ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 3,May 2018,pp. 477－486
DOI:10. 16337/j. 1004-9037. 2018. 03. 011

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

基于置信规则库推理的二择众仓分类方法

方志坚　傅仰耿　陈建华

(福州大学数学与计算机科学学院,福州, 350116)

摘　要:针对线性组合方式所构建的置信规则库存在常常无法准确发挥前件属性权重的效能,且随着评价等级个数的增加,新激活权重公式往往会对结果造成不利影响的不足,本文在现有置信规则库推理分类算法的基础上,提出二择众仓决策法,以此改进置信规则库决策系统。首先仅设置两个规则的后件评价等级,对一个决策问题仅做出二择判定,即回答是与否;其次,设置多个置信规则库同时处理若干个子问题;最后通过众仓决策方式融合多个子问题的结果,进而解决最终的分类问题。实验结果表明,改进后的置信规则库推理分类方法可行有效。

关键词:置信规则库;分类;二择众仓;证据推理;投票

中图分类号:TP18　　**文献标志码**:A

Two-Value Judgment Classification Approach Based on Belief Rule-Base Reasoning

Fang Zhijian, Fu Yanggeng, Chen Jianhua

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou, 350116, China)

Abstract: The weight of antecedent attributes can′t work accurately in the linear combinational belief rule based system usually. Simultaneously, with an increase in the number of evaluation ranks, the new weight activation formula will have negative effects on results. Aiming at the above drawbacks, this paper proposes a two-value and multi-base reasoning method based on the existing belief rule based inference classification algorithm to improve the belief rule based decision system. The evaluation of belief rules in the conclusion is divided into two ranks firstly, which means making a two-value judgment on a decision problem. Then many belief rule bases are set to solve some sub problems simultaneously. Finally results of many sub problems by multi-base reasoning method are mixed to solve the classification problem. Experimental results show the feasibility and effectiveness of the proposed belief rule base reasoning classfication method.

Key words: belief rule base; classification; two-value judgment; evidential reasoning; voting

引　　言

随着互联网络及媒体设备的高速发展,各个行业领域每时每刻均在产生大量的数据。数据信息量的激增导致了海量数据库的产生,如何在海量信息源中提取隐藏和有价值的信息,并应用这些信息构建

基金项目:国家自然科学基金(71501047)资助项目;福建省自然科学基金(2015J1248)资助项目;福州大学科技发展基金(2014-XQ-26)资助项目。
收稿日期:2016-09-06;**修订日期**:2016-10-14

决策支持的模型一直在商务管理、生产控制和市场分析等领域有着强烈的需求。因此数据挖掘如今成为一个热门的研究领域。数据分类是数据挖掘领域中一个重要的分支,这是由于大多数的实际工程问题均能转换成分类问题。分类就是利用已知标签的数据来构建相关的模型,进而确定未知类别数据标签的决策过程。

目前已经有许多经典的分类算法被提出,例如:K 近邻[1]、支持向量机[2]和粗糙集[3]等方法。K 近邻算法原理相对简单,方法易于实现,并且支持增量学习,具有对复杂决策空间进行建模的能力,其缺点在于寻找近邻点需要大量的计算,且当数据中包含弱相关属性时,其分类精度会下降;支持向量机在已知核函数的情况下,能够大大减少对高维问题的求解复杂度,其在一定程度上具有泛化推广的能力,因此也导致了支持向量机的分类精度十分依赖核函数的选择,而如何选择核函数一直是一个公认的难题;粗糙集的优点在于不需要附加信息或先验知识,就能够处理存在不精确乃至不完整数据的问题。但它只能处理离散化的属性,且产生的决策规则不稳定,准确率有待提高。

置信规则库推理方法(Belief rule-base inference methodology using evidential reasoning approach, RIMER)最早由 Yang 等[4]提出,囊括了传统 IF-THEN 规则库[5]、D-S 证据理论[6,7]、决策理论[8]和模糊理论[9]等方面的知识,具有对不完整或不精确信息进行建模的能力。当前,以 RIMER 为核心的置信规则库(Belief rule base, BRB)系统已经广泛应用于输油管道检漏[10]、工程系统安全评估[11]和军事能力评估[12]等工程领域。传统的 RIMER 方法不适用于属性数量过多的问题,这是由于 BRB 在构建过程中需要遍历所有前件属性的各个候选值,因此随着属性数量的增多,BRB 的规则条数将呈指数级增长,这必然会导致"组合爆炸"问题的产生。鉴于此,前人提出了通过线性组合的方式构建规则库,使得规则库的条数不再随属性数量的增长而增长。由于规则库在结构上发生了改变,使得原有方法中的激活公式不再适用于现有方法。因此,Chang 等[13]将原有方法中的激活规则改为激活属性,即不再关注哪条规则被激活,而是注重前件属性的哪些候选值被激活。将 BRB 应用于分类算法中,Ye 等[14]提出设定规则数等于分类数,以输入值和候选值之间距离倒数的归一化值作为个体匹配度,即对于任何输入值,规则库中的每条规则都将被激活,激活权重表现为每条规则对分类结果的贡献度。这些方法不仅解决了"组合爆炸"问题,而且在分类准确性上获得了不错的效果。然而采用线性组合方式也存在两点不足:(1)线性组合迫使每个前件属性候选值的个数必须相等,这也就忽略了各个前件属性间的差异;(2)激活权重公式的改变导致后件置信度受到了来自非激活前件属性候选值的影响,即在原始 RIMER 方法中,参与证据推理(Evidertail reasoning, ER)合成的置信度所对应规则的候选值都存在个体匹配度,而采用线性组合的方式,常常出现个体匹配度为 0 的属性参与激活合成,这必然会影响后件置信度。

本文提出一种改进置信规则库推理的分类方法。首先,在文献[13]研究的基础上,将评价等级个数设定为两个,即后件评价等级个数不再等于分类数。对于一组输入值,RIMER 过程只得出该输入值是否满足某种特定条件,以及以多大的置信度满足该条件。其次,对于某一具体问题不再局限于设定一个规则库,而是同时存在多个规则库进行决策。依据数据的特性,采用 One-versus-One 和有向无环图来构建决策模型。最后,再根据各个规则库的决策结果进行类别"投票",票数最高的即为最终分类结果。

1 置信规则库推理方法

采用规则形式表示相关信息在人工智能领域是一种非常常见的方式,在置信规则库推理方法中,规则化的信息表示体现在了置信规则库中。

1.1 置信规则库表示

置信规则库由传统 IF-THEN 规则库演化而来,Yang 等将置信框架引入 IF-THEN 规则中,使得传统的 IF-THEN 规则能够合理地表示不完整或不确定信息的知识,从而提出新的规则表达式。新的规

则即称为置信规则,其中第 k 条置信规则可表示为

$$R_k:\text{if } X_1 \text{ is } A_1^k \wedge X_2 \text{ is } A_2^k \wedge \cdots \wedge X_{T_k} \text{ is } A_{T_k}^k$$
$$\text{then}\{(D_1,\bar{\beta}_{1,k}),(D_2,\bar{\beta}_{2,k}),\cdots,(D_N,\bar{\beta}_{N,k})\} \tag{1}$$

式中:$A_i^k(i=1,2,\cdots,T_k;k=1,2,\cdots,L)$是第 k 条规则中第 i 个前件属性的候选值;X_i 是第 i 个前件属性;T_k 表示第 k 条规则中前件属性的数量;L 表示置信规则库中规则的数量;$\bar{\beta}_{j,k}(j=1,2,\cdots,N)$表示第 k 条规则中第 j 个评价等级 D_j 上的置信度;N 表示评价等级的数量;当 $\sum_{j=1}^{N}\bar{\beta}_{j,k}=1$ 时,表明第 k 条规则所包含的信息是完整的,否则称第 k 条规则包含不完整的信息。每条置信规则还包括两个重要的参数,规则权重 θ_k 和前件属性权重 $\delta_{i,k}$,其中 θ_k 表示第 k 条规则的权重,$\delta_{i,k}$表示第 k 条规则中第 i 个前件属性的权重。

1.2 置信规则库系统推理过程

BRB 系统的规则推理由计算激活权重、修正后件置信度和合成激活规则 3 步组成。在计算激活权重前,需先计算各个前件属性候选值的个体匹配度,计算方法为

$$a_i^j=\begin{cases}\dfrac{A_i^{k+1}-x_i}{A_i^{k+1}-A_i^k} & j=k,A_i^k\leqslant x_i\leqslant A_i^{k+1}\\ \dfrac{x_i-A_i^k}{A_i^{k+1}-A_i^k} & j=k+1,A_i^k\leqslant x_i\leqslant A_i^{k+1}\\ 0 & \text{其他}\end{cases} \tag{2}$$

其中,输入值向量表示为 $\boldsymbol{x}=\{x_1,x_2,\cdots,x_T\}$。

第 k 条规则激活权重的计算公式为

$$\omega_k=\frac{\theta_k\prod_{i=1}^{T_k}(\alpha_i^k)^{\bar{\delta}_{i,k}}}{\sum_{l=1}^{L}\left(\theta_l\prod_{i=1}^{T_l}(\alpha_i^l)^{\bar{\delta}_{i,l}}\right)},\bar{\delta}_{i,k}=\frac{\delta_{i,k}}{\max\limits_{i=1,\cdots,T_k}\{\delta_{i,k}\}} \tag{3}$$

式中:a_i^k 表示第 k 条规则中第 i 个前件属性候选值的个体匹配度;θ_k 表示第 k 条规则的规则权重;$\delta_{i,k}$表示第 k 条规则的第 i 个前件属性权重。由于输入值可能不完整,需要对后件置信度作进一步的修正。其中,第 k 条规则的结果集中第 i 个评价等级上的置信度修正公式为

$$\beta_{i,k}=\bar{\beta}_{i,k}\frac{\sum_{t=1}^{T_k}\left(\tau(t,k)\sum_{j=1}^{J_t}\alpha_{t,j}\right)}{\sum_{t=1}^{T_k}\tau(t,k)}$$
$$\tau(t,k)=\begin{cases}1 & U_t\in R_k(t=1,\cdots,T_k)\\ 0 & \text{其他}\end{cases} \tag{4}$$

式中:U_t 表示第 k 条规则的第 t 个前件属性;R_k 表示第 k 条规则的前件属性集合。

由激活规则的后件置信度和激活权重可以求得基本属性的基本可信值为

$$m_{j,k}=\omega_k\beta_{j,k} \tag{5}$$

$$\widetilde{m}_{D,k}=\omega_k\left(1-\sum_{j=1}^{N}\beta_{j,k}\right) \tag{6}$$

$$\overline{m}_{D,k}=1-\omega_k \tag{7}$$

式中:$m_{j,k}$表示对于第 i 个等级的基本可信度分配值;$\widetilde{m}_{D,k}$表示由第 k 条规则评价结果的不完整性引起的基本可信度分配值;$\overline{m}_{D,k}$表示由第 k 条规则的激活权重引起未设置给第 i 个评价等级的基本可信值。

最后将激活的规则使用 ER 法则合成,得到相对于评价等级 D_j 的基本可信度分配值,即

$$C_j = k\Big[\prod_{l=1}^{L}(m_{j,l}+\bar{m}_{D,l}+\tilde{m}_{D,l}) - \prod_{l=1}^{L}(\bar{m}_{D,l}+\tilde{m}_{D,l})\Big] \tag{8}$$

$$\tilde{C}_D = k\Big[\prod_{l=1}^{L}(\bar{m}_{D,l}+\tilde{m}_{D,l}) - \prod_{l=1}^{L}\bar{m}_{D,l}\Big] \tag{9}$$

$$\bar{C}_D = k\prod_{l=1}^{L}\bar{m}_{D,l} \tag{10}$$

$$k^{-1} = \sum_{j=1}^{N}\prod_{l=1}^{L}(m_{j,l}+\bar{m}_{D,l}+\tilde{m}_{D,l}) - (N-1)\prod_{l=1}^{L}(\bar{m}_{D,l}+\tilde{m}_{D,l}) \tag{11}$$

$$\beta_j = \frac{C_j}{1-\bar{C}_D} \quad j=1,\cdots,N \tag{12}$$

$$\beta_D = \frac{\tilde{C}_D}{1-\bar{C}_D} \tag{13}$$

1.3 参数学习模型

专家根据历史信息和先验知识给定的初始 BRB 系统存在主观局限性，特别是当 BRB 系统应用于复杂决策问题时，人为方式难以精确地给出这些参数值。故 Yang 等[15]提出了 BRB 系统参数训练模型，通过比较观测输出和推导输出的差值来矫正 BRB 系统的参数，进而提高 RIMER 方法决策的准确性。训练模型如图 1 所示。参数优化模型可表示为

$$\min\{\Delta(\boldsymbol{P})\}$$
$$\text{s.t. } A(\boldsymbol{P})=0, B(\boldsymbol{P})\geqslant 0 \tag{14}$$

式中：$\boldsymbol{p}=(\beta_{i,k},\theta_k,\delta_{k,i})$为待训练的参数向量；$\Delta(\boldsymbol{P})$为目标函数，当 $\Delta(\boldsymbol{P})$越小时，表明该 BRB 系统更符合实际系统，故最小化 $\Delta(\boldsymbol{P})$是参数训练的最终目的；$A(\boldsymbol{P})$和 $B(\boldsymbol{P})$分别是等式和不等式约束条件。在参数训练过程中，文献[15]给出如下规定

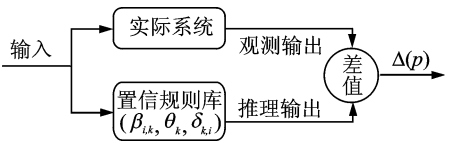

图 1 BRB 参数训练模型

Fig. 1 BRB parameter training model

(1) 标准化前件属性权重$\bar{\delta}_i$，使其不小于 0 且不大于 1，即

$$0\leqslant\bar{\delta}_i\leqslant 1 \qquad i=1,2,\cdots,M \tag{15}$$

(2) 标准化规则权重 θ_k，使其不小于 0 且不大于 1，即

$$0\leqslant\theta_k\leqslant 1 \qquad k=1,2,\cdots,L \tag{16}$$

(3) 任意一条置信规则的后件置信度均不小于 0 且不大于 1，其中第 k 条规则的第 j 个评价等级上的置信度需满足

$$0\leqslant\beta_{j,k}\leqslant 1 \qquad j=1,2,\cdots,N;k=1,2,\cdots,L \tag{17}$$

(4) 假设第 k 条规则是完整的，即输入不包含不确定或模糊信息，则该条规则的后件置信度之和等于 1，即

$$\sum_{j=1}^{N}\beta_{j,k}=1 \qquad k=1,2,\cdots,L \tag{18}$$

2 现有置信规则库推理的分类方法

目前，置信规则库推理方法应用于分类已经在淋巴结疾病诊断[16]和 UCI 分类数据集的测试[17]上取得了一定的进展。采用置信规则库推理方法解决分类问题主要有以下两种构建规则库策略。

(1)采用遍历前件属性候选值的方式，无需改变原始 BRB 系统，在处理分类问题时，只需要将后件置信度转换成分类结果信息，即可直接将 RIMER 方法应用到分类问题中。然而该方法只能应用于数据属性数目较少的情况，这是由于遍历组合方式继承了传统 RIMER 方法中固有的“组合爆炸”问题。

随着数据属性数目的增多,规则库的条数将呈指数级增长。以数据集 Wine 为例,其数据属性个数有 13 个,假设每个前件属性候选值的个数均为 3 个,那么 BRB 的规则条数就达到了 3^{13} 条,此时参数训练所耗费的时间是令人无法忍受的。通过遍历组合方式构建的规则库大小可表示为

$$\text{size}_{\text{BRB}} = \prod_{k=1}^{N} T_k \tag{19}$$

其中 T_k 表示第 k 个前件属性候选值的个数。通过对 UCI 上 210 组分类数据进行统计后发现,前件属性个数小于 10 的分类数据集个数仅为 54 组,而属性个数大于 10 的有 156 组,由此可知分类数据通常为多属性的情况。因此采用遍历组合方式构建规则库并不适用于大部分的分类问题。

(2)采用线性组合方式构建置信规则库中的规则。例如,有 3 个前件属性,每个前件属性的候选值依次为:{1,2,3},{4,5,6}和{7,8,9},则通过线性组合方式构建的置信规则库为

$$\begin{aligned} &R_1: \text{if } A_1 \text{ is } 1 \wedge A_2 \text{ is } 4 \wedge A_3 \text{ is } 7, \text{then}\{D\} \\ &R_2: \text{if } A_1 \text{ is } 2 \wedge A_2 \text{ is } 5 \wedge A_3 \text{ is } 8, \text{then}\{D\} \\ &R_3: \text{if } A_1 \text{ is } 3 \wedge A_2 \text{ is } 6 \wedge A_3 \text{ is } 9, \text{then}\{D\} \end{aligned} \tag{20}$$

可以看出,采用线性组合的方式,规则库中规则的条数只与前件属性候选值的个数有关,而与前件属性个数无关,这样就有效避免了"组合爆炸"问题的产生。然而,采用线性组合方式构建的置信规则库中常常会出现"零激活"问题,这是由于在计算个体匹配度时,至多仅有两个候选值的个体匹配度非零,其余的皆为零。根据式(3)可知,只要规则中存在某个前件属性候选值的个体匹配度为零,那么该条规则的激活权重就为零,即不被激活。因此,激活权重公式可修正为

$$\omega_k = \frac{\theta_k \sum_{i=1}^{T_k} (\alpha_i^k)^{\bar{\delta}_{i,k}}}{\sum_{l=1}^{L} \left(\theta_l \sum_{i=1}^{T_l} (\alpha_i^l)^{\bar{\delta}_{i,l}} \right)}, \bar{\delta}_{i,k} = \frac{\delta_{i,k}}{\max\limits_{i=1,\cdots,T_k} \{\delta_{i,k}\}} \tag{21}$$

即将个体匹配度的累乘形式改为累加形式,这样激活条件变为了只要规则中某个属性候选值的个体匹配度不为零,那么该条规则就会被激活。

采用线性组合方式构建的 BRB 分类方法在一定程度上已能有效地解决分类问题,然而由于线性组合的方式也暴露出以下两点不足:

(1)线性组合的方式迫使每个前件属性候选值的个数必须相等,这样就忽略了各个前件属性间的差异。在原始 BRB 规则库中,各个前件属性候选值个数一般不相等,其个数往往取决于该属性值的区间大小以及所占的权重比例,通常情况下,区间越大,比重越大,则候选值的个数也就越多。文献[13]所提出的方法中,前件属性权重已从激活规则公式中删去,也就是说,该方法认为各个前件属性同等重要,这往往不可取,因为这样会夸大弱属性的效用,而弱化了强属性的分类支持度,导致最终的分类准确性下降。

(2)激活权重公式的改变导致了后件置信度受到了来自非激活前件属性候选值的影响。即原始方法中,参与 ER 合成的置信度,其所对应规则的候选值都存在个体匹配度。从修正后的激活权重公式来看,规则中存在某个属性候选值的个体匹配度不为零,那么该条规则所携带的信息就会参与 ER 合成。倘若只有一个候选值的个体匹配度不为零,那么完全可以弱化该条规则对结果所占比重的影响。特别是当该激活点属于噪音情况时,肯定会对结果造成不利的影响,使得最终的分类准确性下降。从式(8~13)的计算过程可以看出,随着后件评价等级个数的增加,该缺点所带来的影响也会随之增大。

3 基于二择众仓的置信规则库推理的分类方法

针对现有置信规则库推理分类方法的不足,本文提出一种二择众仓决策法。对原有的方法作如下改进:

(1)将规则表达式修正为

$$R_k: \text{if } X_1 \text{ is } A_1^k \wedge X_2 \text{ is } A_2^k \wedge \cdots \wedge X_{T_k} \text{ is } A_{T_k}^k, \text{then}\{(D_1, \bar{\beta}_{1,k}), (D_2, \bar{\beta}_{2,k})\} \tag{22}$$

即每条规则仅设计两个评价等级,置信规则库的推理结果只对输入值作出是与否的置信决策,而不再作多值判定,仅仅是在二者间选择。其好处是规则变得简单,在处理信息时更加快捷高效,而且该设计策略能够很好地改善第 2 节中所提到的第 2 点不足,从而使得分类准确性有所提高。

(2)引入众仓决策模型。在对规则表达式作出修正后可以很明显地看出由于后件评价等级个数只有两个,那么由此类规则所构建的置信规则库只能对二分类问题做出判定,而无法解决多分类问题。因此本文创新性地提出采用多规则库的方式解决同一个分类问题,即每一个置信规则库都是一个二分类决策器,而由多个二分类决策器构成了众仓决策模型,进而解决多分类问题。采用众仓模型后,每一个置信规则库只关心两个类别间或两个大类别间的差异,此时每一个置信规则库都可以有自己的前件属性权重值。例如:存在一个 4 属性、3 类别的分类数据集,可以在两两类别间设置一个置信规则库,假设区分 1,2 类别仅需要前两个属性,那么对于第 1 个置信规则库完全可以将 3,4 属性的权重值设置为零。该优点是现有置信规则库分类方法所不能拥有的,只有一个置信规则库的分类方法,其前件属性权重必须同时考虑区别 1,2,3 类。因此,采用众仓决策模型可以很好地解决第 2 节中所提到的第 1 点不足。

图 2 给出了三类别示例。如图所示,在类别 1 与类别 2 间、类别 1 与类别 3 间找到一个可分平面十分容易,然而在类别 2 与类别 3 间找到一个可分平面就相对比较困难,采用现有的分类方法无法直接了断地解决该瓶颈,只能通过增加规则条数或参数训练复杂度的方法来解决,这必然会对类别 1 产生影响。若是采用二择众仓决策法,类别 1 与类别 2 间、类别 1 与类别 3 间的分类器在训练过程中可以很快地获得精准结果,而对于类别 2 与类别 3 间的分类器,在训练过程中就可以发现这是整个分类问题的瓶颈,这样就不再是"黑箱"操作,而是清楚明了地知道问题瓶颈所在的位置,此时可以通过增加规则条数或增加参数训练复杂度的方法来解决,同时对类别 1 不会造成任何影响。这也是二择众仓决策法的优点之一。

二择众仓决策法不改变 RIMER 方法的整个体系结构,单个置信规则库的所有操作都与原来一致,改变的只是评价等级个数,也就是让单个置信规则库只解决一个小问题,将处理结果返回给该体系外的众仓决策模型,该模型将多个小问题的结果进行融合,从而解决问题。这样不仅继承了原有 RIMER 方法中的优点,而且在一定程度上改善了由线性组合方式所带来的不足。二择众仓决策法的具体实现步骤为

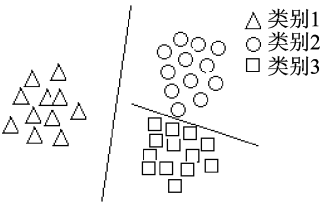

图 2 三类别示例

Fig. 2 Illustraction example of three categories

(1)依据数据的特性选择合适的众仓决策模型。以 Glass 数据集为例,通过查看其类别描述可获得如下信息:

Class Distribution: (out of 214 total instances)

——163 Window glass (building windows and vehicle windows)

-- 87 float processed

-- 70 building windows

-- 17 vehicle windows

-- 76 non-float processed

-- 76 building windows

-- 0 vehicle windows

-- 51 Non-window glass

-- 13 containers

-- 9 tableware

-- 29 headlamps

可以看出 Glass 数据集的类别结构层次分明,能很容易地采用有向无环图来构建众仓决策模型,如图 3 所示。Glass 数据集中类别 4 的个数为零,故未在图中标出。从图 3 可以看出,对 Glass 数据集进行分类需要设计 6 个 BRB 分类器,其中 BRB_1 用来区分 Window glass 与 Non-window glass 这两类,Window glass 有 4 个类别标签:1,2,3 和 4,Non-window 有 3 个类别标签:5,6 和 7。即 BRB_1 是区分 1,2,3,4 类和 5,6,7 类的分类器。BRB_2,BRB_3 和 BRB_4 同理不再赘述。BRB_4,BRB_5 和 BRB_6 采用 One-versus-one 算法,即在两两类间训练一个分类器,当对一个未知样本进行分类时,每个分类器都对其进行判定,并给相应的类别"投上一票",最后票数最多的类别作为该样本的类别。该步骤可用图 4 所示的流程图表示。

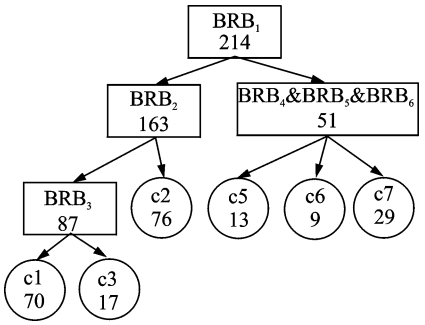

图 3 Glass 数据集的众仓决策模型

Fig. 3 Multiply decision-making model of Glass dataset

(2)由于各个分类器在训练的过程中相互不存在影响,可以采用并行的策略同时进行训练。训练的过程中发现有瓶颈问题,可以不断调整训练参数,以获得更加满意的结果。本文中的实验均采用差分进化算法对 BRB 的参数进行训练。以均方差(Mean squared error, MSE)作为参数训练模型中的目标函数,即

$$\min \mathrm{MSE}(x)=\frac{1}{T}\sum_{t=1}^{T}E$$

$$E=\begin{cases}1 & \hat{c}\neq c\\ 0 & \hat{c}=c\end{cases} \tag{23}$$

式中:T 表示训练集的大小,$\hat{c}$ 表示推理类别,c 表示真实类别。

(3)后件置信度转换成类别信息。由于后件评价等级只有两个,将两个评价等级与两个分类级别对应起来,则最终的分类结果为

$$c=\begin{cases}i & \beta_i\geqslant\beta_j\\ j & \beta_i<\beta_j\end{cases} \tag{24}$$

其中 i,j 表示类别的编号。

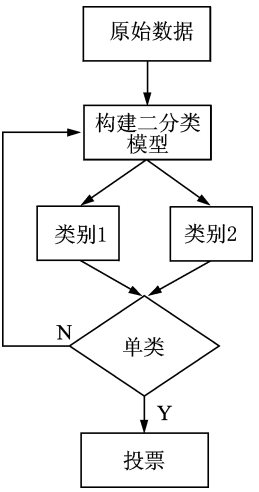

图 4 分类流程图

Fig. 4 Flow chart of classification

4 实验分析与结果对比

通过实验将二择众仓决策法与现有的分类方法进行对比,以差分进化算法作为参数训练的优化算法。一般情况下将种群规模设置在 50,交叉概率为 0.9,缩放因子为0.5。实验环境为:Intel(R)Core(TM)i5-4570 CPU @3.20 GHz 处理器,8 GB 内存,Windows7 操作系统。程序均在 Matlab2014b 中实现。

4.1 实验 1

本实验所使用的 3 个数据集均来自 UCI 公共测试集,分别为:Iris,Wine 和 Glass。表 1 显示了 3 个测试数据集的基本信息。

表 1 数据集基本信息

Tab. 1 Information of data sets

名称	属性数量	类别个数	样本数
Iris	4	3	150
Wine	13	3	178
Glass	9	7	214

采用十折交叉验证法，即将样本数据分成 10 份，每次取其中的 1 份作为测试集，其余的作为训练集。十折交叉验证法的实验结果如表 2 所示。

表 2 十折交叉验证结果

Tab. 2 10-fold cross-validation results

样本数据	Iris		Wine		Glass	
	Acc%	Err	Acc%	Err	Acc%	Err
1	100	0	100.00	0	85.71	3
2	100	0	100.00	0	85.71	3
3	100	0	100.00	0	85.71	3
4	100	0	94.11	1	80.95	4
5	100	0	100.00	0	80.95	4
6	100	0	94.11	1	66.66	7
7	100	0	100.00	0	85.71	3
8	100	0	100.00	0	80.95	4
9	100	0	94.11	1	80.95	4
10	100	0	100.00	0	95.24	1
Average	100	0	98.31	0.3	83.18	3.6

为了进一步验证本文方法的有效性，将本文方法与文献[13,14]同样是采用置信规则库推理的分类方法进行对比，并选取近两年来对这 3 个数据集进行分类的其他方法进行对比，对比结果如表 3 和图 5～7所示。

表 3 不同方法在不同数据集上的分类准确率对比

Tab. 3 Classification accuracy comparison by using different methods on different data sets

方法	备注	Iris/%	Wine/%	Glass/%
LST-KSVC[18]	Neural network	99.27	94.27	65.76
FGGCA[19]	Fuzzy set	97.22	97.10	93.65
WLTSVM[20]	SVM	98.00	96.40	49.91
BRB[13]	Ye, et al.	96.93	—	61.86
BRB[14]	Chang, et al.	100.00	99.44	70.09
BRB	本文	100.00	98.31	83.18

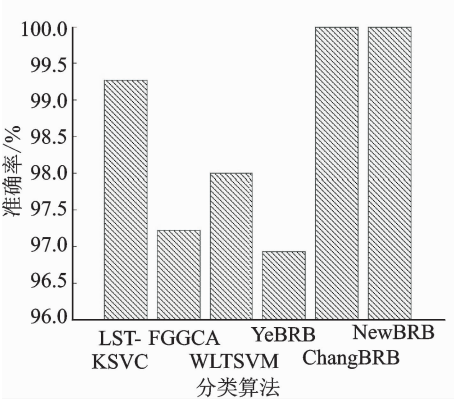

图 5 Iris 数据集的各方法分类准确率对比

Fig. 5 Classfication acouracy cortrast by using different methods on Iris dataset

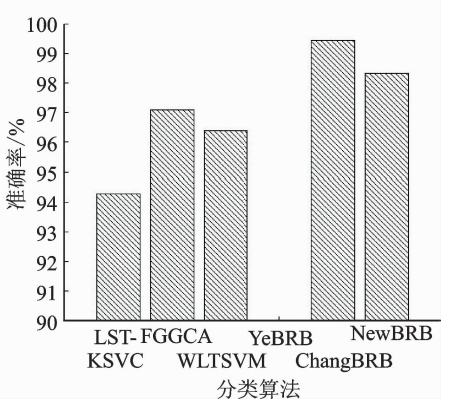

图 6 Wine 数据集的各方法分类准确率对比

Fig. 6 Classification accuracy contrast by using different methods on Wine dataset

将本文方法与非 BRB 方法进行对比,可以看出除了 FG-GCA 的 Glass 数据集外,余下结果都不如本文的方法来得更优。将本文方法与文献[13,14]的方法进行对比,文献[14]的实验缺失 Wine 数据集的结果,然而从 Iris 和 Glass 数据集来看,其结果均不如本文的方法好。而对比文献[13]方法,在 Glass 数据上本文的方法有了很大的提升,这是由于 Glass 数据集带有二分类的特性,特别适用于本文提出的方法。没有任何一种方法能对所有数据集均达到最优的结果,但纵观全局来看,本文的方法在一定程度上提升了分类的准确性。

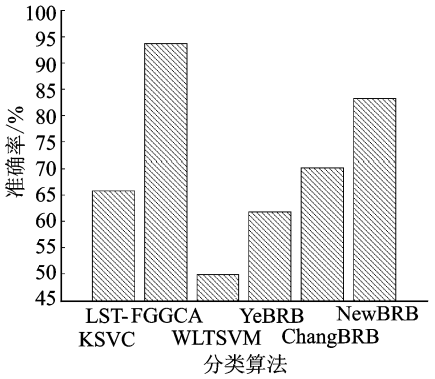

图 7 Glass 数据集的各方法分类准确率对比

Fig. 7 Classification accuracy contrast by using different methods on Glass dataset

4.2 实验 2

本实验在于说明文献[13]提出的方法随着类别个数的增多,其分类准确率将会急剧下降。而采用二择众仓决策方法,将会大幅度地减少准确率的下降程度。实验数据样本来源于 Brodatz 库中的 14 幅纹理图像,其在 Brodatz 库中的编号分别为 D1,D6,D12,D15,D20,D34,D37,D52,D56,D65,D72,D87,D93 和 D110,具体如图 8 所示。

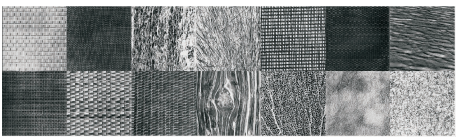

图 8 Brodatz 库中的 14 幅纹理图像

Fig. 8 Fourteen texture images from Brodatz library

这 14 幅纹理图像原始大小均为 640 像素×640 像素,将每幅图像不重叠地切割成 4×4=16 幅子图,每幅子图大小均为 160 像素×160 像素,共获得 14×16=224 幅样本图像。选取每幅图像的前 8 幅子图作为训练数据,后 8 幅子图作为测试数据。使用 Matlab2014b 所提供的 graycomatrix 和 graycoprops 函数获取每幅子图的灰度共生矩阵特征值,并以此作为分类依据。实验类别个数从初始的 4 个逐步增加至 14 个。实验结果如图 9 所示。

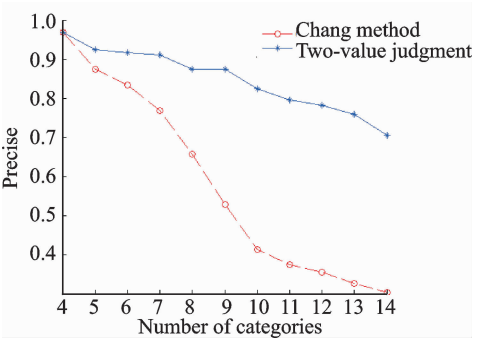

图 9 实验结果对比

Fig. 9 Experimental results comparison

从实验结果可以看出,在类别个数较少时,文献[13]方法与二择众仓决策方法差别不大。但随着类别个数的增多,其准确率会急剧下降,对 14 幅纹理图像同时进行分类时其准确率仅有 30.36%。这是由于线性组合方式并不适用于类别个数较多的情况,这已在第 2 节中做了详细说明。而采用二择众仓决策方法后会大幅度地减少该不足所带来的影响,对 14 幅纹理图像同时进行分类时依然能保持较高的准确率,说明本文所提出的方法具有较强的鲁棒性。

5 结束语

虽然通过线性组合方式构建置信规则库避免了规则条数随着问题属性数量的增多而激增,但该方式仍然存在不足。鉴于此,本文提出一种改进置信规则库推理的分类方法,首先将一个大的分类问题切割成若干个相对独立的分类小问题,每个小问题都是一个二分类问题;其次将规则的后件评价等级设置为两个让每个置信规则库只处理一个小问题,以此减小线性组合所来的的误差;最后,采用众仓决策的方式将若干个小问题的结果进行整合,从而得出最终的分类结果,通过实验分析验证了该方法的可行性。本文在现有置信规则库推理分类算法的基础上,通过改进分类器的设计,从而进一步提高了分类准确率。如何处理

类别数量更多的分类问题将是下一步研究的方向。

参考文献：

[1] Cover T, Hart P. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

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

[3] Bazan J G, Nguyen H S, Nguyen S H, et al. Rough set algorithms in classification problem[J]. Rough Set Methods and Applications, 2000, 56(1): 49-88.

[4] Yang Jianbo, Liu Jun, Wang Jin, 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.

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

[6] Dempster A P. A generalization of Bayesian inference[J]. Journal of the Royal Statistical Society, 1968, 30(2): 205-247.

[7] Shafer G. A mathematical theory of evidence[M]. Princeton: Princeton University Press, 1976: 10-39.

[8] Hwang C L, Yoon K. Methods for multiple attribute decision making[M]. [S. l.]: Sringer Berlin Heidelberg, 981: 58-191.

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

[10] 周志杰，杨剑波，胡昌华，等. 置信规则库专家系统与复杂系统建模[M]. 北京：科学出版社，2011：9-41.
Zhou Zhijie, Yang Jianbo, Hu Changhua, et al. Belief rule base of expert system and complex system modeling[M]. Beijing: Science Press, 2011: 9-41.

[11] Liu Jun, Yang Jianbo Ruan Da, et al. Self-tuning of fuzzy belief rule bases for engineering system safety analysis[J]. Annals of Operations Research, 2008, 163(1): 143-168.

[12] Jiang Jiang, Li Xuan, Zhou Zhijie, et al. Weapon system capability assessment under uncertainty based on the evidential reasoning approach[J]. Expert Systems with Applications, 2011, 38(11): 13773-13784.

[13] Chang Leilei, Zhou Zhijie, You Yuan, et al. Belief rule based expert system for classification problems with new rule activation and weight calculation procedures[J]. Information Sciences, 2016, 336(1): 75-91.

[14] 叶青青，杨隆浩，傅仰耿. 基于改进置信规则库推理的分类方法[J]. 计算机科学与探索，2016，10(5)：709-721.
Ye Qingqing, Yang Longhao, Fu Yanggeng. Classification approach based on improved belief rule-base reasoning[J]. Computer Science and Technology, 2016, 10(5): 709-721.

[15] Yang Jianbo, Liu Jun, Xu Dongling, et al. Optimization models for training belief-rule-based systems[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2007, 37(4): 569-585.

[16] Zhou Zhiguo, Liu Fang, Jiao Licheng, et al. A bi-level belief rule based decision support system for diagnosis of lymph node metastasis in gastric cancer[J]. Knowledge-Based Systems, 2013, 54: 128-136.

[17] Calzada A, Liu J, Wang H, et al. A new dynamic rule activation method for extended belief rule-based systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(4): 880-894.

[18] Nie Qingfeng, Jin Lizou, Fei Shumin, et al. Neural network for multi-class classification by boosting composite stumps[J]. Neurocomputing, 2015, 149: 949-956.

[19] Sanchez M A, Castillo O, Castro J R, et al. Fuzzy granular gravitational clustering algorithm for multivariate data[J]. Information Sciences, 2014, 279: 498-511.

[20] Shao Yuanhai, Chen Weijie, Wang Zhen, et al. Weighted linear loss twin support vector machine for large-scale classification [J]. Knowledge-Based Systems, 2015, 73: 276-288.

作者简介：

方志坚(1990-)，男，硕士研究生，研究方向：智能决策技术、置信规则库推理，E-mail：1350553313@qq.com。

傅仰耿(1981-)，男，博士，副教授，研究方向：决策理论与方法、数据挖掘与机器学习。

陈建华(1959-)，男，副教授，研究方向：多媒体技术、CAD。

（责任编辑：王静）