基于区间数的基本概率指派生成方法及应用

康兵义[1],李 娅[1],邓 勇[1,2],章雅娟[1],邓鑫洋[1]
(1. 西南大学计算机与信息科学学院,重庆 400715;2. 上海交通大学电子信息与电气工程学院,上海 200240)

摘 要: 应用证据理论的一个关键问题是生成基本概率指派(BPA),目前如何生成BPA仍然是一个有待解决的问题.本文提出一种基于区间数的BPA生成方法,首先建立样本属性的区间数模型,然后用区间数的距离表示样本属性之间的差异性,在此基础上提出了一种相似度,最后对相似度进行归一化得到BPA.通过鸢尾花数据集(Iris Data Set)的分类实验验证了该方法的有效性,结论表明整体识别率为96%.本文方法简单实用,数据样本较少情况下也能有效确定BPA.

关键词: 证据理论; BPA; 区间数; 相似度; 数据融合; 分类
中图分类号: TP182 **文献标识码:** A **文章编号:** 0372-2112 (2012)06-1092-05
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.06.004

Determination of Basic Probability Assignment Based on Interval Numbers and Its Application

KANG Bing-yi[1], LI Ya[1], DENG Yong[1,2], ZHANG Ya-juan[1], DENG Xin-yang[1]
(1. *School of Computer and Information Science, Southwest University, Chongqing* 400715, *China*;
2. *School of Electronics and Information Technology, Shanghai Jiao Tong University, Shanghai* 200240, *China*)

Abstract: One of the open issues of Dempster Shafer theory is how to determine basic probability assignment function (BPA). To solve this problem, a method to determine BPA based on interval numbers is proposed in the paper. At first, the model of interval numbers is constructed with the samples. Then the distance of interval numbers is used to represent difference among the attributes of the samples, so the similarity of them is calculated. At last, the similarity is normalized to get the value of BPA. The effectiveness of this method is proved by classifying the Iris Set. It concludes that the total recognition rate is 96%. This method is simple and practical; it can determine BPA in the case of the little number of the samples.

Key words: evidence theory; BPA; interval number; similarity; data fusion; recognition

1 引言

在用DS证据理论进行数据融合的应用系统中,为了使用证据理论的组合规则,第一步就是要求出基本概率指派BPA[1~4],如何生成BPA,一直以来就是研究的焦点,而且还是一个开放的话题,至今没有一致的结论.

BPA生成总体来看可以分成两大类模式:一类是专家根据主观经验加以设定;一类是系统根据一些已知条件自动生成BPA.一般而言,人们无法保证专家主观观点不发生冲突,因此通过多个专家独立设置BPA的方法常常会出现高度冲突的情况[5~8].本文所研究的BPA智能化自动生成方法是特指系统具有一定样本数据的前提下,根据传感器的报告自动生成BPA函数.国内对该方面的研究偏重于应用,以解决实际问题为导向.我们注意到,韩崇昭教授在文献[9]中提出了一种在最大熵原则下确定BPA的方法,这就说明国内信息融合领域领先的团队开始注意到BPA生成的重要性.文献[10]中,基于随机集理论将模糊传感器报告生成BPA并提出一种基于证据距离的融合方法.同时邓勇等提出了基于回转半径而得到相似度,进而得出基本概率指派的方法[11].

分析现有的工作可以看出:BPA生成一般都是建立在相对完备的信息基础之上,但是在一些特殊的应用场合,比如军用目标识别系统中,由于探测手段有限或保密等原因,对敌方目标的观测是有限且是不确定的,所以建立目标属性的描述模型可用的样本数目较少.区间数只要求给定下限和上限两个数据,比较适合描述信息缺乏,不确定度高的应用场合.使用区间数对目标属性进行建模具有简单易行的特点,在其他应用场合也有广泛应用[12].

收稿日期:2011-07-15;修回日期:2012-01-17
基金项目:国家自然科学基金(No.60874105,No.61174022);教育部新世纪优秀人才支持计划(No.NCET-08-0345);重庆市自然科学基金(No.CSCT,2010BA2003)

在目标属性的区间数表示基础上,本文提出了一种新的 BPA 生成方法.这种方法应用区间数的距离来衡量区间数之间的相似度,进而得到命题的基本概率指派,最后本文通过鸢尾花数据集识别实验验证了此方法的有效性,该方法简单、易行、适用于工程.

2 基础理论

2.1 D-S 证据理论

D-S 证据理论是一种广泛被采用处理互补信息和不确定信息的数据融合理论[2].下面介绍一些证据理论的相关概念.

定义 1　辨识框架　(Frame of Discernment).设 Θ 为一个有穷而完备的论域集合,且 Θ 中的各元素相互独立,如果所关心的任一命题对应于 Θ 的一个子集,则称 Θ 为样本空间或辨识框架.

定义 2　基本置信指派函数　(Basic Belief Assignment).设 Θ 为辨识框架,A 为 Θ 的子集,如有集合函数 $m: P(\Theta)\to[0,1]$ 满足下列条件:

$$\text{(i)}\sum_{A\subseteq 2^{\Theta}} m(A)\equiv \sum_{A\subseteq\Theta} m(A)=1;\ \text{(ii)}\ m(\phi)=0 \quad (1)$$

则称 m 为辨识框架 Θ 上的基本置信指派函数(Basic Belief Assignment),也称为基本概率指派函数(Basic Probability Assignment)或 mass 函数.任意 $A\subseteq\Theta$,$m(A)$称 A 的基本置信指派,表示证据支持命题 A 本身发生的程度,而不支持任何 A 的真子集.对于 A 的不知道信息可用 $\bar{A}$ 的基本概率分配来度量,$\bar{A}=\Theta-A$. $m(A)+m(\bar{A})\leqslant 1$,说明 $m(A)$不是概率.条件(i)反应了总的置信度为 1,条件(ii)表明对于空集(空命题)不产生任何信度.

定义 3　焦元　(Focal Element).对于辨识框架的任一子集 A,如果 $m(A)>0$,则称 A 为焦元,一个 mass 函数的所有焦元的集合则成为 mass 函数的核.

定义 4　Dempster 组合规则假设辨识框架上,性质不同的两个证据,其焦元分别为 B_i 和 $C_j(i=1,2,\cdots,n;\ j=1,2,\cdots,m)$,其 mass 函数分别为 m_1 和 m_2,则 Dempster 组合规则:

$$\begin{cases} m(A)=\dfrac{1}{1-K}\displaystyle\sum_{B_i\cap C_j=A} m_1(B_i)m_2(C_j),\ (A\neq\phi,A\subset\Theta)\\ m(\phi)=0 \end{cases} \quad (2)$$

其中,矛盾因子:

$$K=\sum_{B_i\cap C_j=\phi} m_1(B_i)m_2(C_j) \quad (3)$$

2.2 区间数

区间数能够表示不确定信息或不完整信息,自提出以来就受到广泛的研究.下面给出区间数的定义及区间数距离的概念.

定义 4 若 $\tilde{a}=[a^-,a^+]=\{x\mid a^-\leqslant x\leqslant a^+\}$,$a^-$,$a^+\in R$,则称 $\tilde{a}$ 为一个区间数(Interval Number).特别地,若 $a^-=a^+$,则 $\tilde{a}$ 退化为一个实数,如果没有特别说明,本文所讨论的区间数都是正区间数($a^-,a^+>0$).

定义 5 区间数距离.设 $A=[a_1,a_2]$和 $B=[b_1,b_2]$是两个区间数,则区间数 A 和 B 之间距离的二次方 $D^2(A,B)$为[13]:

$$\begin{aligned} D^2(A,B)&=\int_{-1/2}^{1/2}\{[\frac{(a_1+a_2)}{2}+x(a_2-a_1)]\\ &\quad-[\frac{(b_1+b_2)}{2}+x(b_2-b_1)]\}^2\mathrm{d}x\\ &=[\frac{(a_1+a_2)}{2}-\frac{(b_1+b_2)}{2}]^2\\ &\quad+\frac{[(a_2-a_1)+(b_2-b_1)]^2}{12} \end{aligned} \quad (4)$$

3 新的 BPA 生成方法

本节将提出基于区间数的 BPA 生成方法.在 BPA 生成的过程中,涉及待识别样本和模型样本之间的相似度,由于它们都可以表示为区间数,因此本节首先提出区间数的相似度,之后基于所提出的相似度给出区间数 BPA 生成算法.

3.1 区间数相似度

定义 6　设 $A=[a_1,a_2]$和 $B=[b_1,b_2]$是两个区间数,则它们的相似度 $S(A,B)$定义为:

$$S(A,B)=\frac{1}{1+\alpha D(A,B)} \quad (5)$$

其中 $\alpha>0$ 是支持系数,$D(A,B)$为区间数 A 和区间数 B 之间的距离.

从相似度定义可以看出,当区间数 A 和区间数 B 相等时,$S(A,B)=1$.当区间数 A 和 B 差异越大,由于区间数的距离与相似度成反比关系,则相似度越小.同时也容易证明 $S(A,B)=S(B,A)$.

支持系数 α 的作用主要是调节生成相似度数值的离散程度,这一点对应用证据理论融合数据是有影响的.比如对于一组基本概率指派,它们的值相对集中,很容易因为精度的原因造成误差,如果在处理之前,适当调整支持系数 α 以增加数据的离散度,可以避免上述情况,进而提高识别率.

3.2 BPA 生成步骤

用区间数生成 BPA 的主要思想是:首先用收集的样本构造模型区间数,然后求待测样本与模型区间数的距离,在此基础上对区间数的距离取倒数生成相似度,最后对相似度归一化生成 BPA.此过程可以描述成以下步骤:

(1)用收集样本的特征属性的最小最大值构造区

间数模型;

(2)计算待识别样本属性值与区间数之间的距离;

(3)计算待识别样本属性值与模型区间数之间的相似度;

(4)对相似度进行归一化,生成 BPA.

下面用鸢尾花数据集(Iris Data Set)分类问题来介绍新提出的 BPA 生成方法的细节.

在这里,首先对鸢尾花数据集做一个简单的介绍,在鸢尾花数据集中,有三个种类,它们分别是 Setosa, Versicolor 和 Virginica[14],在这个数据集中有 150 个样本,每个种类有 50 个.其中的每一类花都有四种属性特征,它们分别是花萼长度 SL(Sepal Length),花萼宽度 SW (Sepal Width),花瓣长度 PL (Petal Length),花瓣宽度 PW (Petal Width).

第一步,生成区间数模型.对于这三个种类的鸢尾花,都随机选择 40 个样本,建立一个它们的区间数模型.例如对于 Setosa,最小的花萼长度 min(SL) = 4.4,最大的花萼长度 max(SL) = 0.58,那么,可以得到,对于种类 Setosa 的鸢尾花,它的花萼长度的区间数模型为[4.4,5.8],然后对依次考虑其他两种类型,可以得到如表 2 所示的区间数模型.图 1 直观的表示了它们之间的关系.

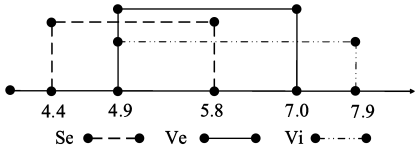

图1 每个种类花萼长度属性的区间数表示 (SL)

为了体现各个属性区间之间的关系,需要把它们之间相交的部分考虑进去.如图 1 所示,各类别的区间数表示模型有一些相交的区域,例如 Setosa 和 Versicolour 相交的区域可以用图 2 阴影部分表示.于是,能够得到 Setosa 与 Versicolor 区间数重合的部分是[4.9,5.8],所有类别相交的区域可以用表 3 表示.

表 2 花萼长度属性的区间数模型(SL)

Hypothesis	Attribute (SL)
Setosa (Se)	[4.4,5.8]
Versicolor (Ve)	[4.9,7.0]
Virginica (Vi)	[4.9,7.9]

表 3 相交区域的区间数(SL)

Hypothesis	Attribute (SL)
{Se,Ve}	[4.9,5.8]
{Se,Vi}	[4.9,5.8]
{Ve,Vi}	[4.9,7.0]
{Se,Ve,Vi}	[4.9,5.8]

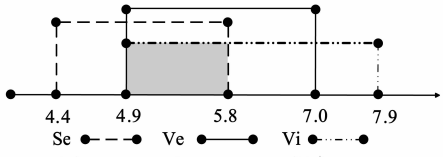

图2 Setosa和Vesicolour相交(SL)

从鸢尾花数据集(Iris Data Set)中随机抽取一个样本,例如,新的样本是(4.5cm,2.3cm,1.3cm,0.3cm),这个样本属于 Setosa 类鸢尾花.

第二步,求待测样本和模型属性之间的距离.把新样本的属性值看成区间数,比如把抽取样本花萼长度(SL)4.5 看成区间[4.5,4.5],于是抽取的样本与模型区间数的关系,可以如图 3 表示.根据式(4)我们能得到样本花萼长度与模板交集区间数的距离,如表 4 所示.

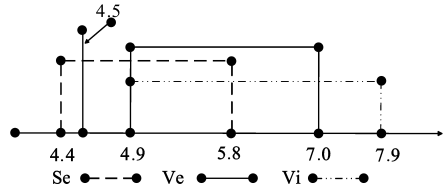

图3 抽取的样本与属性区间数的关系(SL)

第三步,求待测样本和模型属性之间的相似度.设式(5)中的支持系数 $\alpha = 5$,对表 4 中的数根据定义 6 即可求得相似度,如表 5 所示.

第四步,生成 BPA.对求得的相似度归一化,即可求得如表 5 所示的 BPA.

为了验证提出的 BPA 生成方法的有效性,设计了以下实验,该部分主要介绍了实验的总体思想,实验设计的两个目的,以及围绕这几个目的所做实验的细节,每个实验的最后部分都对实验结果做了分析和讨论.

表 4 样本花萼长度和模型区间数交集的距离(SL)

Hypothesis	Distance (SL)
{Se}	0.7234
{Ve}	1.5716
{Vi}	2.0881
{Se,Ve}	0.8888
{Se,Vi}	0.8888
{Ve,Vi}	1.5716
{Se,Ve,Vi}	0.8888

表 5 样本花萼长度和模型属性交集的相似度以及生成的 BPA(SL)

Hypothesis	Similarity	BPA
{Se}	0.2166	0.200
{Ve}	0.1129	0.104
{Vi}	0.0874	0.081
{Se,Ve}	0.1837	0.170
{Se,Vi}	0.1837	0.170
{Ve,Vi}	0.1129	0.104
{Se,Ve,Vi}	0.1837	0.170

4 实验

实验设计的总体思想是:为了介绍新提出的 BPA 生成方法在分类识别方面的应用,我们用鸢尾花数据集(Iris Data Set)作为验证数据库,详细介绍该方法的应用细节;为了探究所提出方法在少量数据下的有效程度,需要对生成样本区间数的样本规模做单变量分析.为此,设计了如下实验.

4.1 实验一:该方法在分类识别方面的应用

实验一可以描述成如下步骤:

(1)一共随机选择鸢尾花数据集(Iris Data Set) 的 120 个样本,其中每一个种类分别选择 40 个,用所得样本的最小值和最大值分别构造区间数模型,如表 6 所示.

(2)剩余的 30 个样本,其中每一个种类还剩 10 个,当作类别是未知的测试样本.

(3)通过求相似度的过程,决定 BPA,通过上例相似的过程,我们可以得到每个属性的 BPA,如表 7 所示.

(4)因为对于四种属性,可以构造四个证据,进而通过 DS 组合规则进行融合.证据理论组合规则满足交换律,结合律等一些优良性质,所以我们首先两两融合,然后融合成最终结果.

(5)未知样本的类型最终由融合后的结果决定,哪个 BPA 的值最大,那么它对应的类别即是未知样本的类别.

举上面的生成 BPA 的例子完成属性信息的融合,在上例中我们选取了样本为(4.5cm, 2.3cm, 1.3cm, 0.3cm).经过建立区间数模型,求相似度,得到 BPA(如表 7),进而运用证据理论融合,融合结果如表 7(Combined BPA)所示,从表 7 的最终融合结果可以明显的看出,这个样本的类别是 Setosa 类鸢尾花,与鸢尾花数据集(Iris Data Set)的数据相符合,这样事例说明了这个方法在鸢尾花数据集分类问题上的应用细节.

表 6　由样本统计的区间数模型

Item	Attributes			
	SL	SW	PL	PW
Se	[4.4,5.8]	[2.3,4.4]	[1.0,1.9]	[0.1,0.6]
Ve	[4.9,7.0]	[2.0,3.4]	[3.0,5.1]	[1.0,1.7]
Vi	[4.9,7.9]	[2.2,3.8]	[4.5,6.9]	[1.4,2.5]

表 7　实例的 BPA 和最终融合结果

Item	Attributes				
	SL	SW	PL	PW	Combined BPA
m(Se)	0.200	0.094	0.710	0.585	0.818
m(Ve)	0.104	0.171	0.118	0.163	0.115
m(Vi)	0.081	0.128	0.076	0.110	0.062
m(Se,Ve)	0.170	0.159	0	0	0
m(Se,Vi)	0.170	0.125	0	0	0
m(Ve,Vi)	0.104	0.164	0.096	0.142	0
m(Se,Ve,Vi)	0.170	0.159	0	0	0.006

为了整体了解该方法尾花数据集(Iris Data Set)分类问题上的有效程度,在支持系数 α 取 5,抽样容量为 120,生成的属性模板区间数如表 6 所示的条件下,对全部 150 个数据集进行测试.经过实验统计,得出整体的识别率为 96%,其中种类 Setosa 的鸢尾花的识别率为 100%,种类为 Versicolor 的鸢尾花的识别率为 98%,种类为 Virginica 的鸢尾花的识别率为 90%.同时通过其它 UCI 分类测试数据集[15,16]测试结果得出,该方法在区分各属性特征值相对集中且区分度较大的数据集上有很好的优势.

在有些目标识别的应用中,由于某些限制,我们得到的数据是有限的,比如军事领域,由于观测的限制或者由于保密的原因,建立目标属性的描述模型所用的样本数目较少,无法建立完整的识别模型,所以在少量数据情况下的识别问题就很有意义,实验二的设计目的是为了探究在数据量少的情况下该方法的有效程度.

4.2　实验二:探索该方法在少量数据情况下识别的有效性

为了测试提出方法在采样量较少的情况下识别率的情况,实验步骤设计如下:

(1)建立区间数模型.在此过程中,取不同规模的样本,考虑到数据量较少,第一次对鸢尾花的每一类分别随机取 1 个样本构建区间数模型.

(2)根据文中第三部分提出的方法生成 BPA.

(3)应用证据理论融合,得到识别结果.

(4)取鸢尾花总体作为测试集,从步骤 1 到步骤 4 重复独立进行实验 50 次,最后对统计的识别率取平均值.

(5)对构建区间数模型的样本规模依次递增,如 2,3,4,5,10,15,20,25,30,35,40,45,然后返回第一步,直到规模值遍历完毕.

实验中取支持系数 $\alpha = 5$,然后取平均值得到不同样本规模情况下的识别率情况,实验结果如表 8 所示,图 4 中更直观的证明了该方法在数据较少的情况下,也能得到较好的识别效果.

表 8　测试过程中建立模板区间数的样本容量(scale)取不同值时的识别率($\alpha = 5$)

	Items			
	Se	Ve	Vi	Total
scale = 1	0.9376	0.8384	0.7364	0.8375
scale = 2	0.9376	0.8924	0.862	0.906
scale = 3	0.9896	0.9256	0.8444	0.9199
scale = 4	0.9852	0.9328	0.8596	0.9259
…	…	…	…	…
scale = 40	1	0.9732	0.856	0.9431
scale = 45	1	0.98	0.856	0.9453

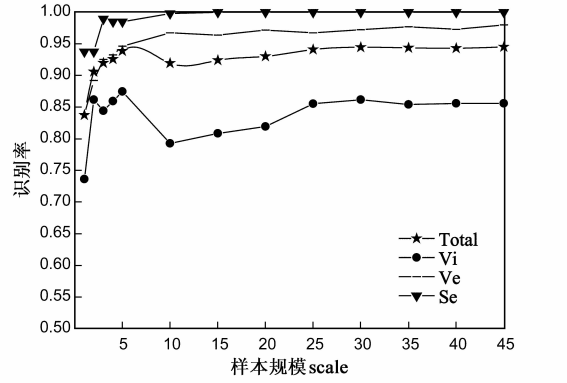

图4　识别率随建立模板区间数的样本规模(scale)的变化趋势

5 结论

用区间数的距离描述不确定信息需要的信息量较其它方法所需的信息量少,这一点显示出该方法对数据的要求较其它方法宽松,通过调整支持系数对识别率的分析能够得到实验中支持系数的最优值,同时实验通过较高的识别率验证了该方法在分类问题上的有效性,最后验证了该方法在少量数据下也有较高的的识别率,所以该方法具有简单、易行、适用于工程的特点.目前我们正在将该方法应用到不确定环境下的机载目标识别中.

参考文献

[1] Hall D L. Mathematical Techniques in Multi-sensor Data Fusion [M]. Boston: Artech House, 1992, 20 – 59.

[2] Dempster A. Upper and lower probabilities induced by multi-valued mapping [J]. Annual Mathematics Statistics, 1967, 38 (6): 325 – 339.

[3] 文成林,周哲,徐晓滨.一种新的广义梯形模糊数相似性度量方法及在故障诊断中的应用[J].电子学报,2011,39(3A):1 – 6.
WEN C L, ZHOU Z, XU X B. A new similarity measure between generalized trapezoidal fuzzy numbers and its application to fault diagnosis[J]. Acta Electronica Sinica, 2011, 39(3A): 1 – 6. (in Chinese)

[4] 刘明,袁宝宗,唐晓芳.证据理论 k-NN 规则中确定相似度参数的新方法[J].电子学报,2005,33(4):766 – 768.
LIU M, YUAN B Z, TANG X F. A new approach to determine the similarity parameters in evidence-theoretic k-NN rule[J]. Acta Electronica Sinica, 2005, 33(4): 766 – 768. (in Chinese)

[5] 何友,王国宏,陆大金,等.多传感器信息融合及应用[M].北京:电子工业出版社,2001.

[6] HAN D Q. Multiple classifiers fusion based on weighted evidence combination[A]. IEEE International Conference on Automation and Logistics [C]. Jinan, China: IEEE, 2007. 2138 – 2143. (in Chinese)

[7] 文成林,周东华,潘泉,张洪才.多尺度动态模型单传感器动态系统分布式信息融合[J].自动化学报,2001,27(2):117 – 124.
WEN C L, ZHOU D H, PAN Q, ZhANG H C. Distributed information fusion algorithm for single sensor dynamic system on the basis of multiscale dynamic models [J]. Acta Automatica Sinica, 2001, 27(2): 158 – 165. (in Chinese)

[8] 邓勇,施文康,朱振福.一种有效处理冲突证据的组合方法[J].红外与毫米波学报,2004,23(1):27 – 32.

[9] 韩崇昭,韩得强,介婧.从生物感知认识到系统工程方法论[J].系统工程理论与实践,2008(增刊):75 – 95.
HAN C Z, HAN D Q, JIE J. From biological cognition and perception to methodologies of system engineering [J]. Systems Engineering-Theory & Practice, 2008 (supplement): 75 – 93. (in Chinese)

[10] DENG Y, SHI W K, ZHU Z F, et al. Combining belief functions based on distance of evidence[J]. Decision Support Systems, 2004, 38(3): 489 – 493.

[11] DENG Y, JIANG W, XU X, et al. Determining BPA under uncertainty environments and its application in data fusion [J]. Journal of Electronics (China), 2009, 26(1): 13 – 17.

[12] WAN S P. Interval number method for object threat assessment[J]. Computer Engineering and Applications (China), 2009, 45(6): 32 – 34.

[13] Tran L, Duckstein L. Comparison of fuzzy numbers using a fuzzy distance measure[J]. Fuzzy Sets and Systems, 2002, 130 (3): 331 – 341.

[14] Iris Data Set. Famous Database for Pattern Recognition from Fisher[OL]. http://archive.ics.uci.edu/ml/datasets/Iris, 2011-3-20.

[15] Blood transfusion service center data set. Data Taken from the Blood Transfusion Service Center in Hsin-Chu City in Taiwan [OL]. http://archive.ics.uci.edu/ml/datasets/Blood + Transfusion + Service + Center, 2012-2-5.

[16] Vertebral Column Data Set. Data Set Containing Values for Six Biomechanical Features used to Classify Orthopaedic Patients into 3 Classes (normal, disk hernia or spondilolysthesis) [OL]. http://archive.ics.uci.edu/ml/datasets/Vertebral + Column, 2012-2-5.

作者简介

康兵义 男,1985 年 10 月生,河南信阳人,硕士研究生.研究方向:信息融合,智能信息处理.在 KNOWLEDGE BASED SYSTEM 等刊物上发表和录用论文 5 篇.
E-mail: kangby@swu.edu.cn

李 娅 女,1981 年 9 月生,四川彭州人,博士研究生.研究方向:信息融合,智能信息处理.
E-mail: jialuoluo99@163.com

邓 勇 男,1975 年 7 月生,湖南长沙人,现为西南大学计算机与信息科学学院教授,博士生导师,入选教育部新世纪优秀人才支持计划和上海市青年科技启明星计划,并获得重庆市杰出青年科学基金资助.研究方向:信息融合,智能信息处理,在国内外发表期刊论文 50 余篇.E-mail: ydeng@swu.edu.cn