.194	0.265	0.325	0.376	0.427	0.458	0.492	0.521	0.548

表7 $Bel(\{h_1\})$ 随 $m_8(\{h_1\})$ 变化而变化的情况

$m_8(h_1)$	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010
$Bel(h_1)$	0.823	0.903	0.933	0.949	0.959	0.965	0.970	0.974	0.977	0.979

由表7可见最终决策为“驱逐舰”。

第4.1节(1)中的 m_2 和(2)中的 m_8 可以认为是相反意见或矛盾证据,有些传统方法简单地将它们从这

组 BPA 中删掉。这样做当然对推理的最终结果很有利,即作为推理结论的那个假设的信度可能很接近1,但是,从证据和信度更新的角度来看,这将带来十分严

重的问题,因为随后进来的 BPA 完全有可能偏向原先个数很少的那些相反意见,但由于证据推理的“0绝对化”,这些相反意见完全被排斥,这种情况并不符合某些工程实际的要求,故不能简单地去掉矛盾证据。如果对这些矛盾证据做适当的、细微的调整,不但能纠正这些矛盾证据的 mass 分配的偏颇,而且也不同程度的吸取了不同的意见,这很符合人类的决策原则,即“求大同,存小异”。

结论 通过以上的分析,可以看出本文提出的解决证据推理中一类“0绝对化”问题的方法是可行的,由于 BPA 的给定带有不同程度的主观性,因而有必要在用 Dempster 合成规则进行证据合成之前,对某些 BPA 做适当的调整。本文为了计算的简化,一开始就将所有被合成的 BPA 都近似转化成贝叶斯 mass 函数,所以最终的信度测度和似然测度是相等的,这样处理是合理的,因为文[2]证明了当人们关心的是单个假设的最终结论时,用贝叶斯近似法能够得到满意的结论。其实如果一开始不进行贝叶斯近似,并不影响本文方法的效果,当然如果那样做的话,将保持证据推理所特有的表示和处理“不确定”和“不知道”的风范,只是

随着假设空间的维数和被合成的 mass 函数个数的增多,其计算量将是非常巨大的。此外,也可以将 mass 函数值中原来的 $m(\emptyset)=0$ 的项赋与很小的值(相应地减少另一个数值很大的 BPA),以达到调整 BPA 的目的,这似乎更符合人的“常识”,更体现 BPA 分配的合理性,但在本文的方法中,即使这样做了,在随后的贝叶斯近似转换中还会将这个 $m(\emptyset)$ 值“分配”到所有的单个假设的 BPA 中,因此不如本文的方法来得直截了当。

参考文献

- Shaler G. A Mathematical Theory of Evidence. Princeton University Press. Princeton, New Jersey, 1976
- Voorbraak F. A Computationally Efficient Approximation of Dempster-Shafer Theory. Int. J. Man-Machine Studies, 1989(30), 525~536
- Dillard R A. Using Data Quality Measures in Decision-Making Algorithms. IEEE Expert, 1992(12): 63~72
- 张文修,梁怡. 不确定性推理原理. 西安交通大学出版社, 1996. 10
- Han Jiawei, Fu Yongjian. Discovery of Multi-level Association Rules From Large Databases. In: Proc of the 21st Intl. Conf. on VLDB. Zurich, Switzerland, Sep. 1995. 420~431
- Park J S, et al. An Effective Hash-based Algorithm for Mining Association Rules. In: Proc. of the 1995 ACM SIGMOD Intl. Conf. On Management of Data. May 1995. 175~185
- O'zden B, et al. Cyclic Association Rules. In: Proc. of the 1998 Intl. Conf. On Data Engineering (ICDE'98) Orlando, FL, Feb. 1998. 412~421
- Agrawal R, et al. Fast Discovery of Association Rules. In: Advances in Knowledge Discovery and Data Mining. AAAI Press, 1995. 1~36
- Han Jiawei, et al. Mining Segment-Wise Periodic Patterns in Time-Related Databases. In: Proc. of 1998 Intl. Conf. on Knowledge Discovery and Data Mining (KDD'98). New York City, NY, Aug. 1998
- 黄益民. 经常性周期关联规则的研究. 计算机科学, 2000, 27(4): 50~53
- Agrawal R, et al. Mining Association Rules between Sets of Items in Large Databases. In: Proc. of the 1993 ACM SIGMOD Intl. Conf. on Management of Data. Washington, DC, 1993. 207~216
- Areaway R, Syrian R. Fast Algorithms for Mining Association Rules. In: Proc. of the 20th Intl. Conf. on VLDB. Santiago, Chile, Sep. 1994. 487~499

参考文献

(上接第52页)

实验结果表明,如果周期 k 的倍数 m 存在,使用方法4求周期 k 的经常性周期关联规则的时间将是直接对周期 k 使用方法2的24%~35%。在实际运用中,周期的取值范围和数据库 D 的具体特性对方法4的效率有很大影响。从文[8]可知,当周期 k 中的经常性周期关联规则较少时,方法1和方法2的运行时间是较小的。因此,在指定周期范围的情况下,只有存在经常性关联规则的周期会花费较多的运行时间。如果在指定的周期范围中,大多数周期中都只有少量经常性关联规则,方法4的优点将更为突出。