参数自适应的析取云模糊置信规则识别方法

李双明[1,2],关 欣[1],王海滨[1]
(1. 海军航空大学,山东烟台 264001; 2. 92941 部队,辽宁葫芦岛 125001)

摘 要: 为获得准确的模糊置信规则结构参数,提出了参数自适应的析取云模糊置信规则识别方法. 为完成模糊域的自适应划分,提出了基于频数统计的双门限检测方法和基于包含度的双门限检测方法. 用云模型作为模糊集,改变熵系数和超熵系数,实现对模糊集形状的调整;前提属性的联接设置为析取逻辑关系,改进了证据的基本概率赋值方式,对规则权重和属性权重进行了优化. 实验结果表明,与其他方法相比,本文方法的正确识别率提高了 5%~15%,规则可解释性更强.

关键词: 参数自适应; 云模糊; 析取置信规则; 识别
中图分类号: TP391;TN957;TP274 **文献标识码:** A **文章编号:** 0372-2112(2022)02-0396-08
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20201253

Disjunctive Cloud Fuzzy Belief Rules for Recognition with Adaptive Parameters

LI Shuang-ming[1,2], GUAN Xin[1], WANG Hai-bin[1]
(1. *Naval Aviation University*, *Yantai*, *Shandong* 264001, *China*; 2. *Unit* 92941 *of PLA*, *Huludao*, *Liaoning* 125001, *China*)

Abstract: In order to obtain the accurate structural parameters of fuzzy rules, a disjunctive cloud fuzzy belief rules based recognition method with adaptive parameters is proposed. In order to complete the adaptive division of fuzzy domain, a double threshold detection method based on frequency statistics and inclusion degree separately are proposed. Cloud model is used as fuzzy set, of which by changing entropy coefficient and super entropy coefficient to adjust the shape of fuzzy set, and the connection of premise attributes is set as disjunctive logic relation. The basic probability assignment of conflict evidence is improved, and a programming model is built to optimize the rule weight and attribute weight. The experimental results show that compared with other methods, the correct recognition rate of our method is improved by 5%~15%, and the rules are more interpretable.

Key words: adaptive parameters; cloud fuzzy; belief rule; recognition

1 引言

数据分类在机器学习和数据挖掘中扮演着重要的角色,已有的数据分类方法大致可分为两种:“黑箱”方法和“白箱”方法. “黑箱”方法有支持向量机[1]、神经网络[2]以及它们的各种扩展方法[3],“白箱”方法有 K 近邻算法[4]、贝叶斯方法[5]、决策树方法[6]、模糊分类法[7]等. 由于复杂的电磁环境、测量设备的系统误差或测量手段的缺乏,不可避免地获得“低质量”的数据,这些数据往往呈现出不确定性,如模糊性、不精确性、不完备性,甚至数据缺失.

使用不确定知识表示和推理的三个最常见的框架是:贝叶斯概率理论[8]、Dempster-Shafer 理论[9]和模糊集理论[10]. 在知识表达系统中,最常见的知识系统为基于规则的系统,其大致分为三种:粗糙集[11]、决策树[6]以及基于“if-then”形式的规则[12]. 在简单“if-then”规则基础上发展而来的模糊规则分类系统(Fuzzy Rule-Based Classification System, FRBCS)已经成为处理分类问题有效的工具之一[13],但是推理过程采用平均加权策略,决策方法采用“单赢”策略,不能够处理不完备的信息,且受样本噪声的影响较大. 文献[14]对 FRBCS 进行了扩展,但该方法也没有实现对不完备信息的建模. 杨剑波教授基于 D-S 理论、决策理论和模糊集理

收稿日期:2020-11-06;修回日期:2021-01-26;责任编辑:梅志强
基金项目:国家自然科学基金青年基金(No.62001503);国防科技卓越青年科学基金(No.2017-JCJQ-ZQ-003);泰山学者工程专项经费(No.ts201712072)

论,提出了以置信结构建立混合规则库、以证据推理(Evidential Reasoning,ER)为推理机的新方法[15],该方法能够实现对不完备信息的建模,但是对于分类问题而言,该方法出现规则数量"爆炸"现象. 文献[16]提出了前提部分也嵌入置信结构的分类系统,该方法的规则数和训练样本数是相等的,在识别时会增加计算负担. 焦连猛提出了带有置信结构的模糊规则分类系统[17],该模型结合了置信结构和模糊集的各自优点,引入了特征权重,提出了数据驱动的置信规则库(Belief Rule Base,BRB)建模方法,该方法充分利用训练数据来映射特征空间和类空间的不确定联系,有效降低噪声数据对分类结果的影响.

尽管上述基于各种模糊规则的识别方法都有各自的优势,但同时也存在不足,其中最重要的问题为规则的可解释性. 文献[17]指出影响规则可解释性的主要原因包括规则结构、规则数量、特征数量、模糊划分的数量、模糊集的形状,其中规则结构包括特征属性的逻辑连接关系、特征属性权重、规则权重、规则前提部分的分布结构、规则结论部分的分布结构、规则结论的生成方式等. 为此,本文提出了参数自适应的云模糊置信规则识别方法.

2　析取云模糊置信规则识别系统

析取命题下的云模糊置信规则为

$$R^q:\ \text{if}\left(x_1\text{is}\tilde{A}_1^q\right)\vee\left(x_2\text{is}\tilde{A}_2^q\right)\vee\cdots\vee\left(x_P\text{is}\tilde{A}_P^q\right)$$
$$\text{then}\ \boldsymbol{C}^q=\left\{\left(\omega_1,\beta_1^q\right),\cdots,\left(\omega_M,\beta_M^q\right)\right\}\tag{1}$$

其中,R^q 表示第 q 条规则,其规则权重为 θ^q,属性权重为 $\delta_1,\delta_2,\cdots,\delta_P$,$q=1,2,\cdots,Q$,$Q$ 为置信规则库中规则的数量,P 为前提属性的数量,M 为推理结论的数量,$\boldsymbol{x}=\left(x_1,x_2,\cdots,x_p\right)^{\mathrm{T}}$ 为模式特征向量,$\tilde{A}_p^q$ 为云模糊集,每个属性模糊划分集为 $\left\{\tilde{A}_{p,1},\tilde{A}_{p,2},\cdots,\tilde{A}_{p,n_p}\right\}$,$\tilde{A}_p^q\in\left\{\tilde{A}_{p,1},\tilde{A}_{p,2},\cdots,\tilde{A}_{p,n_p}\right\}$,规则权重 $0\leqslant\theta^q\leqslant1$,属性特征权重 $0\leqslant\delta_p\leqslant1$,满足 $\sum_{p=1}^{P}\delta_p=1$.

析取云模糊规则识别系统需要解决以下几个问题:模糊集的划分、规则的产生、规则参数的确定. 在无先验知识的前提条件下,本文研究如何从数据自身来实现系统建模,因为系统结构及参数均基于传感器测量的数据而得到,并根据识别结果对其进行调整,故称为参数自适应的析取云模糊规则分类系统. 系统结构如图1所示.

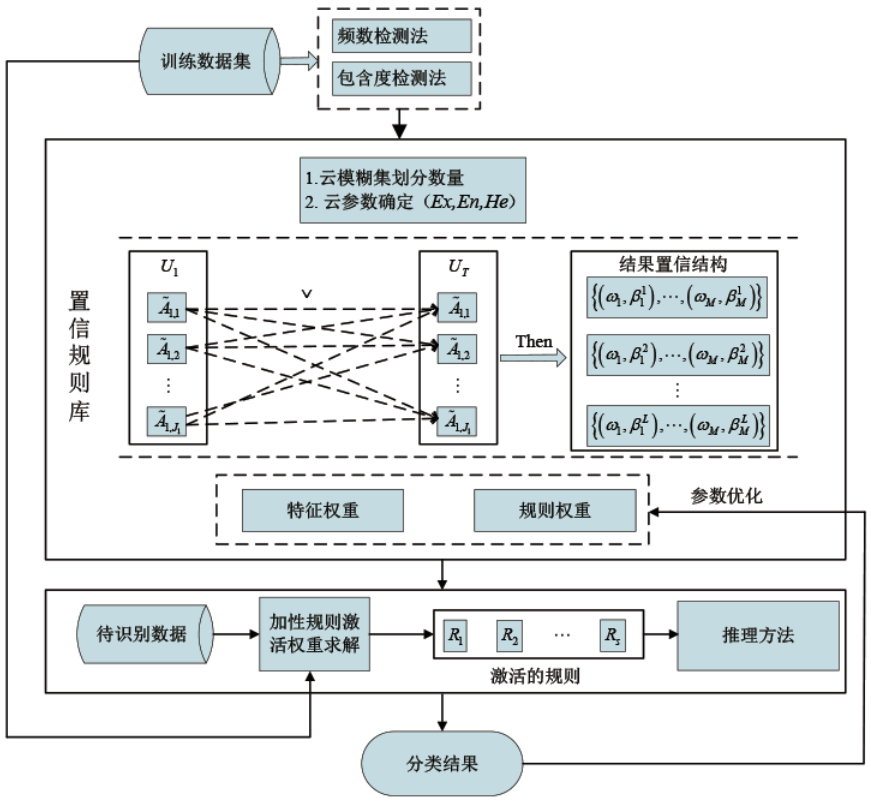


图1　参数自适应的云模糊规则识别系统结构图

2.1 特征域的模糊划分

2.1.1 基于频数的双门限检测方法

定义 1 对于描述某种特征属性的数据集合 $H=\left\{H_i \mid i=1,2,\cdots,m\right\}$，记 y 为集合 H 中的元素 H_i 出现的个数，称 $y=f\left(H_i\right)$ 为数据集合 H 的频数分布函数，则有式(2)成立，即

$$\sum_{i=1}^{m} f\left(H_i\right)=n \tag{2}$$

其中，n 为数据样本总量.

设置频率检测门限(数据点的频数与数据总量的比值)为 δ，当统计数据点的频数满足式(3)时，保留该数据分割点，否则放弃该数据分割点.

$$f\left(H_i\right)>n\cdot\delta \tag{3}$$

当两个数据分割点出现的频数都超过阈值且相距较近，从聚类的角度，这两个数据点应该为同一类数据，因此有必要舍去其中的一个数据点.

定义 2 记相邻的两个数据分割点为 H_i 和 H_{i+1}($i=1,2,\cdots,l,l\leqslant m-1\leqslant n-1$)，称式(4)为两个数据之间的分离度.

$$S\left(H_i,H_{i+1}\right)=\left(H_{i+1}-H_i\right)\Big/H_i \tag{4}$$

设置分离度检测门限为 λ，若经过频率检测门限 δ 检测后相邻两个数据分割点的分离度满足式(5)，即

$$S\left(Q_i,Q_{i+1}\right)\leqslant\lambda \tag{5}$$

那么，舍去其中的一个点，其原则为：将通过频率检测门限的数据点升序排列，首先计算第 1 个点和第 2 个点的分离度，若满足，则舍去第 2 个点，然后计算第 1 个点和第 3 个点的分离度，依次往下；否则，第 1 个点和第 2 个点都保留，然后计算第 3 个点和第 4 个点，依次往下.

2.1.2 基于包含度的双门限检测方法

根据数据聚类的思想，将聚类中心作为模糊域划分点. 文献[18]提出了基于数据包含度的自动聚类算法，该算法是一种基于密度的聚类算法，将自身数据密度大，且离其他数据点相对较远的数据点作为聚类中心. 对于数据个数较多时，上述方法耗时较大，基于上述方法，本文提出了改进的包含度检测方法，步骤如下.

步骤 1：将整个数据集升序排列可得 $x'_1,x'_2,\cdots,x'_n$.

步骤 2：以第一个数据点 x'_1 为起始点，依次计算下一个数据 $x'_i(i>1)$ 与 x'_1 之间的距离 $d(x'_i,x'_1)$，若小于截断距离 d_c，则将 x'_i 和 x'_1 划为一组，记为 S，若 $d(x'_i,x'_1)$ 大于截断距离 d_c，分组停止.

步骤 3：计算包含度 $|S|/n$，若 $|S|/n$ 小于给定的包含度阈值 u_c，则舍去该组数据，否则保留，记为 S_1.

步骤 4：以第 x'_{i+1} 为起始点，依次计算下一个数据 $x'_j(j>i+1)$ 与 x'_{i+1} 之间的距离 $d(x'_j,x'_{i+1})$，执行步骤 3，得到 S_t，遍历整个数据，执行步骤 5.

步骤 5：通过步骤 1~4 后，得到 n' 组数据 $S_t(t=1,2,\cdots,n')$，以 $\mathrm{mean}(S_t)$ 作为模糊域的分割点，n' 作为模糊域的划分数量.

2.2 云模糊集

本文以二阶正态云模型作为模糊集样式[19]，相比于三角形模糊集样式，其具有以下优势：

(1) 能够刻画数据的正态分布特性；

(2) 能够解决模糊集覆盖有限的问题；

(3) 能够调整参数改变模糊集的形状.

设模糊域分割点为 $\left\{p_1,p_2,\cdots,p_l\right\}$，相应地确定了 l 个云模型，按式(6)计算每个云模型的参数.

$$\begin{cases} E_{x,i}=p_i \\ E_{n,i}=k_{\mathrm{en}}\times\left(p_{i+1}-p_{i-1}\right) \\ H_{e,i}=k_{\mathrm{he}}\times E_{n,i} \end{cases} \tag{6}$$

其中，k_{en} 和 k_{he} 为常数，称为熵和超熵系数，决定了云模型的形状.

2.3 模糊规则库

2.3.1 规则的前提部分

对训练样本 $\boldsymbol{x}=\left(x_1,x_2,\cdots,x_n\right)^{\mathrm{T}}$，$x_i$ 对应第 i 个特征属性上的测量值，根据第 2.1 节中特征域上的云模糊集划分，遍历 x_i 隶属于第 i 个特征域的云模糊集合 $\tilde{A}_i=\left\{\tilde{A}_{i1},\tilde{A}_{i2},\cdots,\tilde{A}_{ij}\right\}$ 的隶属度，取每个特征域最大隶属度对应的云模糊集组合为一条规则的前提条件，在规则前提条件确定的过程中，同时也确定了支持该规则所包含的训练样本.

2.3.2 结论部分的置信结构

设第 q 条规则 R^q 包含的训练样本子集为 S^q，类标签集为 $\Omega=\left\{\omega_1,\omega_2,\cdots,\omega_M\right\}$，集合 S^q 中的第 i 个训练样本为 $\boldsymbol{x}_i=\left(x_{i1},x_{i2},\cdots,x_{iP}\right)^{\mathrm{T}}$，在每个特征上的隶属度分别为 $\mu_{\tilde{A}_1^q}\left(x_{i1}\right),\mu_{\tilde{A}_2^q}\left(x_{i2}\right),\cdots,\mu_{\tilde{A}_P^q}\left(x_{iP}\right)$，样本 $\boldsymbol{x}_i$ 与前提部分 $\tilde{\boldsymbol{A}}^q$ 的匹配程度为

$$\mu_{\tilde{A}^q}\left(x_i\right)=\frac{1}{P}\sum_{p=1}^{P}\mu_{\tilde{A}_p^q}\left(x_{ip}\right) \tag{7}$$

文献[17]将 $\Omega=\left\{\omega_1,\omega_2,\cdots,\omega_M\right\}$ 作为辨识框架，对于 S^q 中每一个训练样本 $\boldsymbol{x}_i$，把类别 $\mathrm{Class}(x_i)=\omega_m$ 当作支持 ω_m 为对应规则结论部分的一个证据. 在证据理论框架下，将 $\mu_{\tilde{A}^q}\left(x_i\right)$ 作为 ω_m 类的基本概率分配，由于该证据只支持 ω_m，不支持其它的任何一类，因此除 ω_m 外的其他类基本概率分配为零，将剩余置信 $1-\mu_{\tilde{A}^q}\left(x_i\right)$ 分配

给整个辨识框架Ω,该证据用下面的mass函数m_i^q来表示,即

$$\begin{cases} m_i^q(\omega_m)=\mu_{\tilde{A}^q}(x_i) \\ m_i^q(\Omega)=1-\mu_{\tilde{A}^q}(x_i) \\ m_i^q(A)=0,\forall A\in\Omega\backslash\omega_m \end{cases} \tag{8}$$

其中,$0<\mu_{\tilde{A}^q}(x_i)\leqslant 1$.

同样地,得到S^q中所有样本生成的证据,利用Dempster组合规则进行融合,得到融合后的mass函数m^q,那么规则R^q的结论部分置信度为

$$\begin{cases} \beta_m^q=m^q(\omega_m),m=1,2,\cdots,M \\ \beta_\Omega^q=m^q(\Omega) \end{cases} \tag{9}$$

当一个前提部分包括不同类别的数据样本时,生成的证据之间是高冲突的,用上面的组合规则进行融合是不合适的. 下面以两类数据进行说明.

例1 假设第q个前提组合$\tilde{A}^q$包含n个数据样本,分为ω_1和ω_2两类,ω_1类的样本数为n_1,ω_2类的样本数为n_2,满足$n_1+n_2=n$.

ω_1类的样本x_i生成的mass函数具有如下形式,即

$$\begin{cases} m_{i,\omega_1}^q(\omega_1)=\mu_{\tilde{A}^q}(x_i) \\ m_{i,\omega_1}^q(\omega_2)=0 \\ m_{i,\omega_1}^q(\Omega)=1-\mu_{\tilde{A}^q}(x_i) \end{cases} \tag{10}$$

按照上面的方法,产生ω_2类的mass函数.

Dempster组合规则具有交换律的特点,分两种情况进行组合.

(1)先对ω_1类样本生成的证据进行组合,当$n_1\geqslant N_1$时,所有证据的组合结果为

$$m_{\omega_1}^q(\omega_1)=1,m_{\omega_1}^q(\omega_2)=0,m_{\omega_1}^q(\Omega)=0$$

将$m_{\omega_1}^q$依次和m_{j,ω_2}^q进行组合,最终结果为

$$m^q(\omega_1)=1,m^q(\omega_2)=0,m^q(\Omega)=0$$

显然该证据得到的置信结果是不符合逻辑的,完全丢弃ω_2类的样本对置信度的贡献.

(2)先对两类证据分别组合,当$n_1\geqslant N_1$,$n_2\geqslant N_2$时,有

$$m_{\omega_1}^q(\omega_1)=1,m_{\omega_1}^q(\omega_2)=0,m_{\omega_1}^q(\Omega)=0$$

$$m_{\omega_2}^q(\omega_1)=0,m_{\omega_2}^q(\omega_2)=1,m_{\omega_2}^q(\Omega)=0$$

可见$m_{\omega_1}^q$与$m_{\omega_2}^q$是高冲突的,融合后的结果为

$$m^q(\omega_1)=0,m^q(\omega_2)=0,m^q(\Omega)=0$$

显然这样的结果是错误的.

为此本文提出了一种基于可靠度的置信结构生成方法.

定义3 设第q个前提组合$\tilde{A}^q$包含的样本数为n,类ω_m的样本数为n_{ω_m},则结论部分类别ω_m置信度的可靠度为

$$Sd_{\omega_m}=\frac{n_{\omega_m}}{n} \tag{11}$$

式(11)满足$\sum_{m=1}^{M}Sd_{\omega_m}=1$.

式(10)引入可靠度进行修正,得

$$\begin{cases} m_i^q(\omega_m)=\mu_{\tilde{A}^q}(x_i)\times Sd_{\omega_m} \\ m_i^q(A)=\mu_{\tilde{A}^q}(x_i)\times Sd_A,\forall A\in\Omega\backslash\omega_m \\ m_i^q(\Omega)=1-\mu_{\tilde{A}^q}(x_i) \end{cases} \tag{12}$$

式(12)是一种新的基本概率赋值,可以看做是广义的证据源修正,然后根据Dempster组合规则进行融合,得到结论部分的置信结构分布.

2.3.3 规则权重和特征权重的优化模型

记目标函数为

$$f(\theta_q,\delta_p)=\sum_{i=1}^{T}E_i \tag{13}$$

其中,T为训练数据集的大小. 对每一个样本,若系统识别结果正确,$E_i=0$,否则$E_i=1$. 则优化目标模型为

$$\min f(\theta_q,\delta_p)=\sum_{i=1}^{T}E_i \tag{14}$$

$$\text{s.t.}\begin{cases} 0\leqslant\delta_p\leqslant 1 \\ \sum_{p=1}^{P}\delta_p=1 \\ 0\leqslant\theta_q\leqslant 1 \end{cases} \tag{15}$$

2.4 推理算法

2.4.1 规则激活

设$\boldsymbol{y}=(y_1,y_2,\cdots,y_P)^{\mathrm{T}}$表示要分类的未知目标. 该目标的特征测量值或者是完备的,或者缺失某些特征测量值. 如果某些特征测量值缺失,那么属于相应模糊划分域的匹配度为零,采用加权平均算子获取未知目标在规则R^q模糊域$\boldsymbol{A}^q$上的匹配度为

$$\mu_{A^q}(\boldsymbol{y})=\sum_{p=1}^{P}\delta_p\mu_{A_p^q}(y_p) \tag{16}$$

其中,$\mu_{A_p^q}(\cdot)$为前提云模糊集A_p^q的隶属度函数,δ_p为特征权重.

$\mu_{A^q}(\boldsymbol{y})$的值尽管很小,但总不为零,所以有必要设置规则激活阈值σ,当且仅当$\mu_{A^q}(\boldsymbol{y})>\delta$时,规则$R^q$才被激活,否则不被激活. σ用于控制被激活的规则数量,σ的取值不同,激活的规则数量也不相同. σ的取值越小,被激活的规则数量就越多,直至规则库中所有的规则被激活. 那么该如何选取σ呢? 可以根据实际情况,对σ的取值主观设定. 对于正态云而言,99.7%的云滴都

落在$[Ex-3En, Ex+3En]$的区间内，即云模糊集的绝大部分贡献都处于$[Ex-3En, Ex+3En]$区间内. 所以σ的取值可以设定为边界点$Ex-3En$和$Ex+3En$对应的隶属度，近似地，本文取$\sigma=\mathrm{e}^{-4.5}\approx 0.011\ 1$.

设Q'表示被未知目标$\boldsymbol{y}=\left(y_1, y_2, \cdots, y_P\right)^{\mathrm{T}}$激活的规则集，有

$$Q'=\left\{R^q \mid \mu_{A^q}(\boldsymbol{y}) \geqslant \sigma, q=1,2,\cdots,Q\right\} \tag{17}$$

规则R^q的激活度α^q由两个因素决定：匹配度$\mu_{A^q}(\boldsymbol{y})$和规则权重$\theta^q$. $\mu_{A^q}(\boldsymbol{y})$反应了未知目标与置信规则前提部分的相似程度，$\theta^q$反应了置信规则的稳定程度. 定义

$$\alpha^q=\frac{\mu_{A^q}(\boldsymbol{y})\theta^q}{\sum_{q=1}^{Q'}\mu_{A^q}(\boldsymbol{y})\theta^q}, R^q \in Q' \tag{18}$$

2.4.2 推理决策

用Shafer的折扣算子对激活的置信规则进行折扣，有

$$\begin{cases} m^{\alpha^q}\left(\omega_m\right)=\alpha^q \cdot \beta_m^q \\ m^{\alpha^q}(\Omega)=\alpha^q \cdot \Omega_m^q+1-\alpha^q \end{cases} \tag{19}$$

用Dempster组合规则对激活的规则进行组合，对任意$m^{\alpha^q}(\cdot)$，及$m^{\alpha^q}(\Omega)\neq 0$外，其他元素的基本概率赋值都为零，$Q'$个组合规则的解析表达式为

$$\begin{aligned} m\left(\omega_m\right)&=K\left(\prod_{q=1}^{Q'}\left(m_{m,q}+m_{\Omega,q}\right)-\prod_{q=1}^{Q'}m_{\Omega,q}\right) \\ m(\Omega)&=K\prod_{q=1}^{Q'}m_{\Omega,q} \end{aligned} \tag{20}$$

$$K=\left(\sum_{m=1}^{M}\left(\prod_{q=1}^{Q'}\left(m_{m,q}+m_{\Omega,q}\right)\right)-(M-1)\prod_{q=1}^{Q'}m_{\Omega,q}\right)^{-1}$$

其中，K表示归一化系数，$q=1,2,\cdots,Q', m=1,2,\cdots,M$.

采用置信度最大的原则进行决策，即

$$\omega=\arg\max_{\omega_m}\left\{m\left(\omega_m\right), m(\Omega)\right\} \tag{21}$$

则ω为识别结果.

3 实验验证

3.1 仿真数据验证

以电子侦察系统中的雷达辐射源识别为例，对本文所提方法进行验证. 设有3类雷达，选择射频频率(Radio Frequency，RF)、脉冲重复间隔(Pulse Repetition Interval，PRI)和脉宽(Pulse Width，PW)作为雷达的特征参数，各类雷达每种特征属性上的测量值服从正态分布，各类雷达特征参数见表1. 每类雷达仿真生成两类正态随机数据：一类具有统计特征分布，用来验证频数检测方法；一类不具有统计特征分布，用来验证包含度方法. 在两类数据中，每种雷达各有200个样本，共计600个数据样本，并以该数据作为训练数据，进行系统建模. 为消除量纲的影响，仿真中用到的数据进行了归一化处理.

表1 雷达特征参数

雷达类	RF/MHz		PRI/μs		PW/μs	
	均值	标准差	均值	标准差	均值	标准差
1	4 100	50	3 400	50	0.2	0.2
2	4 200	50	3 500	50	0.6	0.2
3	4 300	50	3 600	50	1.0	0.2

3.1.1 正确识别率分析

当识别系统建好后，用两种测试数据进行测试：一种是以训练数据作为测试数据(无噪声)；另一种是在训练数据集内随机抽取，并分别叠加2%，5%，10%，15%，20%的干扰噪声生成测试数据. 对这两种测试数据，分别进行1 000次Monte Carlo实验，其仿真结果如表2所示. 表2中，数据1表示具有统计分布特征的仿真数据，数据2表示无统计分布特征的仿真数据.

表2 正确识别率/%

	划分	不同噪声等级上的测试数据正确识别结果					
		无	2%	5%	10%	15%	20%
数据1	(3,3,4)	91.70	91.33	90.00	88.60	78.30	73.10
	(5,5,5)	95.17	94.34	91.10	90.50	81.30	79.90
	(7,6,5)	100.00	99.90	96.30	91.50	90.80	84.40
数据2	(3,3,4)	99.50	98.70	97.40	96.30	93.90	89.70
	(5,5,6)	99.67	99.20	99.40	99.00	94.10	91.30
	(6,4,6)	100.00	100.00	99.70	96.90	93.40	86.70

系统云模型的参数设为$k_{\mathrm{en}}=1.2, k_{\mathrm{he}}=0.001$. 不同的门限参数，模糊域划分的数量是不同的. 模糊域划分(3,3,4)表示特征RF、PRI和PW上的模糊分割数为3个模糊集、3个模糊集和4个模糊集.

对于数据1，训练数据集上的识别结果要优于含有噪声的测试数据集，并随着噪声的增大，正确识别率逐渐降低，当加入20%的噪声时，3种模糊划分的正确识别率是最低的，分别为73.1%，79.9%，84.4%. 无论是训练数据集还是含有噪声的测试数据集，随着模糊域划分的精细，即划分的模糊集数量增多，正确识别率逐渐增大，对于数据2也有同样地结论. 此外，在模糊划分数基本相同的情况下，数据1的识别结果要差于数据2的识别结果，例如，当数据1和数据2中的模糊域划分都为(3,3,4)时，数据2的正确识别率要比数据1高7.5%~18.51%，其他的模糊域划分也是如此. 其原因为，尽管数据1和数据2的样本总量是相同的，但在同种特征属性上，样本数量是不同的，数据1的样本量要明显小于数据2的样本量，因此相对地讲，数据1的样

本量是小于数据2的样本量的,所以会出现表中的结果. 表2中只给出了部分不同模糊划分下的识别结果,缺少对相关参数的敏感性分析,在3.1.2节中讲述.

3.1.2　参数敏感性分析

选取3.1.1节中的数据作为训练数据集,可调节的参数包括包含 u_c,d_c,k_{en},k_{he},δ 和 λ. 下面给出各参数的取值范围及变化步长:

(1)$0.07 \leqslant u_c \leqslant 0.16$,变化步长0.005;

(2)$0.07 \leqslant d_c \leqslant 0.16$,变化步长0.005;

(3)$0.8 \leqslant k_{\mathrm{en}} \leqslant 3.8$,变化步长0.25;

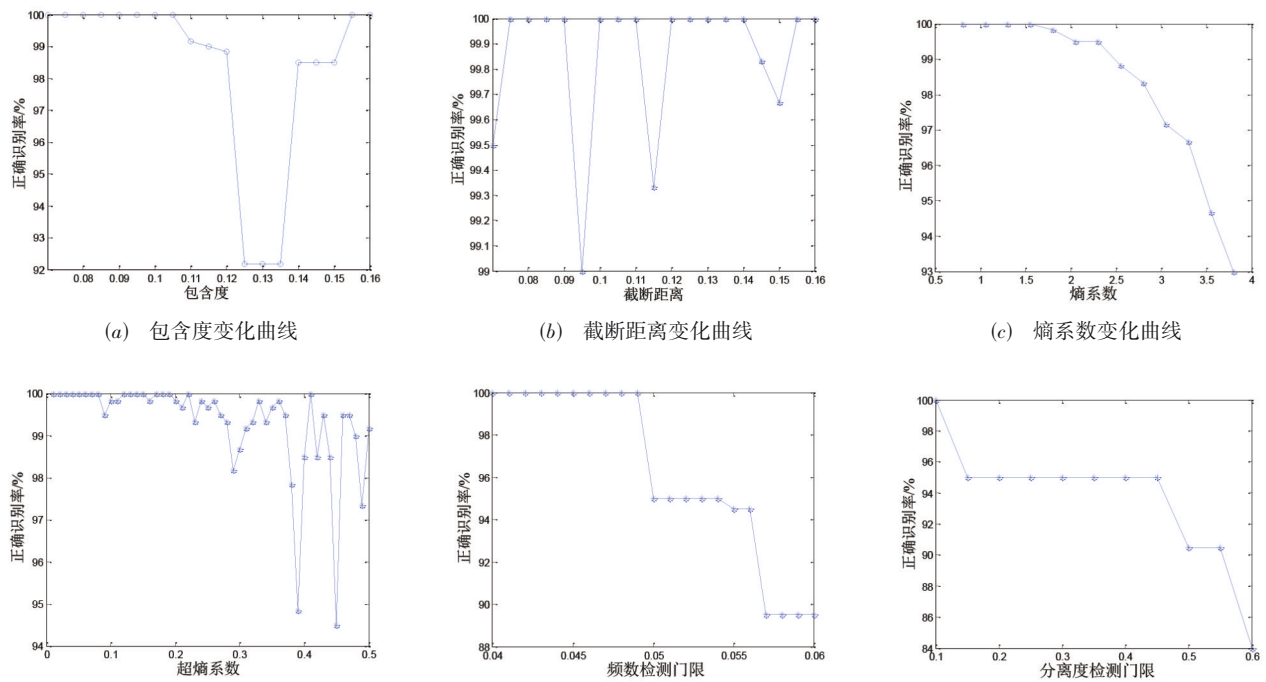
(4)$0.01 \leqslant k_{\mathrm{he}} \leqslant 0.5$,变化步长0.01;

(5)$0.04 \leqslant \delta \leqslant 0.06$,变化步长0.001;

(6)$0.1 \leqslant \lambda \leqslant 0.6$,变化步长0.05.

仿真结果随参数变化情况如图2所示.

图2(*a*)中,随着包含度的增大,正确识别率呈现先下降后上升的"凹陷"现象. 当包含度 u_c 为0.125,0.13和0.135时,正确识别率是最小的,约为92%;当 $0.07 \leqslant u_c \leqslant 0.115$ 和 $0.14 \leqslant u_c \leqslant 0.16$ 时,正确识别率都比较高. 进一步分析,当包含度为0.12,0.125,0.13,0.135和0.14时,其模糊域划分分别为(5,4,3)、(4,4,3)、(4,4,3)、(4,4,3)和(4,3,3),划分(5,4,3)多于划分(4,4,3),识别率高,而划分(4,3,3)少于划分(4,4,3),识别率也高,说明识别率并不随着划分的数量增多而得到改善,而是存在一个最优的组合. 图2(*b*)中,尽管不同的截断距离上的识别结果是不同的,也不存在固定的变化规律,但是识别率都在99%以上,说明截断距离对识别结果的影响是最小的. 图2(*c*)中,正确识别率随着熵系数的增大逐渐降低. 图2(*d*)中,随着超熵系数的逐渐增大,识别结果呈"震荡"式变化,在超熵系数为0.39以及0.45时,出现了明显的"断崖"式下降,因此在选择该参数时,应当尽量地避开这些点,可以选择较小的数值. 图2(*e*)和图2(*f*)中随着检测门限的提高,系统的识别性能都呈现下降趋势,因此在选择这两个参数时,可以考虑选择较小的数值.



(*a*) 包含度变化曲线　(*b*) 截断距离变化曲线　(*c*) 熵系数变化曲线

(*d*) 超熵系数变化曲线　(*e*) 频数检测门限变化曲线　(*f*) 分离度检测门限变化曲线

图2　仿真结果随参数变化情况

3.2　实测数据集验证

利用UCI中的实测数据集,将本文方法与模糊置信规则分类系统(Belief Rule-Based Classification System,BRBCS)、支持向量机(Support Vector Machine,SVM)、核函数极限学习机(Kernel Extreme Learning Machine,KELM)等方法进行对比分析,采用*B*-折交叉验证(*B*-Fold Cross-Validation,*B*-CV)的方法计算正确识别率,本文选用5-折交叉验证. 实验中选用Iris,Banknote,Ecoli,Seeds及Haberman 5类数据集,每类数据集的样本数量、属性数量和类别数量详见表3.

在支持向量机分类方法中,其惩罚系数为1,核系数为0.01,核函数为RBF核. 在核函数极限学习机分类方法中,其惩罚系数为1,核系数为1,核函数为RBF核. 本文中的熵系数和超熵系数分别为1.2和0.001. 仿

表3 数据集信息描述

数据集名称	样本数量	属性数量	类别数量
Iris	150	4	3
Banknote	1 372	4	2
Ecoli	336	5	8
Seeds	210	7	3
Haberman	306	3	2

真实验结果见表4.

在BRBCS方法中,因为无先验知识,采用简单的模糊格主观划分方法,且每个属性上的划分数相同,划分情况分为3种情况:每个属性划分为3个模糊集、5个模糊集和7个模糊集. 从表4中可知,对同种数据集而言,并不是划分的数量越多,识别结果就越好. 在Iris,Haberman以及Banknote数据集上,精细的模糊域划分提高了识别结果;但是在数据集Ecoli和Seeds上,精细的模糊域划分,反而降低了系统的分类性能. 这说明BRBCS分类系统的识别系统与模糊域的划分没有规律可循,若要得到较优的分类性能,需要主观反复地进行验证,以此来确定满足系统较优分类性能的模糊域划分. 在实际应用中,尤其对实时性有一定要求的场景,显然该方法是比较消耗时间的.

表4 正确识别率/%

数据集名称	BRBCS				KELM	SVM	本文方法
Iris	划分	(3,3,3,3)	(5,5,5,5)	(7,7,7,7)			(4,3,4,4)
	识别结果	94.17(4)	95.00(3)	96.67(2)	91.30(5)	55.33(6)	97.33(1)
Banknote	划分	(3,3,3,3)	(5,5,5,5)	(7,7,7,7)			(3,3,3,3)
	识别结果	82.90(5)	83.60(4)	66.62(6)	84.50(3)	99.42(1)	88.30(2)
Ecoli	划分	(3,3,3,3,3)	(5,5,5,5,5)	(7,7,7,7,7)			(6,4,3,4,3)
	识别结果	61.30(2)	59.20(3)	57.44(4)	55.78(5)	43.09(6)	69.64(1)
Seeds	划分	(3,3,3,3,3,3,3)	(5,5,5,5,5,5,5)	(7,7,7,7,7,7,7)			(5,4,5,4,7,5,4)
	识别结果	87.14(2)	86.66(4)	86.67(3)	85.24(5)	83.10(6)	90.48(1)
Haberman	划分	(3,3,3)	(5,5,5)	(7,7,7)			(5,5,3)
	识别结果	40.85(6)	57.84(5)	63.34(3)	76.79(1)	61.36(4)	73.30(2)

KELM、SVM方法在5种数据集上的总体识别结果要差于本文方法,但存在例外,SVM方法在Banknote数据集上的识别结果在所有方法中是最好的,KELM方法在Haberman数据集上的识别结果是最优的. 造成这种结果的原因主要是KELM、SVM方法作为典型基于数据的机器学习方法,用于学习的样本数量要满足一定的数量,较少的训练数据会造成“过拟合”现象.

从识别结果的排名上来看,本文所提方法的识别性能总体上是最优的. 与BRBCS方法相比能够用较少的模糊划分数量而达到较高的准确识别率. 如在Iris数据集上,本文方法的划分数量为(4,3,4,4),正确识别率为97.33%;而BRBCS方法当所用的模糊划分为(7,7,7,7)时的正确识别率为96.67%. 降低模糊划分数量带来的优势为:一是增强系统的可解释性,二是生成的规则数量降低,进而能够降低系统运行的时间,提高系统分类性能的实时性. 在其他4种数据集上,本文方法同样是在较少的模糊划分上获得了较高的识别结果. 与SVM、KELM用于大样本的分类方法相比,本文方法在处理小样本数据识别问题上具有较好的优势.

4 结论

本文提出了参数自适应的析取云模糊置信规则识别方法. 通过两种双门限检测方法,能够有效快速地确定模糊域划分的优化组合方式. 调整云模型的熵和超熵系数,可以改变模糊集的形状. 引入可能度,有效处理冲突条件下置信结果的基本概率赋值问题,并根据优化模型,实现对规则权重和属性权重的优化. 最后,用仿真数据集和实测数据集对所提方法进行验证. 结果表明,设置较少的模糊划分就可获得较高的识别率,能够有效处理小样本识别率低的问题,同时规则的可解释性得到了改善.

参考文献

[1] UTKIN L V. An imprecise extension of SVM-based machine learning models[J]. Neuro Computing, 2019, 331: 18-32.

[2] 吴玉佳, 李晶, 宋成芳, 等. 基于高效用神经网络的文本分类方法[J]. 电子学报, 2020, 48(2): 279-284.
WU Y J, LI J, SONG C F, et al. High utility neural networks for text classification[J]. Acta Electronica Sinica, 2020, 48(2): 279-284. (in Chinese)

[3] DENG C W, HUANG G B, JIA X U, et al. Extreme learning machines: New trends and applications[J]. Science China, 2015, 58(2): 5-20.

[4] MUHAMMAD A. K-nearest neighbor for recognize handwritten arabic character[J]. Mantik: Jurnal Matematika,

2019, 5(2): 83-89.

[5] GUO K, LIU X B, GUO L H, et al. A new constrained maximum margin approach to discriminative learning of Bayesian classifiers[J]. Frontiers of Information Technology & Electronic Engineering, 2018, 19(5): 639-650.

[6] SUN T, ZHOU H. Structural diversity for decision tree ensemble learning[J]. Frontiers of Computer Science, 2018, 12(3): 560-570.

[7] CIRO C, MARIA A. Interpretable fuzzy partitioning of classified data with variable granularity[J]. Applied Soft Computing, 2019, 74: 567-582.

[8] 谭翔元, 高晓光, 贺楚超. 基于马尔科夫毯约束的最优贝叶斯网络结构学习算法[J]. 电子学报, 2019, 47(9): 1898-1904.
TAN X Y,GAO X G,HE C C. Learning optimal Bayesian network structure constrained with markov blanket[J]. Acta Electronica Sinica, 2019, 47(9): 1898-1904. (in Chinese)

[9] YE I M, XU Z H, Gou X J. A new perspective of Bayes formula based on D-S theory in interval intuitionistic fuzzy environment and its applications[J]. International Journal of Fuzzy Systems, 2019, 21(4): 1196-1213.

[10] ZADEH L A. Fuzzy sets[J]. Information and Control, 1965, (8): 338-353.

[11] MENG Z Q, SHI Z Z. On rule acquisition methods for data classification in heterogeneous incomplete decision systems[J]. Knowledge-Based Systems, 2020, 193: 105472.

[12] SUN R. Robust reasoning: Integrating rule-based and similarity-based reasoning[J]. Artificial Intelligence, 1995, 75 (2): 241-295.

[13] SAMANTARAY S R. Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection[J]. Applied Soft Computing, 2013, 13: 928-938.

[14] TANG M, CHEN X, HU W D, et al. Generation of probabilistic fuzzy rule base by learning from examples[J]. Information Sciences, 2012, 217: 21-30.

[15] YANG J B, LIU J, WANG J, et al. Belief rule-base inference methodology using the evidential reasoning approach-RIMER[J]. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2006, 36 (2): 266-285.

[16] LIU J, MARTINEZ L, CALZADA A, et al. A novel belief rule base representation, generation and its inference methodology[J]. Knowledge-Based Systems, 2013, 53: 129-141.

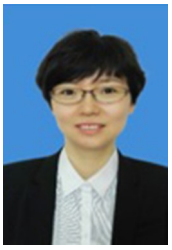
[17] JIAO L M, PAN Q, DENOEUX T, et al. Belief rule-based classification system: Extension of FRBCS in belief functions framework[J]. Information Sciences, 2015, 309: 26-49.

[18] 马云红, 王成汗, 江腾蛟, 等. 一种基于数据包含度的自动聚类算法[J]. 西北工业大学学报, 2016, 34(5): 863-866.
MA Y H,WANG C H, JIANG T J, et al. An automatic clustering algorithm based on data contained ratio[J]. Journal of Northwestern Polytechnical University, 2016, 34(5): 863-866. (in Chinese)

[19] 李双明, 关欣, 赵静, 等. 一种参数区间交叉类型的目标识别方法[J]. 北京航天航空大学学报, 2020, 46(7): 1307-1316.
LI S H, GUAN X, ZHAO J, et al. A methodology for target recognition with parameters of interval cross type[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(7): 1307-1316. (in Chinese)

作者简介

李双明　男,1986年出生,山东梁山人. 博士研究生. 主要研究方向为智能识别、不确定信息处理.
E-mail：aminglishuang@126.com



关　欣　女,1978年出生,辽宁锦州人. 教授. 主要研究方向为智能信息处理、多源信息融合.
E-mail：gxtongwin@163.com

王海滨　男,1982年出生,内蒙古赤峰人. 副教授. 主要研究方向为智能信息处理、多源信息融合.
E-mail：hesonwhb@163.com