基于证据分类的加权冲突证据组合

王进花 吴 迪 曹 洁 李 军
(兰州理工大学电气工程与信息工程学院 兰州 730050)

摘 要 为了有效融合高度冲突的证据,在三角模算子和折扣因子分析的基础上,提出了一种基于证据分类的冲突证据融合规则。采用基于3角模算子定义的平均证据距离与冲突因子将证据分成可信任证据、不冲突证据和冲突证据三类,并赋予可信任证据和不冲突证据折扣因子1,极大程度上保留了证据对正确假设的支持;然后基于证据距离定义了改进的证据权重,基于加权原则对冲突证据进行合成得到修正的证据体,从而消除证据间的冲突;最后利用Dempster规则完成证据组合。算法分析表明所提方法是合理有效的。

关键词 证据理论,冲突,折扣因子,三角模算子
中图法分类号 TP391 **文献标识码** A

Weighted Combination of Conflicting Evidence Based on Evidence Classification

WANG Jin-hua WU Di CAO Jie LI Jun
(College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract In order to combine highly conflicting evidence efficiently, a new evidence combination rule making use of evidence classfication was proposed based on triangular norm and discount factor analyse. First, utilizing average evidence distance and discount factor based on triangular norm, the evidence was classified into three categories: reliability, no conflict and conflict. The discounting factors of the former two categories of evidences were set to one, which keeps the evidence hold of the right hypothesis to a great extant and makes the fusion results focus onto the right hypothesis more strongly. Then the was improved evidence weight was obtained based on evidence distance, and the modified evidence was obtained by verifying the conflicting evidence according to the weighting rule in order to eliminate the conflict. Finally, according to the Dempster's rule, the modified evidence was combined. Numerical examples show the efficiency and rationality of the proposed approach.

Keywords Evidence theory, Conflict, Discount factor, Triangular norm

1 引言

在现实生活中,无论是军事领域还是非军事领域,不确定情况下的决策问题十分常见,而不确定推理方法则为处理不确定和不精确信息提供了有效途径。D-S证据理论能够很好地表示"不确定性"和"无知"等重要概念,尤其在不确定性的表示和组合方面具有一定的优势[1],已广泛应用于决策分析、信息融合和模式识别等领域。

合成规则是D-S证据理论的核心基石之一,尽管其形式比较简单,适合机器实现,但合成的标准化过程可能会导致推理结果出现悖论。自从Zadeh[2]发现这个问题以来,冲突证据合成一直是D-S证据理论所关注的重要问题之一。国内外学者对此做了大量的研究工作,先后提出多种证据合成方法,主要包括3个方面:(1)针对传统D-S证据理论模型框架进行修改,如Smarandache F[3]等人提出的DSmT理论;(2)Yager等[4]、孙全等[5]、张山鹰等[6]、邓勇等人[7]针对组合规则本身进行修正(重新分配冲突证据来解决问题);(3)韩德强等[8]、胡丽芳等[9]、Deng Yong[10]、Murpny[11]等人针对待组合证据进行修正与预处理,即基于模型的改进方法。Haenni在其文章[12]中所述,对数据模型的修改无论是在工程上、数学上、哲学上来说都更为合理。

本文拟采用修正证据的方法来应对冲突证据的组合问题。基于修正证据的方法中,证据折扣是一种比较新颖的方法,该类方法认为冲突证据不可信,故对其进行一定的折扣,可有效削弱冲突证据对融合结果的影响。对此,本文提出一种新的冲突证据融合方法,首先采用基于三角模算子定义平均证据距离与冲突因子,将证据分成可信任证据、不冲突证据和冲突证据3类,并赋予可信任证据和不冲突证据折扣因子1,对于冲突证据进行证据折扣,极大程度上保留了证据对正确假设的支持。然后基于证据距离定义了改进的证据权重,基于加权原则对冲突证据进行合成得到修正的证据体,从而消除证据间的冲突。最后利用Dempster规则完成证据组合。

2 证据理论及冲突证据组合问题

2.1 证据理论概述

在证据理论中,若辨识框架θ中的元素满足互不相容的

到稿日期:2012-04-25 返修日期:2012-08-23 本文受国家自然科学基金项目(61263031),甘肃省自然基金项目(1010RJZA046),甘肃省教育厅研究生导师基金项目(0914ZTB003)资助。

王进花(1978—),女,硕士,讲师,主要研究方向为检测技术与自动化装置,E-mail:770274257@qq.com。

条件，命题 A 对基本概率赋值函数 m 赋值 $m(A)$ 是集合 2^θ 到 $[0,1]$ 的映射，$m:2^\theta\to[0,1]$ 必须满足下列条件[12]：

$$\begin{cases} m(\phi)=0 \\ \sum\limits_{A\subseteq 2^\theta} m(A)=1 \end{cases} \tag{1}$$

式中，$m(A)$ 称为事件 A 的基本概率赋值(Basic Probability Assignment，BPA)，也称 *mass* 函数。

设 BEL_1 和 BEL_2 是同一识别框架 U 上的两个信任函数，m_1 和 m_2 分别是其对应的基本概率赋值，焦元分别为 $A_1,\cdots,A_k$ 和 $B_1,\cdots,B_r$，则组合后的基本概率赋值：$m=m_1\oplus m_2$ 由下式计算获得：

$$m(c)=\begin{cases} \dfrac{\sum\limits_{\substack{i,j \\ A_i\cap B_j=c}} m_1(A_i)m_2(B_j)}{1-k}, & c\neq\phi \\ 0, & c=\phi \end{cases} \tag{2}$$

式中，k 是冲突因子，反映证据之间的冲突程度：

$$k=\sum_{\substack{i,j \\ A_i\cap B_j=c}} m_1(A_i)m_2(B_j) \tag{3}$$

2.2 高冲突证据组合问题

在 Dempster 组合规则中，k 是一个用于衡量融合的各个证据之间冲突的系数，如果 k 的值为 1，就不能用 Dempster 组合规则对证据进行融合；而当 $k\to1$ 时，对高冲突证据进行融合可能会导致与直觉相悖的结果。Zadeh 曾经提出一个很经典的例子：

例 1(Zadeh 反例)[13] 两个医生针对同一病人进行诊断，认为病症可能是脑膜炎(M)、脑震荡(C)及脑肿瘤(T)中的一种，辨识框架可设为 $\Theta=\{M,C,T\}$，两位医生的诊断结果为：

医生 1：$m_1(M)=0.99, m_2(T)=0.01$

医生 2：$m_1(C)=0.99, m_2(T)=0.01$

依据 Dempster 组合规则有如下结果：

$m(M)=0, m(C)=0, m(T)=1$

根据融合结果可以得出该患者所患疾病为脑肿瘤。显然和常理判断是相违的，因为两个医生都判断患者患脑肿瘤的概率极低。两个医生对两种病症的判断几乎是完全冲突的。本例就是证据体冲突的经典极端例子，可以很好地反映出在对高冲突证据体之间利用 Dempster 规则进行组合时可能出现的问题。

为了有效地管理证据冲突，有必要分析冲突的来源。引起冲突的原因是多方面的，主要有：(1)识别框架不完备。对于这种情况，已经超出现有改进的证据合成方法的范畴。冲突的再分配已经不是关键问题，要考虑拒绝某些待考虑的元素或者接受新元素。(2)证据源的检测和识别能力有限。体现在证据源各自给出的 BPA 值上。(3)证据源不可靠。合成规则假定所有参与合成的证据具有相同的重要程度，在证据组合时没有考虑证据的可信度信息。事实上，不同的证据源如传感器、领域专家具有不同的可信度。

3 改进规则中的基本概念

3.1 证据折扣分析

假设证据 e 的基本概率赋值函数为 m，折扣因子为 α，则折扣后的证据为：

$$\begin{cases} m_\alpha(A)=\alpha m(A), & \forall A\subset\Theta \\ m_\alpha(\Theta)=1-\sum\limits_{A\subset\Theta} m_\alpha(A) \end{cases} \tag{4}$$

3.2 三角模算子

三角模算子是在模糊推理思想的指导下引入人工智能的理论方法[15]，是将单源决策映射到另一空间，从而进行比较来完成判决。定义三角模如下：

若映射 $T(S_1,S_2):[0,1]*[0,1]\to[0,1]$，对 $\forall a,b,c,d\in[0,1]$，满足下列条件：

(1) $T(0,0)=0, T(1,1)=1$；

(2)若 $a\leqslant b, c\leqslant d$ 则 $T(a,b)\leqslant T(c,d)$；

(3) $T(a,b)=T(b,a)$；

(4) $T(T(a,b),c)=T(a,T(b,c))$。

则称映射 T 为三角模或三角模算子，常见的三角模算子有以下几种：

(1) Einstein product: $\dfrac{S_1S_2}{2-(S_1+S_2-S_1S_2)}$ (5)

(2) Hamacher: $\dfrac{S_1S_2}{S_1+S_2-S_1S_2}$ (6)

(3) Yager: $\max(1-((1-S_1)^p+(1-S_2)^p)^{1/p},0)$ (7)

3.3 证据分类

一般来说，证据集中的证据可以分为可信任证据、不冲突证据和冲突证据 3 类，其定义分别为：

(1)可信任证据：在证据集中被其它证据支持程度较高的证据，其证据之间分歧小；

(2)不冲突证据：与证据集中可信任证据分歧较大，但合取冲突因子较小的证据，其与可信任证据是相互支持的；

(3)冲突证据：与证据集中可信任证据分歧较大且合取冲突因子较大的证据。

因此，利用证据折扣方法融合证据时，需要尽可能地辨识出可信任证据和不冲突证据，并赋予这两类证据折扣因子 1，对于冲突证据，赋予相应的折扣因子，减小其对融合结果的影响。

4 改进规则的实现

正如前所述，Dempster 组合规则的反直观结果往往与待组合证据的冲突有关，那么采用修正证据解决证据组合反直观结果问题是基于以下出发点：反直观结果不是证据组合规则引起的，而是证据本身存在相关问题。

对于证据集中同一焦元，本文采用 triangular norm 度量证据之间的相似性：

$$S_{ij}=\frac{m_i\cdot m_j}{2-(m_i+m_j-m_i\cdot m_j)} \tag{8}$$

则对于一个包含 k 个焦元和 n 个证据体的识别框架，每两个证据体之间的证据距离为：

$$d(m_i,m_j)=\frac{1}{k}\sum S_{ij} \tag{9}$$

对于证据集 $E=\{e_i,i=1,\cdots,n\}$，相对应的 BPA 为 $M=\{m_i,i=1,\cdots,n\}$。给出证据和证据集的平均证据距离和平均合取冲突的定义。

定义 1 证据 e_i 与证据集 E 中所有其它证据的平均证据距离为：

$$\overline{d(i)}=\frac{1}{n-1}\sum_{j=1,i\neq j}^{n} d(m_i,m_j) \tag{10}$$

定义 2 证据集 E 中所有证据的平均证据距离为：

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n}\overline{d(i)} \tag{11}$$

定义 3 证据 e_i 与证据集 E 中所有其它证据的平均合取冲突为：

$$\overline{k(i)}=\frac{1}{n-1}\sum_{j=1,i\neq j}^{n}k(m_i,m_j) \tag{12}$$

定义 4 证据集 E 中所有证据的平均合取冲突为：

$$\bar{k}=\frac{1}{n}\sum_{i=1}^{n}\overline{k(i)} \tag{13}$$

则对于证据集 E，证据辨识的步骤如下：

可信任证据辨识。如果$\overline{d(i)}\leqslant\bar{d}$，则将证据 e_i 划分到可信任证据集 E_1。

不冲突证据辨识。计算证据集 $E_L=E-E_1$ 中的证据 e_j^L 与证据集 E_1 中所有证据的平均合取冲突$\overline{k_j^L}$和证据集 E_1 的平均合取冲突$\overline{k^L}$，如果$\overline{k_j^L}\leqslant\overline{k^L}$，则将证据 e_j^L 移入到不冲突证据集 E_2。

冲突证据。剩余的证据体即为冲突证据，赋予折扣因子 $\alpha_i=d_{\min}/\overline{d(i)}$。

对证据进行分类并进行折扣分析后，对证据体进行加权组合，权重的具体求解方法如下：

设有 n 个待组合的证据体 m_i，各自所对应的平均证据距离为$\overline{d(i)}$，每个证据体对应的权重如式(14)[8]（α 为负指数函数参数）：

$$w(m_i)=\frac{\exp(-\alpha\cdot\overline{d(i)})}{\sum_j\exp(-\alpha\cdot\overline{d(i)})} \tag{14}$$

则基于加权组合修正证据的公式如式(15)[8]：

$$m_{WAE}=\sum_{i=1}^{n}(w_i\cdot m_i) \tag{15}$$

加权修正时，各证据的 *mass* 值分别乘以各自证据对应的权重，再按照焦元对应关系相加得到修正后的证据 m_{WAE}，最后将 m_{WAE} 组合 $n-1$ 次就得到了最后的证据组合结果。如果遇到两个证据焦元不一致的情况，将各自缺失焦元相应的 *mass* 值赋予 0 即可。

本文所提出的改进规则的实现流程如下：

(1)根据所给证据利用 triangular norm 求解度量证据体之间的相似性，从而求出证据体之间的证据距离；

(2)根据定义 1－定义 4 求解平均证据距离和平均合取冲突；

(3)辨识出可信任证据和不冲突证据，并对冲突证据进行折扣分析；

(4)根据$\overline{d(i)}$求解每一个证据体的权重 w_i；

(5)利用加权融合规则求解待组合的修正证据体 m_{WAE}；

(6)利用 Dempster 组合规则融合修正证据体 m_{WAE}，得到最后的融合结果。

5 实验算例

本节分别对 Zadeh 反例和算例 2 所提供的数据进行合成仿真，从而证明本文所提出方法的有效性。

5.1 Zadeh 反例

Zadeh 反例详见本文例 1，通过本文所提出的方法，可以求得相关数据如下：

$$S(M)=0,S(C)=0,S(T)=\frac{1}{19802}$$

$$\overline{d(1)}=\overline{d(2)}=\frac{1}{59406}\bar{d}=\frac{1}{59406}$$

则可判断出两个证据均为可信任证据，则 $w_1=w_2=0.5$，证据组合结果如下：

$m(M)=0.495,m(C)=0.495,m(T)=0.01$

本例中，该结论相对合理。

5.2 算例 2

各证据的 *BPA* 值如表 1 所列，应用本文及其他合成方法所得的合成结果如表 2 所列。

表 1 算例 2 证据源数据

证据	a	b	c
m_1	0.90	0	0.1
m_2	0.88	0.01	0.11
m_3	0.50	0.20	0.30
m_4	0.98	0.01	0.01

表 2 组合结果

规则	组合结果	m_1m_2	$m_1m_2m_3$	$m_1m_2m_3m_4$
经典规则	m(a)	0.9863014	0.9917355	0.999915
	m(b)	0	0	0
	m(c)	0.0136986	0.0082645	0.0000850
	m(Θ)	0	0	0
Yager[4]	m(a)	0.792	0.396	0.38808
	m(b)	0	0	0
	m(c)	0.011	0.0033	0.000033
	m(Θ)	0.197	0.6007	0.611887
孙全[5]	m(a)	0.8930454	0.6591065	0.6754816
	m(b)	0.0005677	0.0242335	0.0193952
	m(c)	0.0229211	0.0621528	0.0458762
	m(Θ)	0.0834658	0.2545073	0.2592470
邓勇[7]	m(a)	0.8632126	0.6195045	0.6544677
	m(b)	0.0004001	0.0205859	0.0179771
	m(c)	0.0194015	0.0532944	0.0425243
	m(Θ)	0.1169858	0.3043183	0.2850309
本文	m(a)	0.9805	0.9976	0.9997
	m(b)	0.0023	1.1243e−004	5.4132e−006
	m(c)	0.0172	0.0023	3.0528e−004
	m(Θ)	0	0	0

从表 2 的融合结果可以看出，经典 D-S 方法和 Yager 组合规则都存在一票否决的情况，而且 Yager 组合规则的结果中，未知项的概率占主导地位，系统无法做出判断，故融合的意义不大。邓勇和孙全的融合结果差不多，虽然未知项 $m(\Theta)$的值有递减的趋势，但分配到实际目标的精度较低，表明他们正确组合效率较低。而本文所提出的方法不仅收敛速度较快，而且能够正确决策判断。

结束语 针对一票否决和鲁棒性差等冲突证据的合成问题。本文从修正证据体入手，提出了一种基于证据分类的方法用于冲突证据组合。实验结果表明，该方法在证据一致和高度冲突的情况下均表现出良好的适应性，具有较快的收敛速度和高可靠性，明显优于同类其它合成方法。

本文的工作是将证据分为可信任证据、不冲突证据和冲突证据，结合折扣因子分析，将前两者赋予折扣因子 1，最大程度上保留了证据对正确假设的支持，从而使融合结果向正确假设聚焦。对于冲突证据，依据其冲突程度赋予证据折扣是合理的。

本文利用 triangular norm 求解证据体相似性时只利用了一种最普通的 triangular norm，其有多种形式，利用其它形式的 triangular norm 能否提高融合的准确率也是下一步的重点研究方向。

参 考 文 献

[1] Shafer G. A mathematical theory of evidence[M]. Princeton: Princeton University Press, 1976

[2] Zadeh L A. Rview of Shafer's a mathematical theory of evidence [J]. AI Magaine, 1984, 5(3): 81-83

[3] Smarandache F, Dezert J. Applications and Advances of DSmT for Information Fusion [M]. Rehoboth: American Research Press, 2009

[4] Yager R R. Comparing approximate reasoning and probabilistic reasoning using the Dempster-shafer framework [J]. International Journal of Approximate Reasoning, 2009, 50(5): 812-821

[5] 孙全,叶秀清,顾伟康. 一种新的基于证据理论的合成公式[J]. 电子学报,2000,28(08):117-119

[6] 张山鹰,潘泉,张洪才. 证据推理冲突问题研究[J]. 航空学报,2001,22(4):369-372

[7] 邓勇,施文康. 一种改进的证据推理组合规则[J]. 上海交通大学学报,2003,37(8):1275-1278

[8] 韩德强,韩崇昭,邓勇,等. 基于证据方差的加权证据融合[J]. 电子学报,2011,34(3):153-157

[9] 胡丽芳,关欣,邓勇,等. 广义幂集空间中证据冲突的原因分析[J]. 控制理论与应用,2011,28(12):1717-1722

[10] Deng Yong, Shi Wen-kang, Zhu Zhen-fu, et al. Combining belief functions based on distance of evidence [J]. Decision Support Systems, 2004, 38(3): 489-493

[11] Murphy C K. Combining belief functions when evidence conflicts [J]. Decision Support Systems, 2000, 29(1): 1-9

[12] Haenni R. Are alternatives to Dempster's rule of combination real alternative? Comments on about the belief function combination and the conflict management problem[J]. Information Fusion, 2002, 3(4): 237-239

[13] 张捍东,王翠华,强克坤. 基于焦元支持度的合成规则[J]. 控制理论与应用,2011,28(5):741-744

[14] 邓勇,蒋雯,韩德强. 广义证据理论的基本框架子[J]. 西安交通大学学报,2010,44(12):119-124

[15] Jousselme A L, Liu Chun-sheng, Grenier D, et al. Measuring ambiguity in the evidence theory[J]. IEEE Transactions on Systems, Man and Cybernetics, Part A, 2006, 36(5): 890-903

[16] Klement E P, Mesiar R, Endre P, et al. Triangular norms, Position Paper I: Basic analytical and algebraic properties[J]. Fuzzy Sets Systems, 2004, 143: 5-26

(上接第 232 页)

目标领域是箱包时,本文的方法比自学习方法平均提高 2.34 个百分点,比随机子空间协同学习方法提高 1.12 个百分点,比基准系统高 4.4 个百分点。总体来看,在 8 组实验中,我们的方法表现出了较好的性能,同时验证了从一个新的角度-评价对象类别来进行跨领域情感分析是有效的。

为了进一步体现本文方法的优势,还将我们的方法同跨领域情感分类中比较流行的 SCL[3] 方法进行了比较研究。值得一提的是,SCL 在很大程度上依赖于枢轴的选择,合适的枢轴才能取得最佳效果。因此,通过不断地调整参数,使得 SCL 方法取得最佳性能。表 5 给出了本文提出的方法与 SCL 方法的性能对比结果。

表 5 各种跨领域情感分类方法比较

源领域->目标领域	COTC	SCL	Baseline
酒店->笔记本	0.795	0.732	0.722
酒店->家具	0.712	0.680	0.662
酒店->箱包	0.792	0.739	0.725
酒店->数码相机	0.749	0.732	0.697
笔记本->酒店	0.785	0.765	0.742
笔记本->家具	0.766	0.740	0.727
笔记本->箱包	0.819	0.797	0.782
笔记本->数码相机	0.740	0.740	0.725

根据表 5 可以看出,在源领域是酒店,目标领域是其他领域时,基于评价对象类别的方法平均比 SCL 方法高出 4.12 个百分点,在源领域是笔记本,目标领域是其他领域时,比 SCL 方法高出 1.68 个百分点。此实验结果再次验证了基于评价对象类别进行跨领域研究的有效性。

结束语 本文主要研究了从评价对象类别的角度进行跨领域情感分类。首先人工标注了各种评价对象类别,包括整体、硬件、软件和服务类别。然后,在源领域利用以上 4 类评价对象构建分类器,通过将不同评价对象类别当作不同的视图加入目标领域的样本来实现跨领域情感分类。实验结果表明,提出的基于评价对象类别的方法对跨领域情感分类的性能有显著提高。

参 考 文 献

[1] Aue A, Gamon M. Customizing Sentiment Classifiers to New Domains: A Case Study[C]// Proceeding of RANLP. 2005

[2] 吴琼,谭松波. 跨领域倾向性分析相关技术研究[J]. 中文信息学报,2010,24(1):77-83

[3] Blitzer J, Dredze M, Pereira F. Biographies, Bollywood, Boomboxes and Blenders: Domain Adaptation for Sentiment Classification[C] // Proceedings of ACL. 2007: 432-439

[4] Pan S J, Ni X C, Sun J T, et al. Cross-domain Sentiment Classification via Spectral Feature Alignment [C] // Proceedings of WWW. 2010: 751-760

[5] Li S, Huang C, Zong C. Multi-domain Sentiment Classification with Classifier Combination [J]. Journal of Computer Science and Technology (JCST), 2011, 26(1): 25-33

[6] 候锋,王传廷,李国辉. 网络意见挖掘、摘要与检索研究综述[J]. 计算机科学,2009,36(7):15-19

[7] Hu M, Liu B. Mining and Summarizing Customer Reviews[C] // Proceedings of KDD. 2004: 168-177

[8] Zhuang L, Jing F, Zhu X. Movie Review Mining and Summarization[C] // Proceedings of CIKM. 2006: 43-50

[9] Jakob N, Gurevych I. Extracting Opinion Targets in a Single and Cross-Domain Setting with Conditional Random Fields [C] // Proceedings of EMNLP. 2010: 1035-1045

[10] 宗成庆. 统计自然语言处理[M]. 北京:清华大学出版社,2008: 341

[11] Jacob C. A Coefficient of Agreement for Nominal Scales [J]. Educational and Psychological Measurement, 1960, 20(1): 37-46

[12] Ganchev K, Graça J, Blitzer J, et al. Multi-View Learning over Structured and Non-Identical Outputs [C] // Proceedings of UAI. 2008