第46卷第9期 中国科学技术大学学报 Vol. 46, No. 9
2016年9月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Sep. 2016

文章编号:0253-2778(2016)9-0764-10

基于差分进化算法的置信规则库推理的分类方法

刘莞玲[1],王韩杰[1],傅仰耿[1],杨隆浩[2],吴英杰[1]

(1.福州大学数学与计算机科学学院,福建福州 350116;2.福州大学决策科学研究所,福建福州 350116)

摘要:现有基于置信规则库参数学习的分类系统存在着一些问题,如分类准确度受模糊子区间划分数量约束,成非严格正相关关系;参数学习方法需人为给定规则数量;推理过程未体现特征与分类结果关联度等.为解决这些问题,提出基于差分进化算法的置信规则库推理的分类方法,该方法包括置信规则库分类系统构建及参数训练.首先引入置信规则库分类系统构建策略确定规则数;然后使用置信推理方法作为分类查询推理机;最后结合差分进化算法建立训练模型.在实验分析中,首先通过与现有分类方法进行对比,验证该方法的有效性;再通过对比不同区间划分数的置信规则库分类系统,说明参数训练的合理性.实验结果表明,该方法合理有效.

关键词:置信规则库推理;分类系统;参数学习;差分进化算法

中图分类号:TP18 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2016.09.008

引用格式:刘莞玲,王韩杰,傅仰耿,等.基于差分进化算法的置信规则库推理的分类方法[J].中国科学技术大学学报,2016,46(9):764-773.

LIU Wanling, WANG Hanjie, Fu Yanggeng, et al. Belief rule based inference methodology for classification based on differential evolution algorithm[J]. Journal of University of Science and Technology of China, 2016,46(9):764-773.

Belief rule based inference methodology for classification based on differential evolution algorithm

LIU Wanling[1], WANG Hanjie[1], FU Yanggeng[1], YANG Longhao[2], Wu Yingjie[1]

(1. *College of Mathematics and Computer Science, Fuzhou University, Fuzhou* 350116, *China*;
2. *Decision Sciences Institute, Fuzhou University, Fuzhou* 350116, *China*)

Abstract: A new method, based on belief rule base(BRB) was proposed for constructing a classification system with high performance for classification problems. The existing belief rule base classification system (BRBCS) is flawed because its classification accuracy is limited to the partition number, the parameter training method needs the number of rules given in advance, and the reasoning process does not reflect the correlation between characteristics and results. The belief rule base inference method was thus proposed for classification based on the differential evolutionary algorithm (DEBRM) to solve the classification problems. The proposed method consists of two procedures: belief rule base classification system (BRBCS) and parameter training method. The new method first introduced the construction

收稿日期:2016-03-01;**修回日期**:2016-09-17

基金项目:国家自然科学基金(70925004,61300026,71501047),福建省自然科学基金(2015J01248),福建省教育厅科技项目(JA13036),福州大学科技发展基金(2014-XQ-26)资助.

作者简介:刘莞玲,女,1991年生,硕士生.研究方向:数据挖掘.E-mail:380509981@qq.com

通讯作者:傅仰耿,博士/副教授.E-mail:ygfu@qq.com

strategy of BRBCS to determine the number of rules. Then, belief reasoning method was adopted as the inference engine. Finally, the training model for classification which is combined with differential evolutionary algorithm was built. In the experiment analysis, the effectiveness of the method was validated by comparing it with the existing parameter training method, and the rationality of parameters training in comparison of other belief rule base methods for different number of partitions. The classification results show that the proposed method is effective and reasonable.

Key words: belief reasoning method; classification system; parameters training; differential evolutionary algorithm

0　引言

随着大数据时代的到来，数据挖掘已经成为一个热门的研究领域. 从海量的数据中抽取隐含的、先前未知、有价值的信息具有重大现实意义. 数据分类是数据挖掘中极为重要的一部分，其主要是通过数据间的特征以带有已知类别的数据集为依据对未知类别的数据进行区分的过程[1]. 目前，数据分类的应用已涉及多个领域，包括机器学习、模式识别、统计学、神经网络、遗传算法、数据库、专家系统等. 并且，在应用过程中众多经典的分类方法相继被提出，诸如基于统计的分类方法、基于机器学习的分类方法以及神经网络分类方法. 此外，支持向量机法、粗糙集、模糊集、贝叶斯方法、K 近邻分类法、决策树、神经网络、遗传算法和最大熵模型都是常见的分类方法.

不同的分类方法在实际应用中，对不同的数据类型和应用领域具有不同的优缺点. 支持向量机法是 Cortes 等[2]于 1995 年提出的，SVM 可以最大化类与类的间隔，因而适应能力好、分类准确率高，但是该方法过度依赖于内积函数(核函数)的选择，有限样本下学习机器的复杂性与推广性之间经常存在矛盾. 粗糙集法是 Pawlak[3]首次提出的，粗糙集理论主要是通过知识约简，导出问题的决策或分类规则. 粗糙集分类方法在处理带有噪声、不精确甚至不完整的数据时具有明显的优势. 该方法不需要提供问题所需处理的数据集之外的任何先验信息，而且思路清晰便于操作. 虽然，引入粗糙集理论，可以对原始信息系统进行数据的处理以及属性约简、关联权重计算等，但这样做或多或少改变了原始信息. 因此该方法也并不是完美无瑕的. 模糊集(fuzzy sets)是由 Zadeh 等[4]在 1965 提出的，模糊集分类方法主要适用于对定义不严格的数据进行分类. 在自然学科或社会学科研究中，存在许多模糊性的概念，因此把模糊数学理论用于决策研究能获得更好的分类效果. 但是，基于模糊集理论的分类方法也存在一些缺陷，如何进行模糊划分以及规则数量是其难点. 贝叶斯分类方法早在 20 世纪 50 年代就已得到广泛研究[5]. 是一种基于统计学的分类方法，可以预测类成员关系的可能性. 它是一种有监督的学习方法，与其他分类方法相比较具有出错率较小的优点，但是朴素贝叶斯方法存在先验概率定义困难以及在实际问题中条件属性的独立假设一般不成立等缺陷，因此在解决实际问题中分类效果并不理想. K 近邻(K-nearest neighbors)分类法[6]是一种非参数的分类技术[7]，该方法简单有效，能适应类域分布较复杂的情况，但该方法在分类过程中相似量计算量过大导致分类速度慢，对样本库过于依赖、度量相似性的距离函数不适应以及 k 值选择不当等都会降低分类的效率和准确率. 神经网络分类器是 Guo 等[8]提出的，神经网络的优点是对噪声数据具有较强的适应能力，并且能较好地对未知数据进行预测分类，该方法在学习过程无需人为干预，预测结果具有叠加性. 该方法的缺陷是其过程不具备透明性，方法需要经过多次迭代较为复杂，容易出现早熟的现象. 其他分类方法也各有优劣，在此不一一做详细介绍. 总之，没有哪种分类方法是最优的，只能说针对特定分类问题，某些方法会更适合于解决该分类任务而已.

针对传统经典方法存在的不足，研究者提出了很多改进的方法. 其中模糊规则库分类系统(fuzzy rule-based classification systems, FRBCS)[9-11]已经成功地应用于解决分类问题. 但是基于模糊集理论的分类方法，模糊划分和规则数量是其解决问题的难点. Yang 等[12]提出了基于证据推理算法的置信规则库推理方法(belief rule-base inference methodology using the evidence reasoning approach, RIMER)，该方法更适用于复杂分类问题的建模. 它是以模糊理论、传统 IF-THEN 规则库、D-S 证据理

论和决策理论等为基础，能够有效地利用不完整或不精确的信息对复杂决策问题进行建模，可以更好地描述存在模糊、不精确或不完整信息的数据集. 之后 Lian 在置信函数框架上扩展模糊规则库分类系统(FRBCS)，提出置信规则库分类系统(belief rule base classification system，BRBCS)[13]. 置信规则库分类系统采用了数据驱动方式由数据集自动生成置信规则，建立特征空间与类空间之间的不确定关系，利用基于置信函数理论的置信推理方法(belief reasoning method，BRM)对查询模式进行分类推理. BRBCS 由于其推理性能依赖于内部参数取值，所以当模糊区间划分不同时，其推理性能也会发生改变. 在置信规则库的研究中，为提高 RIMER 的推理性能，Yang 等[14]提出 BRB 系统的参数训练方法. Chang 等[15]将参数训练方法应用到分类问题的求解上提出基于传统置信规则库参数学习的分类方法. Chang 的方法可以有效地解决推理性能强依赖于内部参数问题，但该方法仍存在不足，比如规则数量由人为给定，受主观因素制约；待优化参数不涉及特征权重，无法体现出不同特征对分类结果的影响.

针对上述三类基于规则库分类方法的不足，本文对现有的方法提出了改进，首先引入 BRBCS 的置信规则库构建策略，通过数据驱动的方式，根据数据集与模糊划分的关联，自动构建初始置信规则库，摒弃文献[15]人为给定规则数的方式. 接着使用置信推理作为推理机，结合群智能算法提出新的参数学习模型. 并将前提属性权重作为待优化参数，进行学习矫正，提高分类准确度. 实验中，使用 University of California at Irvine 中的 4 个公共数据集，以平均误差(mean error，ME)公式作为目标函数，进行分析验证，并与 Chang 等的方法进行对比. 实验结果表明，该方法的分类准确率更高，分类效果更好.

1 置信规则库的基础知识

1.1 置信规则的表示

置信规则库(belief rule base，BRB)中置信规则是由传统 IF-THEN[16] 规则扩展而来，其除了分布式置信框架外，还加入了规则权重和前提属性权重. 其中第 k 条置信规则表示如下：

$$R_k: \text{if } A_1^k \wedge A_2^k \wedge \cdots \wedge A_{T_k}^k \text{ then } \{(D_1, \bar{\beta}_{1,k}), (D_2, \bar{\beta}_{2,k}), \cdots, (D_N, \bar{\beta}_{N,k})\} \quad (1)$$

式中，$A_i^k (i=1,\cdots,T_k; k=1,\cdots,L)$ 表示第 k 条规则中第 i 个前提属性的候选值；L 表示 BRB 中的总规则数；T_k 表示第 k 条规则中包含前提属性的个数；$\bar{\beta}_{j,k}$ 表示第 k 条规则中第 j 个评价等级 D_j 上的置信度，当 $\sum_{j=1}^{N} \bar{\beta}_{j,k} = 1$ 时表示第 k 条规则包含完整的信息，当 $\sum_{j=1}^{N} \bar{\beta}_{j,k} < 1$ 表示该条规则包含的信息不完整.

此外，第 k 条规则中还包含前提属性权重 $\sigma_{i,k} (i=1,\cdots,T_k, k=1,\cdots,L)$ 和规则权重 θ_k 两类权重参数，这两个参数用于区分前提属性和规则的重要性，提升了置信规则表示不确定信息的能力，可以使得 RIMER 方法在解决实际问题时具有更高的准确性.

1.2 置信规则推理

RIMER 方法的推理过程主要包含三个步骤：激活权重的计算、置信度的修正和置信规则的激活.

1.2.1 计算激活权重以及修正置信度

激活权重的计算与 RIMER 方法的输入值、规则权重和前提属性权重相关. 激活权重的计算，首先需依据 RIMER 方法的输入值，假设第 i 个前提属性的输入值为 x_i，则可算得如下各个候选值的个体匹配度：

$$S(x_i) = \{(A_{i,j}, \alpha_{i,j}); i=1,\cdots,M; j=1,\cdots,J_i\} \quad (2)$$

式中，$\alpha_{i,j}$ 表示第 i 个前提属性中第 j 个候选值 $A_{i,j}$ 的个体匹配度；M 表示 BRB 中总的前提属性个数.

接着，计算第 k 条规则的激活权重：

$$\left.\begin{aligned} \omega_k &= \frac{\theta_k \prod_{i=1}^{T_k} (\alpha_i^k)^{\bar{\delta}_{k,i}}}{\sum_{l=1}^{L} \left(\theta_l \prod_{i=1}^{T_k} (\alpha_i^l)^{\bar{\delta}_{l,i}}\right)} \\ \bar{\delta}_{k,i} &= \frac{\delta_{k,i}}{\max_{i=1,\cdots,T_k}\{\delta_{k,i}\}} \end{aligned}\right\} \quad (3)$$

式中，α_i^k 表示第 k 条规则中前提属性 x_i 的候选值对应的个体匹配度，当 $\alpha_i^k = 0$ 时，$\omega_k = 0$，此时第 k 条规则未被激活.

由于输入值可能不完整，因此还需对激活规则结果集的置信度进行修正，其中第 k 条规则的第 i 个评价等级上的置信度修正公式如下：

$$\beta_{i,k} = \bar{\beta}_{i,k} \frac{\sum_{t=1}^{T_k} \left(\tau(t,k) \sum_{j=1}^{J_t} \alpha_{t,j}\right)}{\sum_{t=1}^{T_k} \tau(t,k)},$$

$$\tau(t,k)=\begin{cases}1, & U_t\in R_k(t=1,\cdots,T_k)\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

1.2.2　合成激活规则

通过规则的激活权重可确定置信规则库中激活规则的数量，然后由证据推理(ER)[17]算法解析公式可将其合成.

当求得 BRB 中所有激活规则的激活权重及修正的置信度后，便可用证据推理算法中的解析公式一次合成所有激活规则，其中各评价等级的置信度合成公式如下：

$$\beta_n=\frac{u\Big[\prod_{k=1}^{k}\big(w_k\beta_{n,k}+1-w_k\sum_{i=1}^{N}\beta_{i,k}\big)-\prod_{k-1}^{k}\big(1-w_k\sum_{i=1}^{N}\beta_{i,k}\big)\Big]}{1-u\Big[\prod_{k=1}^{k}(1-w_k)\Big]} \tag{5}$$

式中，

$$u^{-1}=\sum_{j=1}^{N}\prod_{k=1}^{k}\big(w_k\beta_{n,k}+1-w_k\sum_{i=1}^{N}\beta_{i,k}\big)-(K-1)\prod_{k=1}^{k}\big(1-w_k\sum_{i=1}^{N}\beta_{i,k}\big) \tag{6}$$

推理的输出是合成所有激活规则的分布式置信结构：

$$S(f(x))=\{(Dn,\beta_n(x)),n=1,\cdots,N\} \tag{7}$$

为了让 BRB 系统的输入更加直观，假设结果属性中第 i 个评价等级 D_i 的等级效用值为 uD_n，则可进一步推理得到 BRB 系统的数值型输出为：

$$f(x)=\sum_{n=1}^{N}(\mu(D_n)\beta_n(x)) \tag{8}$$

2　基于规则的分类方法

2.1　模糊规则的分类方法

模糊集合(fuzzy set)理论由 Zadeh 等[4,18]于 1965 年提出. 模糊分类是模糊集合理论的一个重要应用，基于模糊规则的分类系统(fuzzy rule-based classification systems，FRBCS)[9-11]能有效地从专家知识等定量和定性信息中构建语义模型. 为提高分类性能，Nozaki 等[18]提出自适应的模糊规则库分类系统. 自适应模糊规则分类系统可以通过学习自动产生其模糊规则，迭代地调整 CF 取值，并基于遗传算法约减模糊规则.

模糊规则与人类表达知识相似，符合人类思维逻辑，可读性及可解释性较强. Ishibuchi 等[10]使用格状划分均分获得模糊子区间，进而依次对应地构造模糊分类规则，表示如下：

$$\begin{aligned}&\text{Rule } R_{ij}^{K}: if\ x_{p1}\ is\ A_i^K\ \wedge\ x_{p2}\ is\ A_j^K\\ &\text{then } x_p \text{ belongs to } C_{ij}^K \text{ with } CF=CF_{ij}^K\\ &i=1,2,\cdots,K;j=1,2,\cdots,K;\end{aligned} \tag{9}$$

式中，K 表示模式空间子集划分数，R_{ij}^K 表示规则标签，A_i^K 和 A_j^K 是单位区间[0,1]的模糊子集. C_{ij}^K 表示分类结果，CF 表示规则权重.

当未知类别的查询模式到来时，使用模糊推理方法(fuzzy reasoning method，FRM)在规则集合中对其进行分类. 首先在不同的模糊子区间内，计算查询模式的隶属度；接着根据隶属度和区间规则权重相乘计算得到取值最大的规则；最后得到规则对应的类标即为查询模式的分类结果.

2.2　基于置信规则库的分类方法

Jiao 等在置信函数框架上扩展 FRBCS，提出基于数据驱动规则构建的 BRBCS 方法. 前提属性表现形式两者相类似，结果部分扩展为置信分布形式，并新增不属于任一基本类别的全局模糊评价等级. 引入特征权重表示不同特征与分类结果的关联度，最后使用置信推理方法作为分类查询方法. 置信分类规则表示如下：

$$\begin{aligned}&\text{Belief Rule } R^{q}: if\ X1\ is\ A_1^q\ \wedge\ X2\ is\ A_2^q\ \wedge\cdots\wedge\ X_P\ is\ A_P^q\\ &\text{then the } C^{q}=\{(\omega_1,\beta_1^q),(\omega_2,\beta_2^q),\cdots,(\omega_M,\beta_M^q),(\omega_\Omega,\beta_\Omega^q)\}\\ &\text{With a rule weight } \theta^q \text{ and feature weights } \delta_1,\delta_2,\cdots,\delta_P\end{aligned} \tag{10}$$

式中，$\theta^q,q=1,2,\cdots,Q$ 表示规则权重，$\delta_1,\cdots,\delta_p$，是特征权重. 结果置信度 β，表示当输入激活前提属性候选值集合时的置信度，体现倾向于各类别的程度.

在置信分类规则中，规则权重用于评估规则分类能力，由置信度和支持度相互制约决定. 当置信度较小，而支持度较大时，规则冲突较大，应当缩减规则权重；当支持度较小，而置信度较大时，结果易受噪音数据影响，同样应当缩减规则权重. 则规则权重可描述成置信度与支持度的乘积形式：

$$\theta^q \propto c(R^q) \cdot s(R^q) \tag{11}$$

实际问题中，不同特征或对分类结果产生相异影响. 为体现其关联度，于是引入特征权重.

Jiao 等首先根据每个特征的划分区间将 BRB 划分成多个子 BRB，分别将子 BRB 结果部分合成；再使用 Jousselme's 距离[19]计算相邻子 BRB 合成结果的差异，并累加；最后计算平均差异，进行标准化处理，即可得到每个特征的特征权重.

2.3 基于置信规则库参数学习的分类方法

BRB 推理性能主要依赖于内部参数的取值，当实际问题较为复杂时，专家给定参数的方式受限于专家知识等主观因素. 基于利用输入和系统输出矫正 BRB 内部参数，从而提高推理性能的思想，Yang 等[14]提出 BRB 系统的参数训练方法. Chang 在求解分类问题时，结合群智能算法和证据推理方法构建参数学习模型，提出基于置信规则库参数学习的分类问题求解方法，对基于 BRB 的分类系统内部参数进行优化，待优化的参数主要包括前提属性候选值、规则权重以及结果置信度. Chang 的方法解决了 BRBCS 存在数据依赖的问题的，在 4 个公共测试数据集上进行验证，均取得不错的分类效果；但其方法需人为给定规则数量，存在主观局限性，当实际问题比较复杂时，无法确定合适的规则数；并且待优化参数不包括前提属性权重，推理过程中各前提属性权重并无差异，不同特征对分类结果的影响无法区分.

本文提出了基于差分进化算法的置信规则库推理分类方法. 首先引入 BRBCS 的置信规则库构建策略，通过数据驱动的方式，根据数据集与模糊划分的关联，自动构建初始置信规则库，摒弃人为给定规则数的方式；接着将置信推理作为推理机，结合群智能算法提出新的参数学习模型，并将前提属性权重作为待优化参数，进行学习矫正，提高分类准确度. 由于本文方法是基于参数训练模型，因此无需通过增加划分数来提高分类的准确性.

3 基于 DE 的置信规则库推理分类方法

针对现有基于规则库分类方法中存在的不足，本文提出了基于差异进化（differential evolutionary，DE）的置信规则库推理分类方法，主要包括：基于 DE 的参数训练和置信推理过程，该方法步骤流程如图 1 所示.

3.1 基于 DE 的参数训练

3.1.1 数据驱动方式构建待训练置信规则库

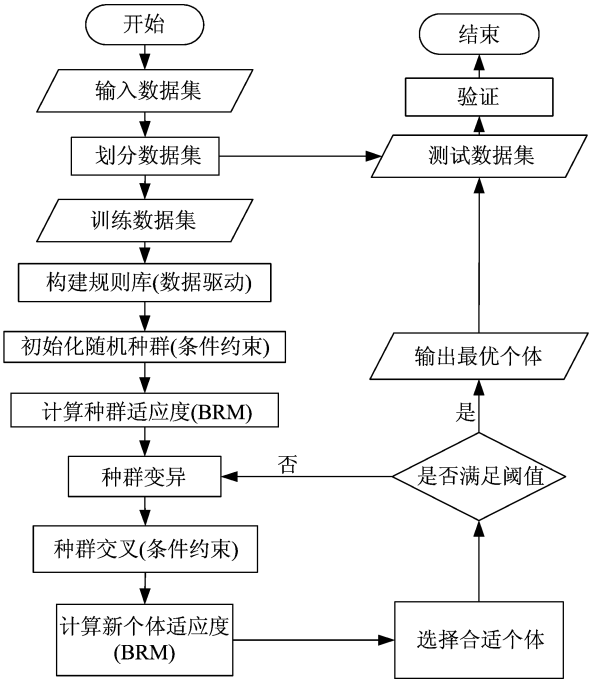

图 1 基于差分进化算法的置信规则库推理分类方法

Fig. 1 Process of classification for DEBRM

针对 Chang 方法中人为给定规则数存在的主观局限性，本文采用数据驱动的方式进行置信规则库构建. 在 BRM 方法中，当查询模式到来时，只有关联度与规则权重的乘积不为 0 时，规则才能被激活. 由此可见，在构建规则时，可删去置信度或支持度为 0 的规则. 这样的构建方式可以对规则进行约简，避免了 Chang 考虑的组合爆炸问题的出现.

用于实验分析的数据集，先随机划分为训练集和测试集，使用训练集构建规则库. 由于训练样本有限，可能出现查询数据不激活规则的情况，那么激活距离它所在模糊子区间最近的规则.

3.1.2 基于 DE 优化置信规则库

差分进化算法是一种全局启发式搜索算法，具有收敛速度快、强鲁棒性以及受控参数少的特点，能够有效解决复杂优化问题[6-7]. 本文基于 DE 提出新的参数训练分类的算法步骤如下所示：

步骤 1 将数据集分为训练数据集与测试数据集，根据划分数均分模糊区间. 输入训练数据集，根据其与模糊子区间关联度，确定规则库轮廓，构建初始置信规则库.

步骤 2 依照文献[14]中参数训练模型，利用 DE 对特征权重、规则权重、结果置信度进行参数训练. 训练步骤如下：

①初始化种群. 严格遵循等式及不等式约束条

件来初始化种群中的每一个个体，个体间相互独立，包含所有待优化参数，并设定种群大小、进化代数等 DE 算法参数. 接着使用 BRM 计算适应性.

②变异操作. 随机抽取种群中的个体，通过差分方式对个体进行变异处理，产生中间变异个体，变异公式如下：

$$v_i = x_{r1} + F \cdot (x_{r2} - x_{r3}) \tag{12}$$

式中，$r1,r2,r3$ 随机生成，且满足 $r1 \neq r2 \neq r3 \neq i$，$F$ 是常量，一般取值 0.5.

③交叉操作. 根据交叉概率，在中间变异个体与当前个体之间进行交叉操作，产生中间交叉个体，用 BRM 计算中间交叉个体的适应性. 交叉公式如下：

$$u_i(j) = \begin{cases} v_i(j), & \text{if rand} \leqslant \text{CR} \\ x_i(j), & \text{otherwise} \end{cases} \tag{13}$$

④选择操作. 对比当前个体与中间交叉个体的适应性，即分类的准确性，贪婪地选择更优个体作为下一代个体，选择公式如下：

$$x_i' = \begin{cases} u_i, & \text{if } f(u_i) \leqslant f(x_i) \\ x_i, & \text{otherwise} \end{cases} \tag{14}$$

本文选择指标为平均误差 ME 公式为如下：

$$\text{ME} = \frac{1}{T}\sum_{t=1}^{T} E(t) \tag{15}$$

$$E = \begin{cases} 1 & \text{if } \hat{n} \neq n \\ 0 & \text{if } \hat{n} = n \end{cases} \tag{16}$$

式中，n 为测试数据集中的真实类别，$\hat{n}$ 为推理分类结果.

当种群进化代数未达到预先设定的进化代数或者分类准确率阈值时，继续执行训练步骤②、③、④；否则，结束训练算法.

步骤 3　输入测试数据集，对差分进化算法学习后生成的置信规则分类系统进行验证.

3.2　置信推理方法(BRM)

置信规则库分类系统规则激活方式与 RIMER 规则激活方式存在差异，前者通过隶属函数映射输入数据隶属度来确定激活规则；并且在证据推理合成方法中，不同规则的权重表示决策者的主观偏好，而在分类问题中，规则权重应当视作规则正确分类能力的可靠性度量. BRM 推理过程取最大值，这种方式分类的有效性易受到噪音数据的影响，而 Lian 提出的 BRM，通过信息融合的方式，对激活规则进行合成，可克服这一缺陷.

本文使用 BRM 作为推理方法. 其主要思想是：当一个查询模式需要分类时，首先计算查询模式与激活置信规则结果的关联度，然后基于置信函数理论通过关联度来合成结果置信度.

假设 $x_i = (x_{i1},\cdots,x_{ip}), i = 1,2,\cdots,N$ 为一个 p 维的训练模式. 对于每个训练模式 x_i：通过乘积运算操作计算不同模糊域的匹配度 $\mu(x_i)$：

$$\mu_{A^q}(x_i) = \prod_{p=1}^{P} \mu_{A_p^q}(x_{ip}) \tag{17}$$

式中，$\mu_{A_p^q}(\cdot)$ 是模糊集 A_p^q 的隶属度函数. 通过取最大的隶属度来分配查询模式 x_i 模糊域.

本文使用对称三角隶属函数计算隶属度，假设每一个模式空间都均匀划分为 K 个模糊子集 A_p^1，$A_p^2,\cdots,A_p^k$，则 A_p^j 的表示为：

$$\mu_{A_p^j}(x_{ip}) = \max\left(1 - \frac{|x_{ip} - a_j^K|}{b^K}, 0\right), \quad j = 1,2,\cdots,K \tag{18}$$

式中，$a_j^K = \dfrac{j-1}{K-1}$，$b^K = \dfrac{1}{K-1}$.

3.2.1　置信规则结果关联度计算

首先定义：$y = (y_1,\cdots,y_p)$ 是一个分类查询模式，在不同模糊区间内计算得到查询模式的各前提属性的隶属度，再利用简单的加权乘法聚合函数来计算得到总隶属度：

$$\mu_{A^q}(y) = \left[\prod_{p=1}^{P} \mu_{A_p^q}(y^p)^{\delta_p}\right] 1/P \tag{19}$$

式中，δ^p 是第 p 个前提特征的属性权重.

隶属度反映了查询模式的前件与置信规则前件的相似性，规则权重表示了规则的可靠性，由此可见，查询模式与激活规则的关联度和这两个因素密切相关. 当匹配度不等于 0 时，该模糊区间对应的规则被激活，因此关联度的计算公式定义为：

$$\alpha^q = \mu_{A^q}(y) \cdot \theta^q, R^q \in S' \tag{20}$$

3.2.2　基于置信函数理论的推理

置信规则被激活后，在进行合成之前需对不可靠信息进行折扣处理，将上一步计算得到的置信规则结果关联度 α 作为可靠性因子. 在置信函数理论中，通常使用 Shafer 的折扣定律公式对所有激活规则进行处理：

$$\left.\begin{aligned} m^{\alpha}(\{\omega_m\}) &= \alpha \cdot \beta_m, m = 1,2,\cdots,M \\ m^{\alpha}(\Omega) &= \alpha \cdot \beta_{\Omega} + (1-\alpha) \end{aligned}\right\} \tag{21}$$

接着，使用递归的证据合成算法迭代合成结果，作为决策基础：

$$m_{I(i+1)}(\{\omega_q\}) = K_{I(i+1)}[m_{I(i)}(\{\omega_q\}) \cdot m^{\alpha}_{i+1}(\{\omega_q\}) + m_{I(i)}(\Omega) \cdot m^{\alpha}_{i+1}(\{\omega_q\}) + m_{I(i)}(\{\omega_q\}) \cdot m^{\alpha}_{i+1}(\Omega)]$$
$$m_{I(i+1)}(\Omega) = K_{I(i+1)}[m_{I(i)}(\Omega) \cdot m^{\alpha}_{i+1}(\Omega)] \tag{22}$$

归一化因子：

$$K_{I(i+1)} = \left[1 - \sum_{j=1}^{M} \sum_{p=1, p\neq j}^{M} m_{I(i)}(\{\omega_j\}) \cdot m^{\alpha}_{i+1}(\{\omega_p\})\right]^{-1} \quad i = 1,2,\cdots,L-1 \tag{23}$$

因此，

$$\sum_{q=1}^{M} m_{I(i+1)}(\{\omega_q\}) + mI(i+1)(\Omega) = 1 \tag{24}$$

式中，$m_{I(1)}(\{\omega_q\}) = m^q_1(\{\omega^q\}), q = 1,2,\cdots,M$，$m_{I(1)}(\Omega) = m^q_1(\Omega)$.

最后，基于合成的 BBA 使用置信函数 Bel(·) 进行决策，对于每个分类 ω_q，置信函数计算公式如下：

$$\mathrm{Bel}(\{\omega_q\}) = m(\{\omega_q\}) \tag{25}$$

则查询模式的最终分类结果可由如下公式得到

$$\omega = \arg\max_{\omega q \in \Omega} \mathrm{Bel}(\{\omega_q\}) \tag{26}$$

3.3 参数的约束条件与优化

BRBCS 基于置信规则库，影响其分类准确性的参数主要包括结果置信度、规则权重及特征权重. Lian 通过一系列公式计算得到这些参数的取值，但是当区间划分数发生改变时，参数取值也会发生改变，这必将影响推理的准确性. 人为给定的方式，无法确定合适的划分数.

本文使用差分进化算法进行求解，由于算法具有随机性，使其不受初始解的影响，由此可设置较小区间划分数，将这些内部参数列为待优化参数，在训练过程中进行优化，使其满足特定约束条件，即可取得较好效果. 其中约束条件如下说明：

(Ⅰ)假设第 q 条规则是完整的，则该条规则的结果置信度之和等于 1：

$$\sum_{j=1}^{\Omega} \beta_j = 1 \tag{27}$$

(Ⅱ)任意一条置信规则的结果置信度均不小于 0 或不大于 1，其中第 j 个分类结果上的置信度需满足：

$$0 \leqslant \beta_j \leqslant 1,\ j = 1,2,\cdots,M,Q \tag{28}$$

(Ⅲ)规则权重 θ^q，使其不小于 0 或不大于 1，即：

$$0 \leqslant \theta^q \leqslant 1,\ q = 1,2,\cdots,Q \tag{29}$$

式中，Q 表示 BRB 中规则的数量.

(Ⅳ)特征权重 δ_i，使其不小于 0 或不大于 1，即

$$0 \leqslant \delta_i \leqslant 1,\ i = 1,2,\cdots,p \tag{30}$$

式中，p 表示 BRB 中特征的数量.

4 实验分析

本节中将对参数训练分类方法的有效性进行验证分析，分别将本文方法(DEBRM)与 Chang 的方法以及 BRBCS 进行对比. 同时通过各数据集多组实验的分类准确率对比验证本文方法的鲁棒性. 实验中方法均在 VC++6.0 中运行. 由于差分进化算法存在随机特性，实验中采取独立运行 30 次获取实验结果. 此外，实验环境的基本信息为：Intel(R) Core(TM) i3-3210 CPU @ 3.20GHz 处理器、4GB 内存，Windows 8 操作系统.

4.1 数据集及 DE 参数设置

对于实验中所用的数据集，本文选自 University of California at Irvine 分校网页中获取的公共测试集. 数据集主要包括鸢尾花特征数据 Iris、红酒化学成分特征数据 Wine、乳腺癌数据 Cancer 和玻璃类型数据 Glass. 其中，表 1 列举了上述 4 个测试数据集中前件属性数量、类别数量和数据集大小的信息.

表 1 测试数据集的基本信息

Tab. 1 Base information of test data set

数据集	前件属性数量	类别数量	数据量
Iris	4	3	150
Cancr	9	2	683
Wine	13	3	178
Glass	9	6	214

表 1 中，Cancer 原数据集中存在缺失数据，实验中已将缺失部分的 16 条数据删除，保留 683 条数据.

4.2 实验准确度对比

实验中，为减少训练数据与参数学习方法的关联性，随机选择 80% 的数据作为训练数据，剩余 20%作为测试数据；差分进化算法种群规模大小设置为 100，迭代次数为 3 000.

为了验证本文方法的有效性，本文设计了该方法和 Chang 方法以及该方法和 BRBCS 分类方法的分类准确度对比实验.

4.2.1　DEBRM 与 Chang 方法进行对比

置信规则库推理方法是白盒模型，透明性优于神经网络，相比模糊规则能够更有效地处理模糊、不精确或不完全信息. 针对使用基于置信规则库的专家系统建模求解分类问题，需要离散化前提属性、对初始解敏感以及规则库可能出现组合爆炸的难点，Chang 提出了置信规则库参数学习方法. 主要思想是结合差分进化算法和证据推理方法，对置信规则库内部参数进行优化，其中待优化参数包括前提属性候选值、规则权重以及结果置信度，并未涉及前提属性权重，即本文中提及的特征权重. 在构建规则库方面，Chang 提前设定规则数量，Iris 为 3 条，另外 3 个数据集为 4 条，这种人为给定的方式，存在主观局限性，复杂决策问题无法给定合适的规则数量.

本文方法求解分类问题的模型与 Chang 一致，同为参数训练方法. 基于数据驱动的构建策略，根据训练数据集与模糊区间的关联度，构建置信规则库. 由于本文方法与 Chang 使用的置信规则库激活方式不同，推理机制方面使用置信推理方法. 在训练过程中，拓宽参与训练的内部参数，加入特征权重，体现不同特征对分类结果的影响. 表 2 中给出了 Chang 方法与本文方法关于求解 Iris、Wine 和 Glass 三组数据分类问题的准确性对比.

表 2　本文方法与 Chang 方法分类准确性对比

Tab. 2　Classification accuracy of DEBRM and Chang

方法		数据集	
		DEBRM	Chang
Iris	准确度	99.33%	100%
	失误个数	1	0
Wine	准确度	99.44%	96.63%
	失误个数	1	6
Glass	准确度	80.84%	69.16%
	失误个数	27	66

如表 2 所示，本文方法应用在 Iris 数据集中，分类失误个数比 Chang 方法多 1 个，Wine 数据集中失误个数仅为 1 个，而 Glass 数据集中失误个数减少了 39 个.

Chang 方法在确定规则数时，需要多次进行实验验证，才能得到合适的取值，而本文基于数据驱动的构建方式，只需设置一个较小的划分数，即可由随机的训练数据自动生成规则，同时规则数不会超过数据量，避免组合爆炸的问题. 置信推理方法的过程涉及特征权重，经过参数训练后，合理的取值可以提高关键属性对分类结果的影响. 实验结果表明，本文方法相对于 Chang 的方法能够有效降低失误个数，得到良好准确性.

4.2.2　DEBRM 与 BRBCS 分类方法进行对比

置信规则库分类系统由模糊规则库分类系统扩展而来，在规则的前提属性部分两者相似，而结果部分更改为置信分布形式. 由于置信规则库分类系统的规则构建是数据驱动的方式，其轮廓取决于模糊子区间划分以及数据集，所以不同的划分方式将影响分类结果的准确性. 一般情况下，规则越多，能够得到更准确的结果. 本文方法对置信规则库分类系统内部参数进行训练优化，因此不需要依赖增加划分数量提高分类准确度. 表 3-5 中对比了本文方法与 BRBCS 分类方法解决 Iris 数据集、Cancer 数据集和 Glass 数据集分类问题的准确性. 实验中针对每组数据集，我们随机把 80% 的数据作为训练集，另外 20%作为测试数据.

表 3　Iris 数据集分类准确性对比

Tab. 3　Classification accuracy of Iris

Iris 数据集				
	DEBRM	BRBCS		
划分数	3	3	5	7
训练集	99.17%	94.17%	95.00%	96.67%
测试集	96.67%	93.67%	96.33%	96.67%

表 4　Cancer 数据集分类准确性对比

Tab. 4　Classification accuracy of Cancer

Cancer 数据集				
	DEBRM	BRBCS		
划分数	3	3	5	7
训练集	100%	97.44%	99.82%	98.35%
测试集	98.54%	95.82%	96.74%	92.47%

表 5　Glass 数据集分类准确性对比

Tab. 5　Classification accuracy of Glass

Glass 数据集				
	DEBRM	BRBCS		
划分数	3	3	5	7
训练集	83.04%	83.14%	94.77%	90.12%
测试集	72.09%	69.04%	71.84%	64.29%

由上表可见，BRBCS 在划分数量发生变化时，

准确度随之发生改变，但并非呈现严格的正相关. 如 Iris 数据集，准确度随着划分数增加而递增；Cancer 和 Glass 数据集，在划分数由 3 增加到 5 时，准确度有一定提升，而由 5 增加到 7 时，准确度有下降趋势，因此不能盲目追求准确度而增加划分数量. 本文方法将划分数量设置为 3，对置信分类规则库待优化参数进行训练优化. 在 Iris 和 Cancer 数据集中，训练后分类准确度高于 BRBCS 的最优准确度；在 Glass 的训练数据集中，分类准确度只能达到 83.04%，略显不足，不过在测试数据集中，能够达到不错效果，同样具备良好的预测能力.

通过分析对比，BRBCS 的分类准确性与模糊子区间的划分数量并非严格的正相关，数量的增加反而可能导致准确性下降，同时加重了计算负担. 而划分数量的确定，同样是人为给定，依旧存在主观局限性. 本文为与其对比，挑选其实验中的最小划分数作为本文方法的划分数. 通过参数训练，得到了推理性能更优的规则库. 可见本文方法是有效可行的.

4.3 与其他方法进行对比

为了进一步验证本文方法求解分类问题的有效性. 本节将对比使用本文提出算法求解分类问题与现有文献的分类结果. 表 6 列出了与 Fallahnezhad 等[20]总结的部分前人成果.

表 6 不同方法的最优分类准确率对比

Tab. 6 Comparison of accuracy with different methods

分类方法	数据集			
	Iris	Cancer	Wine	Glass
Naïve Bayes(%)	96.30	96.10	99.07	44.60
C4.5(%)	95.33	94.80	96.26	68.40
AMO(%)	99.27	98.48	99.98	65.43
Fuzzygain measure (%)	99.28	99.04	99.62	73.83
Fallahnezhad (%)	99.77	98.34	100	64.08
BRB-based(%)-testing	100	98.92	97.73	71.70
BRB-based(%)-total	100	98.77	97.19	68.24
DEBRM(%)-testing	100	99.82	99.44	83.04
DEBRM(%)-total	100	99.56	99.44	80.84

根据表 6 可知，基于差分进化算法的置信规则库推理分类方法针对 4 个测试数据集的分类准确性都较为理想. 其中 Iris 取得了 100%的分类准确率，在 Glass 上取得了最高精度，在 Cancer 和 Wine 中也取得较优精度的结果.

由于置信规则库的参数学习方法以差分进化算法作为参数训练方法，因此还需要进一步对比 30 次运行实验的统计数据以验证该方法的稳定性. 表 7 列举了针对各测试数据集 30 次运行实验的统计数据.

表 7 使用基于差分进化算法的置信规则库推理分类方法的统计结果

Tab. 7 DEBRM Classification accuracy

精度	数据集			
	iris	cancer	wine	glass
Max(%)	100	97.99	100	76.33
Min(%)	97.50	96.52	99.29	69.23
Avg(%)	98.75	97.25	99.65	72.78
Vara	1.56E-04	5.4E-05	1.26E-05	1.26E-03

为了更加直观地展现各测试数据集 30 次运行试验的结果，绘制如图 2.

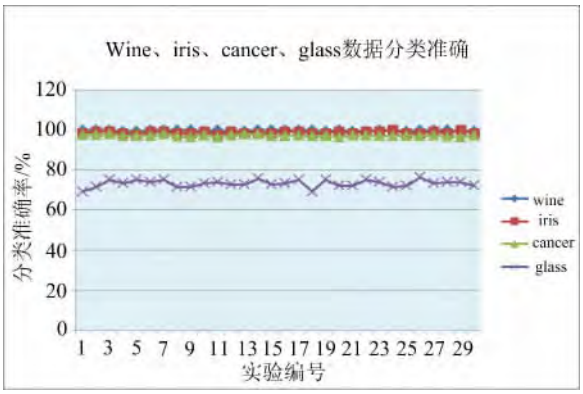

图 2 iris、wine、cancer、glass 数据集分类准确率

Fig. 2 Classification accuracy of iris、wine、cancer and glass

通过对比可知，所有示例中的精度的均值和最大值/最小值都十分接近，且精度的方差都远小于其均值，说明该方法的有效性. 同时通过各数据集分类准确率折线图也能直观地说明该分类方法可以得到稳定的解，具有较强的鲁棒性.

5 结论

针对现有基于置信规则库参数学习的分类问题求解方法中存在的不足，本文引入置信规则库分类系统数据驱动方式的构建策略，同时将前提属性权重作为待优化参数使用差分进化算法进行训练，体现了不同特征的重要程度. 并在此基础上将更适合分类问题的置信推理方法作为推理机对查询模式进行分类推理. 实验分析中，本文方法分析对比了基于

置信规则库参数学习的分类问题求解方法和置信规则库分类系统的分类性能，实验结果表明，本文方法是可行有效的. 在今后的研究工作中，将对合理设置阈值、规则约简以及多目标问题求解等问题作进一步的研究，以期提出分类性能良好且更合理的分类系统.

参考文献(References)

[1] 崔彩霞. 智能分类方法[M]. 北京：气象出版社，2009：1-216.

[2] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.

[3] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[4] ZADEH L A. Fuzzy sets [J]. Information and Control, 1965, 8(3): 338-353.

[5] RISHI. An empirical study of the naive Bayes classifier [J]. Journal of Universal Computer Science, 2001, 1(2): 41-46.

[6] COVER T M, HART P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[7] DASARATHY B V. Nearest neighbor (NN) norms [A]// NN Pattern Classification Techniques. IEEE Computer Society, 1991, 9: 1-30.

[8] GUO N R, LI T H S. Construction of a neuron-fuzzy classification model based on feature-extraction approach [J]. Expert Systems with Applications, 2011, 38(1): 682-691.

[9] CHI Z, YAN H, PHAM T. Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition [M]. Singapore: World Scientific Publishing, 2014.

[10] ISHIBUCHI H, NOZAKI K, TANAKA H. Distributed representation of fuzzy rules and its application to pattern classification [J]. Fuzzy Sets and Systems, 1992, 52(1): 21-32.

[11] SUN C T. Rule-base structure identification in an adaptive-network based fuzzy inference system [J]. IEEE Transactions on Fuzzy Systems, 1994, 2(1): 64-73.

[12] 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, 2006, 36(2): 266-285.

[13] 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.

[14] YANG J B, LIU J, XU D L, et al. Optimization models for training belief-rule-based systems [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2007, 37(4): 569-585.

[15] CHANG L L, ZHOU Z H, YOU Y, et al. Belief rule based expert system for classification problems [J]. Information Sciences, 2015, 336: 75-91.

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

[17] WANG Y M, YANG J B, XU D L. Environmental impact assessment using the evidential reasoning approach [J]. European Journal of Operational Research, 2006, 174(3): 1885-1913.

[18] NOZAKI K, ISHIBUCHI H, TANAKA H. Adaptive fuzzy rule-based classification systems [J]. IEEE Transactions on Fuzzy Systems, 1996, 4(3): 238-250.

[19] JOUSSELME A L, GRENIER D, BOSSÉ É. A new distance between two bodies of evidence [J]. Information Fusion, 2001, 2(2): 91-101.

[20] FALLAHNEZHAD M, MORADI M H, ZAFERANLOUEI S. A hybrid higher order neural classifier for handling classification problems [J]. Expert Systems with Applications, 2011, 38(1): 386-393.