ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 6,Nov. 2015,pp. 1318－1324
DOI:10.16337/j.1004-9037.2015.06.022

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于正态云模型的基本概率指派生成方法及应用

崔家玮[1]　李　冰[2]　李弼程[1]

(1.解放军信息工程大学信息系统工程学院,郑州,450002;2.中国电子系统设备工程公司,北京,100010)

摘　要: 基本概率指派(Basic probability assignment, BPA)生成是应用D-S证据理论的关键环节和第一步,而如何生成BPA仍然是一个有待解决的问题。本文提出一种基于云模型的BPA生成方法,首先,采用逆向云发生器生成每类样本在某属性下的正态云模型。其次,利用前件云发生器得到待测样本在该属性下对每类样本的确定度期望。再次,给出一种正态云模型交叠度计算方法,用确定度最大类的正态云模型与其他种类的最大交叠度作为对全集的信任度。最后,对确定度进行归一化得到待测样本的BPA。实验结果验证了该方法的有效性,此外,在样本数据较少情况下也能有效生成BPA。

关键词: 信息融合;云模型;证据理论;基本概率指派

中图分类号: TP391　　**文献标志码:** A

Determination of Basic Probability Assignment Based on Cloud Model and Application

Cui Jiawei[1], Li Bing[2], Li Bicheng[1]

(1. Institute of Information System Engineering, PLA Information Engineering University, Zhengzhou, 450002, China;
2. Chinese Institute of Electronics Equipment Corporation, Beijing, 100010, China)

Abstract: Determination of the basic probability assignment (BPA) is the first and main step to the evidence theory application. How to generate BPA is still an open issue. To solve the problem, a method for determining BPA based on the cloud model is proposed. Firstly, the normal cloud model of each sample under the property is constructed based on the backward cloud generator. Secondly, through the antecedent cloud generator repeatedly, the average certainty of the test sample under this property is obtained. Thirdly, a method for measuring the similarity of normal cloud models is proposed, and the maximal similarity of the normal cloud model is made, which has the maximal certainty as the belief of the universal set. Finally, the certainty is normalized to obtain the BPA of each class. The effectiveness of the method is proved by experiments, and it can generate BPA in the case of little samples numbers.

Key words: information fusion; cloud model; evidence theory; basic probability assignment (BPA)

引　　言

证据理论由Dempster于1967年提出,后经他的学生Shafer进一步发展完善,所以又被称为D-S证据理论[1],目前已成为不确定性信息处理的重要工具[2]之一,被广泛应用于模式分类[3]、风险评估[4]和

基金项目: 国家高技术研究发展计划("八六三"计划)(2012AA7032030D)资助项目。
收稿日期: 2013-11-11;**修订日期:** 2013-12-19

金融安全[5]等领域。在应用 D-S 证据理论的过程中,为了使用 D-S 合成规则,其关键一环就是获得传感器输出的基本概率指派 BPA,而 BPA 的自动生成一直以来就是研究的焦点,引起了许多学者的关注[6-10]。

通过多个传感器获取的不同信息综合判断某一事物特征时,存在一定的不确定性,需要采用有效方法进行多信息融合的识别,比如地震灾害的预测、事故隐患的判断和机电故障的识别等,而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。在多源异质信息融合中,由于信息的各异性,很多情况下无法在数据级或特征级进行融合,必须将数据转化为决策级信息来进行融合,而 BPA 的自动生成其实是将数据转换为决策级信息的过程,该过程的优劣很大程度上影响了融合的准确度和有效性,因此许多学者对合理地自动生成 BPA 开展了研究。蒋雯等[6-8]采用三角模糊数进行建模获得 BPA,但隶属度函数的确定过于主观;Xu 等[9]提出基于正态分布的 BPA 生成,但并未将随机性和模糊性综合考虑;康兵义等[10]利用数据样本上、下限生成区间数,该方法仅使用样本数据的上、下限,模型受样本数据影响较大。综上所述,现有的 BPA 生成方法存在所需样本规模过大、缺少合理建模依据和主观介入程度过强等问题。云模型是一种定量数值与定性概念之间的不确定性转换模型[11],能够更好地兼顾模糊性和随机性,在舆情预警[12]、算法改进[13]和网络安全[14]等领域也有广泛的应用。本文提出了一种新的 BPA 生成方法,这种方法将目标属性进行正态云模型转换,进而根据待测样本确定度和正态云模型交叠度获得 BPA。最后,本文通过鸢尾花数据集[15]分类实验验证了此方法的有效性。

1 相关基础知识

1.1 证据理论

定义 1 辨识框架辨识框架 Θ 表示人们对于某一判决问题所能认识到的所有可能的结果集合,人们所关心的任一命题都对应于 Θ 的一个子集。Θ 包含 N 个互斥且穷举的假设,即

$$\Theta=\{H_1,H_2,\cdots,H_N\} \tag{1}$$

由辨识框架 Θ 所有子集组成的集合称为 Θ 的幂集,记为 2^Θ,可表示为

$$\{\varphi,\{H_1\},\{H_2\},\cdots,\{H_n\},\{H_1\cup H_2\},\cdots,\Theta\} \tag{2}$$

对于辨识框架 Θ,问题域中任意命题 A 都应属于幂集 2^Θ,即 A 是 Θ 的子集。

定义 2 基本概率指派幂集 2^Θ 上的基本概率指派 m 定义为:$m:2^\Theta\to[0,1]$,满足

$$m(\Phi)=0 \tag{3}$$

$$\sum_{A\subseteq\Theta}m(A)=1 \tag{4}$$

式中:$m(A)$表示证据支持命题 A 发生的程度。$m(A)$表示证据对 A 本身的信任度大小,不能再细分给 A 的真子集(由于缺乏进一步的信息)。条件式(3)表示证据对于空集 Φ(空命题)不产生任何信任度,条件式(4)表示所有命题的信任度值之和等于 1,即总信任度为 1。

定义 3 焦元对于辨识框架 Θ,若 $m(A)>0(A\subseteq\Theta)$,则称 A 为证据的焦元。

定义 4 D-S 合成规则 假设辨识框架 Θ 下的两个证据 E_1 和 E_2,其相应的基本概率指派函数为 m_1 和 m_2,焦元分别为 A_i 和 B_j,则 D-S 合成规则为

$$m(A)=\begin{cases}\dfrac{\sum\limits_{A_i\cap B_j=A}m_1(A_i)m_2(B_j)}{1-K} & A\neq\phi\\ 0 & A=\phi\end{cases} \tag{5}$$

式中

$$K=\sum_{A_i\cap B_j=\phi}m_1(A_i)m_2(B_j) \tag{6}$$

它反映了各个证据之间的冲突程度，系数 $1/(1-K)$ 称为正则化因子，其作用是避免在合成时将非 0 的概率赋给空集 ϕ。

1.2 云模型

定义 5 设 U 是一个用精确数值表示的定量论域，C 为 U 上的定性概念。若定量值 $x\in U$，且 x 是定性概念 C 的一次随机实现，x 对 C 的确定度 $\mu(x)\in[0,1]$ 是具有稳定倾向性的随机数

$$\mu:U\rightarrow[0,1],\quad \forall x\in U,\quad x\rightarrow\mu(x) \tag{7}$$

则 x 在论域 U 上的分布称为云，每一个 x 称为一个云滴，表示为 $\mathrm{Drop}(x,\mu(x))$。云是由云滴组成，一个云滴是定性概念在数量上的一次实现，云滴越多越能反映该定性概念的整体特征。其中，云滴的确定度类似于模糊集合的隶属度，反映了模糊性，同时这个值自身也是个随机值，也可以用其概率分布函数描述。因此，云是模糊性和随机性的有机结合。

云可以用期望 E_x、熵 E_n 和超熵 H_e 等 3 个数字特征来表征一个概念，并将云用 $C(E_x,E_n,H_e)$ 表示。其中，期望 E_x 是云滴在论域空间上分布的期望，是该概念语言值量化的最典型样本。熵 E_n 为该定性概念语言值的不确定性度量，由该语言值的模糊性和随机性共同决定，表示在论域空间可以被定性概念接受的取值范围大小。超熵 H_e 为熵的不确定性度量，即熵的熵，由 E_n 的模糊性和随机性共同决定。正态云模型是重要的一种云模型，它具有普适性[16]，图 1 示意了正态云模型。$E_x=10$；$E_n=2$；$H_e=0.2$；$N=1\ 500$。由于正态分布是正态云在 $H_e=0$ 时的特例，在一个正态云中，$[E_x-3E_n,E_x+3E_n]$ 区间对云所表示概念的贡献达到 99.74%（即正态云的“$3E_n$ 规则”[11]）。本文方法将用到无确定度信息的逆向云发生器和前件云发生器。

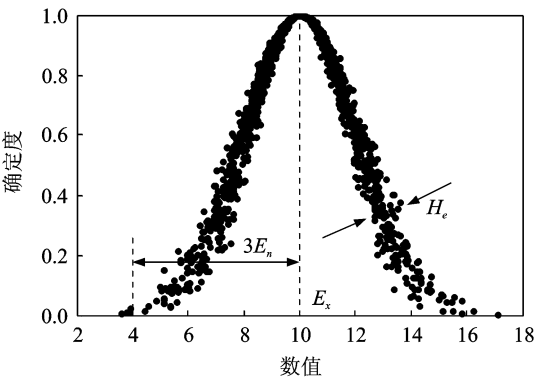

图 1 正态云模型

Fig. 1 Normal cloud model

2 新的 BPA 生成方法

2.1 正态云模型交叠度

定义 6 设 $C_1(E_{x1},E_{n1},H_{e1})$ 和 $C_2(E_{x2},E_{n2},H_{e2})$ 是两个正态云模型，则它们的交叠度 $S(C_1,C_2)$ 定义为

$$S(C_1,C_2)=\begin{cases}\dfrac{3(E_{n1}+E_{n2})-|E_{x1}-E_{x2}|}{3(E_{n1}+E_{n2})+|E_{x1}-E_{x2}|} & d>0\\ 0 & d\leqslant 0\end{cases} \tag{8}$$

式中：$d=3(E_{n1}+E_{n2})-|E_{x1}-E_{x2}|$。由式(8)可知，当 $d\leqslant 0$ 时，正态云 C_1 和 C_2 没有交叠部分，其交叠度为 0；当 $d>0$ 时，C_1 和 C_2 有交叠部分，且 $|E_{x1}-E_{x2}|$ 值越小，交叠的部分就越大；当 $E_{x1}=E_{x2}$ 时，C_1 和 C_2 期望相同，此时认为交叠度为 1。

2.2 基于正态云模型的 BPA 生成

本文方法的基本思想是：由于不同属性所具有的区分度不同，将属性下正态云模型的交叠度作为对全集 Θ 的赋值，交叠度大的属性不确定性大，交叠度小的属性不确定性小。该方法按照区分度对 BPA 的全集 Θ 赋予不同的值，使得生成的 BPA 更加合理有效。本文方法基本流程如图 2 所示，具体步骤如下。

步骤 1:将训练样本输入逆向云发生器,计算出每个种类的正态云模型数字特征。

步骤 2:将待测样本和正态云模型数字特征输入前件云发生器,计算出该样本对于每个种类的确定度。多次实现求期望,消除超熵 H_e 对确定度的影响。

步骤 3:根据每个种类的云数字特征计算出每个种类正态云模型与其他种类的交叠度,将待测样本确定度最大种类的正态云模型交叠度作为对全集 Θ 的 BPA。

步骤 4:根据全集 Θ 的 BPA,将确定度归一化,输出每个种类的 BPA。

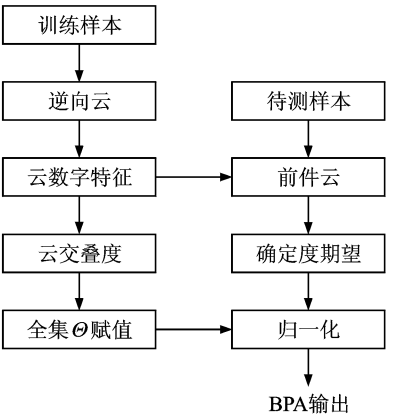

图 2 本文方法基本流程图

Fig. 2 Basic flowchart of method

在鸢尾花数据集中有 3 个种类的花,分别是 Setosa,Versicolor 和 Virginica。每个种类有 50 个样本,其中每一类花具有 4 种属性,分别是花萼长度(Sepal length,SL),花萼宽度(Sepal width,SW),花瓣长度(Petal length,PL)以及花瓣宽度(Petal width,PW)。

(1)生成正态云模型,即计算出云数字特征。对于鸢尾花的 3 个种类,随机抽取 40 个样本作为训练样本,剩余 10 个样本作为待测样本。将样本数据输入逆向云发生器,分别计算出 3 个种类在 4 种不同属性下的云数字特征。图 3 直观地给出了 3 类花在花瓣长度属性下的正态云模型。

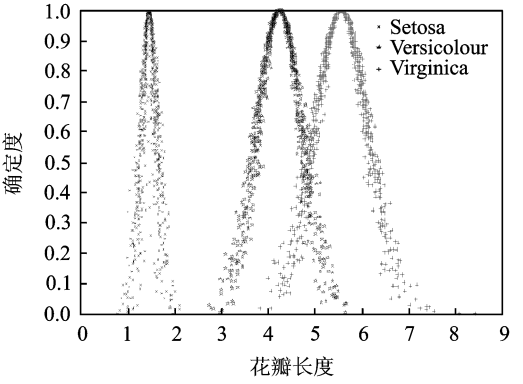

图 3 3 类鸢尾花在花瓣长度属性下的正态云模型

Fig. 3 Normal cloud model of three types on petal length

(2)将待测样本和云数字特征输出前件云发生器,多次计算后获得确定度期望值。从鸢尾花数据集中随机抽取一个待测样本,例如,待测样本是(5.2,2.7,3.9,1.4) cm,这个样本属于 Versicolor 类鸢尾花。将样本数据输入前件云发生器,由于 H_e 的存在,每次实现得到的确定度均不相同,经过多次随机实现,将确定度的期望值作为该待测样本属于该类的确定度。相比文献[6-8]的模糊数方法,该方法生成的 BPA 始终不为 0,避免了证据理论中的"0 悖论"。该待测样本对每类的确定度如表 1 所示。

表 1 待测样本对 3 类鸢尾花的确定度

Table 1 Membership of test sample for three types of iris

项目	Settosa	Versicolor	Virginia
SL	0.891 0	0.308 9	0.136 0
SW	0.132 1	0.983 5	0.642 8
PL	$5.332\ 1\times10^{-12}$	0.624 4	0.026 2
PW	$3.287\ 9\times10^{-16}$	0.969 7	0.114 8

(3)计算 4 种属性下确定度最大的鸢尾花种类正态云模型与其他两类花交叠度,将交叠度的最大值作为该类属性下赋给全集 Θ 的 BPA。4 种属性中确定度最大的鸢尾花种类与其他两类的交叠度如表 2 所示。

(4)将同一种属性下交叠度的最大值作为待测样本在该属性下对全集 Θ 的赋值,然后对 3 类鸢尾花的确定度进行归一化处理最终生成 BPA。

表 2 最大确定度种类与其他类正态云模型交叠度

Table 2 Similarity of maximum degree of species with other species

特征	最大确定度种类	与 Set 交叠度	与 Ver 交叠度	与 Vir 交叠度
SL	Settosa	Null	0.466 2	0.331 8
SW	Versicolor	0.505 6	Null	0.811 4
PL	Versicolor	0	Null	0.426 8
PW	Versicolor	0	Null	0.390 7

3 实验结果与性能分析

本文使用鸢尾花数据集作为验证数据库,采用经典 Demspter 证据合成公式对生成的 BPA 进行合成,通过融合分类实验验证该方法的有效性,并对训练样本规模对识别率的影响做了分析。

3.1 新方法在目标识别中的应用

实验 1 的步骤如下。

(1)随机选取鸢尾花数据集中的 120 个样本,其中每一个种类分别选择 40 个,将样本数据输入逆向云发生器,分别计算出 3 个种类在 4 种不同属性下云数字的特征。

(2)每一个种类的剩余 10 个样本作为待测样本,用来检测识别率。

(3)通过正态云模型交叠度确定全集 Θ 的 BPA,归一化得到每一个种类的 BPA。

(4)对于 4 种属性,构造 4 个证据,采用经典 Demspter 证据合成公式进行合成。

(5)待测样本的类型由证据融合的结果确定,使用基于 BPA 的简单决策方法,即所得 BPA 最大的种类就是待测样本的种类。

仍然以上面生成 BPA 的例子,选取样本为(5.2,2.7,3.9,1.4)cm,经过 Demspter 证据合成公式获得最终的融合结果如表 3 所示。在相同测试集情况下,将本文方法与其他分类方法如支持向量机、决策树方法、最邻近方法、文献[9,10]方法进行对比实验,结果如表 4 所示。其中,本文方法识别率与文献[10]方法同为 96%,识别率高于其他方法。

由表 4 的计算结果可知,该待测样本属于 Versicolor 类,与实际相符。为了充分验证该方法的有效性,在训练样本为 120 个的情况下,对全部 150 个鸢尾花数据进行测试。经过实验统计,得出整体的识别率为 96%。其中 Settosa 类识别率为 100%,Versicolor 类识别率为 94%,Virginia 类识别率为 94%。

表 3 待测样本 BPA 最终融合结果

Table 3 Final results of BPA test sample

项目	Settosa	Versicolor	Virginia	Θ
花萼长度	0.356 0	0.123 4	0.054 3	0.466 2
花萼宽度	0.014 2	0.105 5	0.068 9	0.811 4
花瓣长度	$4.698\ 1\times10^{-12}$	0.550 1	0.023 1	0.426 8
花瓣宽度	$1.959\ 5\times10^{-12}$	0.544 8	0.064 5	0.390 7
融合结果	0.087 2	0.749 6	0.053 5	0.109 7

表 4 多种分类方法识别率对比

Table 4 Comparison of recognition rate of multiple classification

分类方法	SVM-RBF	REPTree	1NN	文献[9]	文献[10]	本文方法
识别率/%	92.7	94.7	94	95.5	96	96

3.2 新方法在少量数据样本情况下的有效性

实验 2 步骤如下:(1)建立正态云模型。由于鸢尾花数据集中,花瓣宽度这一属性变化较小,样本过少无法生成正态云模型,所以第一次随机使用 5 个样本进行测试。(2)根据本文方法生成 BPA。(3)应用 Demspter 合成公式进行证据融合,取 BPA 最大种类为识别结果。(4)取所有鸢尾花数据作为测试样本,从步骤(1)到步骤(3)重复独立实验 100 次,取识别率平均值作为识别结果。(5)返回第(1)步,依次递增样本规模,从 5 个样本递增至 50 个样本,直到遍历完毕。

图 4 更加直观地证明了该方法在少量训练样本下的有效性。实验结果如表 5 所示。由表 5 可以看出,总体识别率最终收敛于 96%,当训练样本增加到 7 个以上,整体识别率与最终识别率波动小于 1%。与文献[9]相比,本文方法在同等识别率情况下所需训练样本更少。此外,文献[9]方法对 Versicolor 和 Virginia 类的识别率分别为 98%和 90%,本文均为 94%,可见本文方法性能较文献[9]方法更为稳定。

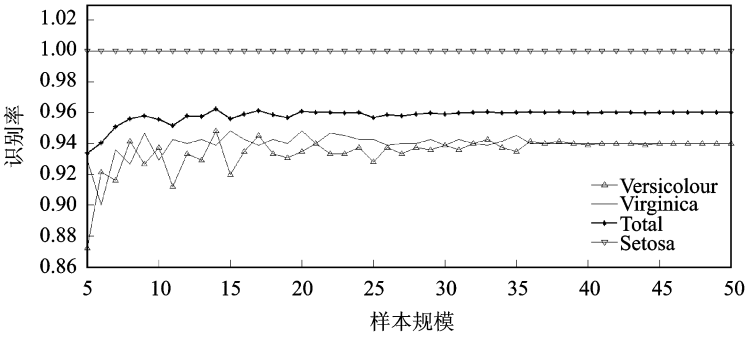

图 4 识别率随样本规模的变化趋势

Fig. 4 Trends of recognition rate with sample sizes

表 5 不同训练样本规模下的识别率

Table 5 Recognition rate of different training sample sizes

训练样本数	Settosa	Versicolor	Virginia	Total
5	1.000 0	0.872 0	0.929 3	0.933 8
6	1.000 0	0.921 3	0.900 0	0.940 4
7	1.000 0	0.916 0	0.936 0	0.950 7
8	1.000 0	0.941 3	0.926 7	0.956 0
⋮	⋮	⋮	⋮	⋮
17	1.000 0	0.945 3	0.938 7	0.961 3
50	1.000 0	0.940 0	0.940 0	0.960 0

4 结束语

云模型能够很好地表达数据的模糊性和随机性,而正态云模型自身的普适性保证了正态云模型在大多数情况下建模的合理性。本文将正态云模型交叠度作为对全集的 BPA,体现不同区分度属性生成 BPA 的不确定度不同。本文方法可用于异质信息融合中特征级向决策级的转换以使用 D-S 证据理论来实现融合。实验验证本文方法在目标识别问题上的有效性,并对训练样本依赖度小。该方法具有合理、易行和适用于工程实现的特点。

参考文献:

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

[2] 李弼程,黄洁,高世海,等. 信息融合技术及其应用[M]. 北京:国防工业出版社,2010:8-10.

Li Bicheng, Huang Jie, Gao Shihai, et al. Information fusion technology and application[M]. Beijing: National Defence Industry Press, 2010:8-10.

[3] Tabassian M, Ghaderi R, Ebrahimpour R. Combining complementary information sources in the Dempster-Shafer framework for solving classification problems with imperfect labels[J]. Knowledge-Based Systems, 2012(27):92-102.

[4] 崔建国,张杰,陈希成,等.信息融合在飞行器智能健康诊断中的应用[J].数据采集与处理,2012,27(2):236-240.
Cui Jianguo, Zhang Jie, Chen Xicheng, et al. Application of information fusion in aircraft intelligent health diagnosis[J]. Journal of Data Acquisition and Processing, 2012,27(2):236-240.

[5] Xiao Z, Yang X L, Pang Y, et al. The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster-Shafer evidence theory[J]. Knowledge-Based Systems, 2012(26):196-206.

[6] 蒋雯,张安,杨奇.一种基本概率指派的模糊生成及其在数据融合中的应用[J].传感器技术学报,2008,21(10):1717-1720.
Jiang Wen, Zhang An, Yang Qi. Fuzzy approach to construct basic probability assignment and its application in multi-sensor data fusion systems[J]. Chinese Journal of Sensors and Actuators, 2008,21(10):1717-1720.

[7] 邓勇,韩德强.广义证据理论中的基本概率指派生成方法[J].西安交通大学学报,2011,45(2):34-38.
Deng Yong, Han Deqiang. Methods to determine generalized basic probability assignment in generalized evidence theory[J]. Journal of Xi'an Jiaotong University, 2011,45(2):34-38.

[8] Deng Y, Jiang W, Xu X, et al. Determining BPA under uncertainty environment and its application in data fusion[J]. Journal of Electronics, 2009,26(1):13-17.

[9] Xu Peida, Deng Yong, Su Xiaoyan, et al. A new method to determine basic probability assignment from training data[J]. Knowledge-Based Systems, 2013(46):69-80.

[10] 康兵义,李娅,邓勇,等.基于区间数的基本概率指派生成方法及应用[J].电子学报,2012,6(40):1092-1086.
Kang Bingyi, Li Ya, Deng Yong, et al. Determination of basic probability assignment based on interval numbers and its application[J]. Chinese Journal of Electronics, 2012,6(40):1092-1086.

[11] 李德毅,杜鹢.不确定性人工智能[M].北京:国防工业出版社,2005:143-158.
Li Deyi, Du Yi. Artificial intelligence with uncertainty[M]. Beijing: National Defence Industry Press, 2005:143-158.

[12] 王振兴,郭毅,张连成,等.一种基于云模型的突发事件预警方法[J].信息工程大学学报,2012,13(1):120-128.
Wang Zhenxing, Guo Yi, Zhang Liancheng, et al. Early warning of sudden events on the cloud model[J]. Journal of Information Engineering University, 2012,13(1):120-128.

[13] 刘禹,李德毅,张光卫,等.云模型雾化特性及在进化算法中的应用[J].电子学报,2009,37(8):1651-1658.
Liu Yu, Li Deyi, Zhang Guangwei, et al. Atomized feature in cloud based evolutionary algorithm[J]. Chinese Journal of Electronics, 2009,37(8):1651-1658.

[14] 张仕斌,许春香.基于云模型的信任评估方法研究[J].计算机学报,2013,36(2):422-431.
Zhang Shibin, Xu Chunxiang. Study on the trust evaluation approach based on cloud model[J]. Chinese Journal of Computer, 2013,36(2):422-431.

[15] Iris Data Set. Famous database for pattern recognition from Fisher[EB/OL]. http://arehive.ics.uci.edu/ml/datasets/Iris, 2011-3-20.

[16] 李德毅,刘常昱.论正态云模型的普适性[J].中国工程科学,2004,6(8):28-34.
Li Deyi, Liu Changyi. Study on the universality of the normal cloud model[J]. Engineering Sciences, 2004,6(8):28-34.

作者简介:

崔家玮(1989-),男,硕士研究生,研究方向:信息融合和证据理论,E-mail:106798411@qq.com。

李冰(1983-),女,工程师,研究方向:智能信息处理。

李弼程(1970-),男,教授、博士生导师,研究方向:智能信息处理、信息融合。