一种改进的案例推理分类方法研究

张春晓 [1] 严爱军 [1] 王 普 [1]

摘 要 特征属性的权重分配和案例检索策略对案例推理 (Case-based reasoning, CBR) 分类的准确率有显著影响. 本文提出一种结合遗传算法、内省学习和群决策思想改进的 CBR 分类方法. 首先, 利用遗传算法得到多组属性权重, 再根据内省学习原理对每组权重进行迭代调整; 然后, 通过案例群检索策略得到满足大多数原则的群决策分类结果; 最后, 以典型分类数据集的对比实验证明了本文方法能进一步提高 CBR 分类的准确率. 这表明内省学习可以保证权重分配的合理性, 案例群检索策略能充分利用案例库的潜在信息, 对提升 CBR 的学习能力有显著作用.

关键词 案例推理, 案例检索, 遗传算法, 内省学习, 群决策

引用格式 张春晓, 严爱军, 王普. 一种改进的案例推理分类方法研究. 自动化学报, 2014, **40**(9): 2015−2021

DOI 10.3724/SP.J.1004.2014.02015

An Improved Classification Approach by Case-based Reasoning

ZHANG Chun-Xiao[1] YAN Ai-Jun[1] WANG Pu[1]

Abstract The distribution of feature attribute weights and the strategy of case retrieval have significant impacts on the classification accuracy of case-based reasoning (CBR). An improved CBR classification approach is proposed, which is combined with genetic algorithms, introspective learning, and group decision-making theory. First, multiple attribute weights are given by a genetic algorithm. Then each group of weights is iteratively adjusted in accordance with the introspective learning principle. After that the group decision-making retrieval result which satisfies the plurality rule can be obtained according to the case group-retrieval strategy. At last, the classification comparison experiments prove that the proposed method could improve the classification accuracy of CBR. The results indicate that the introspective learning could guarantee the rationality of weight allocation, and that the case group-retrieval strategy could make full use of the potential knowledge of case base, having remarkable effects on promoting the learning ability of CBR.

Key words Case-based reasoning (CBR), case retrieval, genetic algorithms, introspective learning, group decision-making

Citation Zhang Chun-Xiao, Yan Ai-Jun, Wang Pu. An improved classification approach by case-based reasoning. *Acta Automatica Sinica*, 2014, **40**(9): 2015−2021

案例推理 (Case-based reasoning, CBR) 是人工智能领域一种较新的问题求解和机器学习方法, 崛起于上世纪 80 年代初[1]. 它起源于对人类认知心理学的研究, 基本思想是利用过去解决类似问题的经验案例 (或称源案例) 进行推理求解新问题 (或称目标案例)[2]. 1994 年, Aamodt 等将 CBR 方法归纳为 4 个阶段: 检索、重用、修正和存储[3], 即通过检索得到一个或几个先前的案例; 然后, 重用相应的解决方案或者以此为基础修正案例来解决新问题; 最后, 存储新的案例供未来的问题求解使用. 作为人工智能技术的重要组成部分, CBR 的应用研究得到了广泛关注, 比如, 在智能优化控制[4−5]、分类与诊断[6−7] 以及信息管理[8] 等领域取得了阶段性成果. 然而, CBR 在复杂环境中的决策和学习能力仍有很大发展空间.

在 CBR 模型中, 案例检索处于首位, 检索策略的合理与否直接影响整个推理系统的实现效果[9]. 在检索策略方面主要有知识导引策略[10]、模板检索策略[11] 以及最近邻策略[12−13] 等. 其中, 以相似度为基础的最近邻策略得到了广泛关注, 但这种策略与其他检索策略均未提及如何充分利用特征属性的潜在知识, 当样本提供的信息量不足时, 利用此检索策略的求解准确率并不高, 因此, 改进的检索策略需要关注如何充分挖掘样本信息.

此外, 最近邻检索策略通过计算属性间的加权距离来衡量案例的相似程度, 虽然简单易行, 但对于噪声或者不相关的数据比较敏感, 这一问题的解决通常依赖于合理分配特征属性的权重. 很多文献讨论了权重的分配方法, 特别是遗传算法在此得到了

收稿日期 2013-05-29 录用日期 2014-05-15
Manuscript received May 29, 2013; accepted May 15, 2014
国家自然科学基金 (61374143)
Supported by National Natural Science Foundation of China (61374143)
本文责任编委 刘德荣
Recommended by Associate Editor LIU De-Rong
1. 北京工业大学电子信息与控制工程学院 北京 100124
1. College of Electronic Information & Control Engineering, Beijing University of Technology, Beijing 100124

广泛关注[14−15]. 然而, 由于遗传算法的随机迭代过程存在易陷于局部极小的局限性, 这一缺陷可能导致它分配的权重不合理, 继而影响 CBR 问题求解的效果. 因此, 如何利用遗传算法的寻优能力有待于进一步研究.

在认知科学的理论中, 人类常常通过内省即自我观察实现自我提高. 将内省学习的概念引入机器学习中, 通过检查智能系统自身的知识处理和推理方式来改进存在的问题, 实现修正自身的学习目标, 可以提高一个学习系统的性能[16−18]. 因此, 针对上述问题, 本文在遗传算法优化分配特征属性权重的基础上, 通过内省学习原理对这些权重进行迭代调整; 然后, 借鉴群决策中的简单多数原则, 利用多组内省学习后的权重作为群权重去实现案例群检索; 最后, 通过分类实验对本文方法的有效性进行验证.

本文余下部分的结构安排如下: 第 1 节介绍 CBR 推理方法及存在的问题; 第 2 节讨论如何利用内省学习对遗传算法分配的权重进行迭代调整; 第 3 节介绍基于简单多数原则的案例群检索策略; 第 4 节通过 UCI (University of Californialrvine) 数据集进行分类对比实验, 并给出结果分析; 最后, 对本文方法和未来研究方向进行了总结.

1 CBR 及问题分析

CBR 是人工智能领域一种较新的问题求解和机器学习方法, 经典的 CBR 循环如图 1 所示[3]. 历史案例库中存储的是问题描述及相应的解答组成的源案例. 目标案例出现时, 通过案例检索评估源案例与目标案例的相似度以获取相似案例; 再通过案例重用得到目标案例的建议解; 案例修正阶段对此建议解进行评价, 若评价为失败解则需修正, 直至获得正确解; 最后通过案例保存将目标案例及其正确解

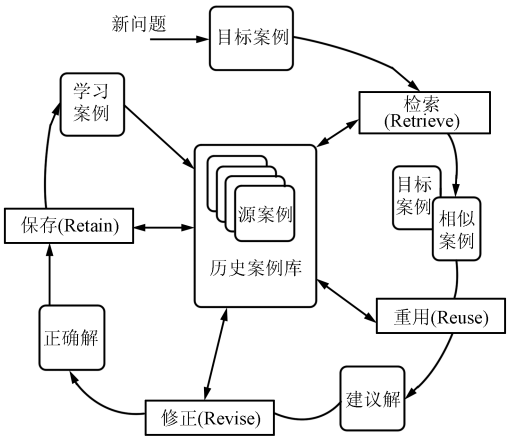


图 1 CBR 循环

Fig. 1 The CBR cycle

(也称为学习案例) 储存于案例库中, 用于将来的问题求解, 从而实现 CBR 的增量式学习过程. CBR 用于分类问题时, 其案例表示和求解过程如下所述:

历史案例库中的源案例可采用如下的二元组表示形式:

$$C_k = (X_k; Y_k), \qquad k = 1, 2, \cdots, p \tag{1}$$

其中, p 是源案例总数; X_k 是第 k 条源案例的问题描述; Y_k 是 X_k 所属的类别. X_k 可表示为

$$X_k = (x_{1,k}, x_{2,k}, \cdots, x_{i,k}, \cdots, x_{m,k}) \tag{2}$$

其中, $x_{i,k}$ (i=1, 2, $\cdots$, m) 表示第 k 条源案例中的第 i 个特征属性的描述值; m 是特征属性的个数.

以模式分类问题为例, 设目标案例的问题描述为 X= (x_1, x_2, $\cdots$, x_i, $\cdots$, x_m), 需推理求出相应的类别 Y.

按最近邻检索策略[13], 计算 X 与 X_k 的相似度

$$S(X, X_k) = 1 - \left(\sum_{i=1}^{m} \omega_i (x_i - x_{i,k})^2 \right)^{\frac{1}{2}} \tag{3}$$

其中, ω_i 是第 i 个特征属性的权重, 满足

$$\begin{cases} \sum\limits_{i=1}^{m} \omega_i = 1 \\ \omega_i \geq 0 \end{cases} \tag{4}$$

根据式 (3) 的计算, 可从中获得最大相似度 $S_{\max}$:

$$S_{\max} = \max(S(X, X_k)), \qquad k = 1, 2, \cdots, p \tag{5}$$

设最大相似度 $S_{\max}$ 对应于第 q 个案例记录, 则与目标案例 X 最为相似的案例记录为 C_q = (X_q; Y_q).

对于分类问题, 通常直接重用 C_q 的类别 Y_q 作为目标案例 X 的建议类别, 即

$$Y = Y_q \tag{6}$$

上式表明, 通过重用具有最大相似度的相似案例的类别可以得到目标案例的建议类别. 当建议类别与实际情况不完全符合时, 还需要进行案例修正, 目前一般是通过人机交互平台完成. 最后, 将目标案例的描述 X 及其相应的解答 Y 存储于案例库中, 用于下一次的分类求解. 至此, 源案例的总数由 p 变为 $p+1$ 个.

上述的最近邻检索策略强调了加权作用, 很多文献讨论了权重的分配方法, 其中遗传算法的应用得到了广泛关注[14−15]. 但遗传算法存在易陷于局部极小的自身缺陷, 会影响其分配权重的效果. 而且在

样本量不足的情况下, 如果只根据一组优化权重进行单次最近邻检索, 检索的准确率可能不高. 这些因素的存在限制了 CBR 分类的准确性, 因此, 如何提高权重分配的合理性, 以及如何利用多组权重的检索策略还需进一步研究.

2 权重调整方法

基于上述分析, 本节从认知科学的角度探索权重的迭代调整方法. 首先将遗传算法得到的权重视为初始权重, 然后借助于内省学习原理, 对这些初始权重按成功驱动策略作进一步调整, 以保证 CBR 具有积极学习能力.

2.1 遗传算法分配权重

对于优化问题, 遗传算法运行简单、鲁棒性强, 根据定义的适应度函数可有效地进行全局随机搜索[19]. 利用遗传算法对属性权重进行迭代搜索主要包括以下环节:

1) 编码和解码. 对每个特征属性权重采用 n 位二进制编码, 不同的编码值表示权重的大小不同. 计算相似度的时候, 需将二进制码解码成十进制数, 然后映射至区间 [0, 1]. 以第 i 个特征属性的权重计算为例:

$$\omega_i = \frac{t_i}{\sum\limits_{i=1}^{m} t_i} \tag{7}$$

其中, ω_i 是第 i 个特征属性的权重; t_i 是二进制的解码.

2) 适应度函数. 遗传算法在迭代搜索过程中以适应度函数作为进化的依据, 对分类问题, 可定义如下的适应度函数:

$$F = \frac{\sum\limits_{j=1}^{Q} q_j}{Q} \tag{8}$$

其中, Q 表示测试集中的目标案例总数, q_j 表示第 j 个目标案例分类是否正确的标志, 用 0 和 1 表示, 分类正确则 $q_j = 1$; 否则为 0, F 表示分类的准确率. 该函数可以保证向准确率高的方向对权重进行迭代搜索.

3) 选择. 按照适应度比例法 (又称轮盘赌法) 选择合适的权重分布, 每组权重的选择概率与其适应度成比例. 设初始权重的规模数量为 N (这里的权重为二进制编码形式, 下同), 根据式 (8) 得到的每组权重的适应度为 F_l ($l = 1, 2, \cdots, N$), 则第 l 组权重被选择的概率为

$$P_l = \frac{F_l}{\sum\limits_{l=1}^{N} F_l} \tag{9}$$

然后根据轮盘赌规则确定每组权重是否被选择复制到下一次的迭代过程.

4) 交叉. 采用单点交叉方式, 在相邻的两组权重的二进制码串上随机选择某位, 然后以此点为界将其分为左右部分, 根据设定的交叉概率大小决定是否将两组权重的左右部分互相交换, 交叉后会形成两组新的权重.

5) 变异. 为了提高遗传算法的全局搜索能力, 还需要执行变异操作. 从交叉后的每组权重中, 随机选取某位, 根据设定的变异概率大小决定是否将该位取反, 若满足取反条件, 则产生新的权重分布.

通过上述的迭代搜索, 在规定的迭代次数内或以设定的分类准确率为终止条件, 可以得到最终的一组权重参与式 (3) 的相似度计算.

2.2 内省学习调整权重

在相似度计算过程中, 权重的不同分配会得到不同的相似度, 进而影响分类的结果. 遗传算法通过领域知识的获取对属性权重进行分配, 而现实问题的复杂多样性导致的结果是: 即使一个推理者具备了正确的领域知识也难免会推理失败. 内省学习在分析任务执行成功或失败的基础上形成学习目标, 可以根据系统自身的性能进行学习. 具有自省能力的系统具备改变知识处理和推理方法的能力, 更具灵活性和适应性. 因此, 本节将内省学习的概念引入到权重的调整中, 基于成功驱动策略对遗传算法搜索得到的权重进行调整.

1) 匹配属性和不匹配属性

根据两个案例相同属性的描述值, 可将案例属性分为匹配属性和不匹配属性, 以下给出其定义:

定义 1. 设同一个案例库中的两个源案例为 $C_I = (X_I; Y_I)$ 和 $C_{II} = (X_{II}; Y_{II})$, 其中, $X_I = (x_{1,I}, x_{2,I}, \cdots, x_{i,I}, \cdots, x_{m,I})$, $X_{II} = (x_{1,II}, x_{2,II}, \cdots, x_{i,II}, \cdots, x_{m,II})$, 若满足 $0 \le |x_{i,I} - x_{i,II}| \le \xi_{\text{match}}$ (其中, ξ_{match} 为判断属性是否匹配的阈值, 表示两个案例同一属性的接近程度), 则称源案例 C_I 与 C_{II} 的第 i 个属性为匹配属性, 否则, 称为不匹配属性.

2) 权重学习策略

基于内省学习原理, 采取如下成功驱动的权重学习策略:

存在一条案例分类成功时, 匹配属性的权重增加, 同时不匹配属性的权重减少 (Good matching up & Good un-matching down, 简记为 GMU+GUD). 对目标案例 X 来说, 如果检索到的

相似案例为 A, 当 A 和 X 的类别相同 (即分类成功) 时, 增大 A 与 X 中的匹配属性的权重, 同时减小 A 与 X 中的不匹配属性的权重. 如此, A 与 X 的相似度将增大, 保证在求解目标案例 X 时, 案例 A 能被成功检索.

3) 权重调整方式

通过内省学习对权重进行调整的基本方法是增减或乘除一个固定值[18−19]. 前者保证在学习进程中权重的变化是一个常数, 而后者则会使权重随学习进程发生衰变. 采用如下公式进行权重的调整:

权重增加:

$$\omega_i(t+1)=\omega_i(t)+\frac{\Delta_i}{m},\qquad \omega_i(t+1)\geq 0 \tag{10}$$

权重减少:

$$\omega_i(t+1)=\omega_i(t)-\frac{\Delta_i}{m},\qquad \omega_i(t+1)\geq 0 \tag{11}$$

其中, $\omega_i(t)$ 是第 i 个特征属性第 t 次迭代的权重; $\omega_i(t+1)$ 是第 i 个特征属性第 $t+1$ 次迭代的权重; Δ_i 是第 i 个权重的调整因子, m 是特征属性的个数. 对于不同的分类问题, 如果特征属性的个数不同, 即使 Δ_i 相同, 权重的变化量也会有所差别. 但对同一个案例库而言, m 固定不变, 所以权重的变化主要与 Δ_i 有关. 为实现权重的有效调整, 且保证其调整值不超过所有权重之和, Δ_i 在区间 $(0,\, m)$ 内取值, 并且需满足式 (10) 和式 (11) 更新后的权重不小于零.

当所有特征属性的权重都被调整之后, 为保证权重之和为 1, 需进行归一化操作:

$$\omega_i(t+1)'=\frac{\omega_i(t+1)}{\sum\limits_{i=1}^{m}\omega_i(t+1)} \tag{12}$$

其中, $\omega_i(t+1)'$ 是第 i 个属性第 $t+1$ 次迭代调整后的最终权重.

3 案例群检索策略

如前所述, 传统的最近邻检索策略未提及如何充分利用案例特征属性的潜在知识, 会影响分类的准确率, 本节引入群决策方法改进检索策略以弥补这一不足.

3.1 案例群检索策略

群决策是集中群中成员的意见以形成群的意见[20]. 群决策的群体认知能力可弥补个体经验和知识的不足, 做出更合理的决策. 案例检索问题也可看作一种决策问题, 当样本量不足时, 单一最近邻检索的结果有时会不准确, 因此可利用群决策进行多次检索, 发挥其群体认知的优势去获得检索结论. 群决策中的社会选择理论提出采用某种与群中成员的偏好有关的数量指标来反映群, 其中一种规则称为简单多数准则[21], 此准则可保证群决策的连续性与稳定性. 本文采用此准则进行群意见的集结. 以最近邻检索策略[13] 为基础, 采用简单多数原则的案例群检索策略描述如下:

设有 s 位专家, 则需得到 s 组权重构成群权重 GW. 群权重可由第 2 节介绍的方法得到: 首先利用遗传算法得到多组权重, 再对遗传算法得到的每组权重分别利用内省学习进行调整, 调整后的各组权重即构成群权重 GW, 以矩阵形式表示为

$$GW=\begin{bmatrix}\omega_{11} & \omega_{12} & \cdots & \omega_{1i} & \cdots & \omega_{1m}\\ \omega_{21} & \omega_{22} & & & & \omega_{2m}\\ \vdots & & \ddots & & & \vdots\\ \omega_{j1} & & & \omega_{ji} & & \omega_{jm}\\ \vdots & & & & \ddots & \vdots\\ \omega_{s1} & \omega_{s2} & \cdots & \omega_{si} & \cdots & \omega_{sm}\end{bmatrix} \tag{13}$$

其中, GW 的每一行表示一组权重分配结果, 代表一位专家分配的权重. ω_{ji} 是第 j 位专家为第 i 个特征属性分配的权重.

每位专家分配权重后, 以第 j 位专家分配的权重为例, 将式 (3) 改写, 计算目标案例 X 与第 k 条源案例 X_k 的相似度:

$$S_j(X,X_k)=1-\left(\sum_{i=1}^{m}\omega_{ji}(x_i-x_{k,i})^2\right)^{\frac{1}{2}} \tag{14}$$

其中, $S_j(X,\, X_k)$ 是第 j 位专家权重得到的第 k 条源案例 X_k 与目标案例 X 的相似度.

求相应于每位专家权重的最大相似度 $S_{j\max}$:

$$S_{j\max}=\max(S_j(X,X_k)),\qquad k=1,2,\cdots,p \tag{15}$$

设最大相似度对应的源案例是第 q 个, 则每位专家权重得到的与目标案例 X 最为相似的源案例记录为 $C_q=(X_q;\, Y_q)$.

对于分类问题, 则第 j 位专家权重的个体决策结果为 $Y_{j,q}=Y_q$. 下面以二分类问题为例, 概述如何采用简单多数准则得到目标案例 X 的群决策结果.

考虑两种类别 a 和 b, 则 $Y_{j,q}$ 表示 a 或 b, 根据第 j 位专家的个体决策 $Y_{j,q}$, 可统计得到 a 和 b 出现的次数, 分别记为 N_a 和 N_b. 若 N_a 和 N_b 不相

等, 则对于目标案例 X, 群决策的最终分类结果为

$$Y=\begin{cases} a,\ N_a>N_b \\ b,\ N_a<N_b \end{cases} \tag{16}$$

若 N_a 和 N_b 相等, 则根据最大相似度得到群决策结果:

$$Y=\arg\max_{Y_{j,q}}(S_{j\max}),\qquad j=1,2,\cdots,s \tag{17}$$

3.2　算法收敛性分析

案例群检索策略的核心体现于式 (14), 该式与最近邻算法中对案例相似度的计算公式 (3) 类似, 因此可直接以最近邻算法的收敛条件分析案例群检索策略的收敛性. Cover 等[13] 给出了最近邻算法收敛性的条件: 如果 $X_k\to X$ 的概率为 1, 则算法是收敛的. 据此, 要证明案例群检索策略的收敛性, 只需证明利用案例群检索策略检索到的最近邻案例 X_k 以概率 1 收敛于目标案例 X.

引理 1. 设 X 和 $X_1, X_2, \cdots, X_k$ 是在可分离的度量空间 $\mathbf{C}$ 上取值的独立恒等分布的随机变量, X_k 表示目标案例 X 的最近邻, 取自集合 $\{X_1, X_2, \cdots, X_p\}$, 则 $X_k\to X$ 的概率为 1.

证明. 设 $S_X(r)$ 是以 X 为中心, r 为半径的球体, 假定对于给定的点 $X\in\mathbf{C}$, 每一个球体 $S_X(r)$ $(r>0)$ 均有非零的概率密度.

那么, 任意近邻 X_k 落在以 X 为中心的球体 $S_X(\delta)$ $(\delta>0)$ 里的概率为一个正数, 记为 $P(S_X(\delta))$; 某一近邻 X_k 落在球体 $S_X(\delta)$ $(\delta>0)$ 外的概率为 $1-P(S_X(\delta))$, 则 K 个近邻 $X_1, X_2, \cdots, X_k$ 落在球体 $S_X(\delta)$ $(\delta>0)$ 外的概率为 $(1-P(S_X(\delta)))^K$.

当 $K\to\infty$ 时, $(1-P(S_X(\delta)))^K\to 0$.

由于 δ 可以任意小, 所以, 当 $n\to\infty$ 时, X_k 落在以 X 为中心无限小区域中的概率趋近于 1. 也就是说 X_k 以概率 1 收敛于 X. □

3.3　算法步骤

本文利用遗传算法得到多组初始权重, 然后通过内省学习对每组权重进行迭代调整, 最后根据群决策理论中的简单多数原则对多组权重的检索结果进行综合判断与决策, 得到最终的群检索结果. 具体算法步骤如下:

步骤 1. 参数初始化, 设定专家个数为 s;

步骤 2. 构造群权重: 首先, 得到 s 组遗传算法分配的初始权重; 然后, 对每一组权重分别按照 GMU+GUD 的权重学习策略和式 (10) ∼ (11) 所示的权重调整方式进行迭代调整, 并对得到的权重进行归一化处理. 迭代调整后的 s 组权重共同构成式 (13) 所示的群权重 GW;

步骤 3. 将 GW 中的 s 组权重分别代入式 (14) 和 (15), 得到每位专家的个体决策结果 $Y_{j,q}$;

步骤 4. 由 s 位专家的个体决策 $Y_{j,q}$, 结合式 (16) 和 (17), 得到群决策结果 Y.

4　实验结果

4.1　实验方案

为验证方法的有效性, 选择 UCI 资源库[22] 的 10 种分类数据集进行分类实验, 各数据集的名称缩写、样本数、属性个数、类别数等信息参见表 1.

表 1　数据集信息

Table 1　Data sets for experiments

数据集	缩写	样本数	属性数	类别数
Fertility	FE	100	9	2
Connec. Bench	CB	208	60	2
Glass	GL	214	9	6
Ionosphere	IO	351	33	2
WDBC	WD	569	30	2
PIDD	PI	768	8	2
Statlog (Ve. Sil.)	VS	946	18	4
Mammographic Ma.	MM	961	5	2
Steel Plates Fau.	SPF	1 941	27	7
Statlog (Ima. Seg.)	IS	2 310	18	7

对于每个数据集, 进行五折交叉实验, 将数据集均分为 5 份, 每一折实验中, 取其中的 4 份作为训练集, 剩余的 1 份为测试集. 为了考察内省学习权重调整以及案例群检索策略对 CBR 分类准确率的影响, 分别采用以下 4 种 CBR 分类器进行对比实验:

1) 采用均权重和最近邻检索策略考察 CBR 分类器的性能, 记为 EWCBR 分类器;

2) 采用遗传算法分配权重和最近邻检索策略考察 CBR 分类器的性能, 记为 GACBR 分类器;

3) 采用内省学习调整权重和最近邻检索策略考察 CBR 分类器的性能, 记为 IGCBR 分类器;

4) 采用内省学习调整遗传算法权重和案例群检索策略考察 CBR 分类器的性能, 记为 GIGCBR 分类器.

4.2　参数设置

本文提到的算法在参数设置方面有如下要求:

遗传算法需保证全局搜索能力和快速收敛性; 群决策专家个数的设置需充分利用案例库的信息; 内省学习中的参数设置要保证对权重稳定有效的调整. 基于此, 本文算法中的参数设置情况是: 遗传算法分配特征属性权重时的交叉概率取 0.4, 变异概率取 0.05, 初始种群规模为 20, 迭代次数为 10. 案

例群检索策略中专家个数取 5. 内省学习策略采用 GMU+GUD 策略, 权重调整因子 Δ_i 为 0.015, 匹配属性的阈值 ξ_{match} 取 0.01.

4.3 实验结果

为了考察内省学习调整权重方法以及案例群检索策略对 CBR 分类准确率的影响, 分别采用 EWCBR、GACBR、IGCBR、GIGCBR 分类器对表 1 所示的 10 个数据集进行分类对比实验, 并记录五折交叉验证的平均值. 表 2 是采用这 4 种分类器对各数据集的平均分类准确率.

表 2 不同分类器的平均准确率 (%)

Table 2 Average accuracy (%) of different classifiers

	EWCBR	GACBR	IGCBR	GIGCBR
FE	81	81	83	84
CB	77.83	78.87	81.27	82.18
GL	67.26	69.62	72.41	75.7
IO	85.75	90.01	91.44	94.3
WD	95.09	95.1	95.45	96.66
PI	69.92	69.39	70.96	72
VS	68.92	71.52	74.94	76.48
MM	73.26	73.68	75.66	75.34
SPF	67.23	68.16	70.43	72.23
IS	97.27	97.23	97.84	98.14

由表 2 可见, 相对于 EWCBR, GACBR 在多数数据集上的准确率有所提高, 但是在少数数据集上 (FE、PI、IS、WD) 有所降低或提高较少; IGCBR 的准确率高于 EWCBR 和 GACBR, 而 GIGCBR 在这 4 个分类器中的准确率最高. 该结果说明就本文的大多数数据集而言, 遗传算法分配权重可以提高 CBR 的分类准确率, 但是由于遗传算法本身的局限性, 它对于某些数据集不能得到合理的属性权重, 导致 GACBR 在这些数据集上的准确率较低. 利用内省学习调整遗传算法得到的权重后, 分类器的准确率有所提高, 说明内省学习的迭代调整可以使属性权重的分配更为合理. 而利用群决策改进案例检索策略后, 可以更充分地利用案例库的潜在信息, CBR 分类器的准确率得到进一步提高.

图 2 是所有数据集的平均准确率. 由图 2 可见, 各分类器在所有数据集上的平均准确率由高到低依次是: GIGCBR、IGCBR、GACBR、EWCBR. 该结果表明内省学习调整遗传算法权重, 可以降低遗传算法自身局限性的影响, 得到更合理的权重分配, 有效提高了 CBR 分类准确率; 另外, 简单多数原则改进的案例群检索策略利用多组权重进行检索后, 可以充分利用案例库的信息, 能够进一步提高 CBR 的分类准确率.

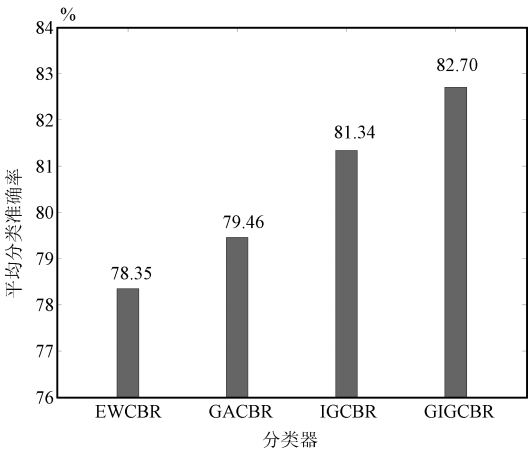


图 2 所有数据集的平均准确率

Fig. 2 Average accuracy (%) of all data sets

上述实验说明了遗传算法分配权重后, 利用内省学习原理对权重进一步调整可以得到更高的分类准确率. 此外, 群决策改进的案例检索策略能够进一步提高分类准确率.

5 结语

本文针对 CBR 中的特征属性权重分配和案例检索策略问题, 讨论了利用内省学习原理对权重进行迭代调整的方法, 并将群决策思想融入案例检索环节, 实现了一种改进的 CBR 分类方法, 对比实验结果表明该方法能够有效提高 CBR 分类准确率, 说明内省学习的引入能够避免遗传算法易陷入局部极小的缺陷, 得到更为合理的属性权重, 同时, 群决策改进的案例群检索能够充分利用多组权重蕴含的潜在信息, 充分挖掘案例库中的知识, 不仅提高了 CBR 的问题求解能力, 而且有助于其学习性能的改进. 以后的研究将进一步完善算法中相关参数的理论和实验研究, 以明确各种参数对方法的影响, 进而明确方法的适应性和可靠性.

References

1 Schank R C. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. New York: Cambridge University Press, 1982.

2 Shokouhi S V, Skalle P, Aamodt A. An overview of case-based reasoning applications in drilling engineering. *Artificial Intelligence Review*, 2014, **41**(3): 317−329

3 Aamodt A, Plaza E. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, 1994, **7**(1): 39−59

4 Pian Jin-Xiang, Chai Tian-You, Li Jie-Jia. Application of case-based reasoning and iterative learning to laminar cooling process control. *Acta Automatica Sinica*, 2012, **38**(12): 2032−2037

(片锦香, 柴天佑, 李界家. 案例推理及迭代学习在层流冷却控制中的应用. 自动化学报, 2012, **38**(12): 2032−2037)

5 Zhao Hong-Wei, Xie Yong-Fang, Jiang Zhao-Hui, Xu De-Gang, Yang Chun-Hua. An intelligently optimal setting approach based on froth features for level of flotation cells. *Acta Automatica Sinica*, 2014, **40**(6): 1086−1097
(赵洪伟, 谢永芳, 蒋朝辉, 徐德刚, 阳春华. 基于泡沫图像特征的浮选槽液位智能优化设定方法. 自动化学报, 2014, **40**(6): 1086−1097)

6 Chattopadhyay S, Banerjee S, Rabhi F A, Acharya U R. A case-based reasoning system for complex medical diagnosis. *Expert Systems*, 2013, **30**(1): 12−20

7 Yan A J, Wang W X, Zhang C X, Zhao H. A fault prediction method that uses improved case-based reasoning to continuously predict the status of a shaft furnace. *Information Sciences*, 2014, **259**(2): 269−281

8 Lao S I, Choy K L, Ho G T S, Tsim Y C, Poon T C, Cheng C K. A real-time food safety management system for receiving operations in distribution centers. *Expert Systems with Applications*, 2012, **39**(3): 2532−2548

9 Mantaras R L D, Plaza E. Case-based reasoning: an overview. *AI Communication*, 1997, **10**(1): 21−29

10 Subramanyam Rallabandi V P, Sett S K. Knowledge-based image retrieval system. *Knowledge-Based Systems*, 2008, **21**(2): 89−100

11 Chen Ling, Cheng Zhong-Hua, Zeng Hui-Yan. Study on case retrieval of case-based RCM analysis system. *Computer and Engineering and Design*, 2012, **33**(2): 581−585
(陈凌, 程中华, 曾慧燕. 基于案例推理的 RCM 分析系统中案例检索研究. 计算机工程与设计, 2012, **33**(2): 581−585)

12 Guo Xiao-Ping, Yuan Jie, Li Yuan. Feature space k nearest neighbor based batch process monitoring. *Acta Automatica Sinica*, 2014, **40**(1): 135−142
(郭小萍, 袁杰, 李元. 基于特征空间 k 最近邻的批次过程监视. 自动化学报, 2014, **40**(1): 135−142)

13 Cover T M, Hart P E. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 1967, **13**(1): 21−27

14 Ahn H, Kim K, Man I. Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system. *Expert Systems*, 2006, **23**(5): 290−301

15 Kim M, Lee S, Woo S, Shin D H. Approximate cost estimating model for river facility construction based on case-based reasoning with genetic algorithms. *KSCE Journal of Civil Engineering*, 2012, **16**(3): 283−292

16 Cox M T, Ram A. Introspective multistrategy learning: on the construction of learning strategies. *Artificial Intelligence*, 1999, **112**(1−2): 1−55

17 Bonzano A, Cunningham P, Smyth B. Using introspective learning to improve retrieval in CBR: a case study in air traffic control. In: Proceeding of the 2nd International Conference on Case-Based Reasoning. Berlin Heidelberg, USA: Springer, 1997. 291−302

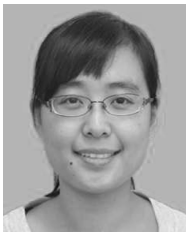
18 Zhang Z, Yang Q. Feature weight maintenance in case bases using introspective learning. *Journal of Intelligent Information Systems*, 2001, **16**(2): 95−116

19 Zhang Gui-Mei, Jiang Shao-Bo, Chu Jun. Affine registration based on chord height point and genetic algorithm. *Acta Automatica Sinica*, 2013, **39**(9): 1447−1457
(张桂梅, 江少波, 储珺. 基于弦高点和遗传算法的仿射配准. 自动化学报, 2013, **39**(9): 1447−1457)

20 Wan Jun, Xing Huan-Ge, Zhang Xiao-Hui. Algorithm of adjusting weights of decision-makers in multi-attribute group decision-making based on entropy theory. *Control and Decision*, 2010, **25**(6): 907−910
(万俊, 邢焕革, 张晓晖. 基于熵理论的多属性群决策专家权重的调整算法. 控制与决策, 2010, **25**(6): 907−910)

21 Xu Jiu-Ping, Chen Jian-Zhong. *The Theory and Methods of Group Decision Making with Its Realization*. Beijing: Tsinghua University Press, 2009.
(徐玖平, 陈建中. 群决策理论与方法及实现. 北京: 清华大学出版社, 2009.)

22 Frank A, Asuncion A. UCI machine learning repository [Online], Available: http://archive.ics.uci.edu/ml, May 15, 2014



张春晓 北京工业大学电子信息与控制工程学院博士研究生. 主要研究方向为案例推理及其应用. E-mail: zaichunfei123@emails.bjut.edu.cn
(**ZHANG Chun-Xiao** Ph. D. candidate at the College of Electronic Information & Control Engineering, Beijing University of Technology. Her research interest covers case-based reasoning and its applications.)

严爱军 北京工业大学电子信息与控制工程学院副教授. 2006 年于东北大学获得博士学位. 主要研究方向为人工智能, 过程建模与优化控制, 本文通信作者. E-mail: yanaijun@bjut.edu.cn
(**YAN Ai-Jun** Associate professor at the College of Electronic Information & Control Engineering, Beijing University of Technology. He received his Ph. D. degree from Northeastern University in 2006. His research interest covers artificial intelligence, process modeling, and optimization control. Corresponding author of this paper.)

王 普 北京工业大学电子信息与控制工程学院教授. 1988 年于中国矿业大学获得博士学位. 主要研究方向为信息处理, 智能控制. E-mail: wangpu@bjut.edu.cn
(**WANG Pu** Professor at the College of Electronic Information & Control Engineering, Beijing University of technology. He received his Ph. D. degree from China University of Mining Technology in 1988. His research interest covers information process and intelligent control.)