文章编号: 1001-0920 (2016) 07-1253-05 DOI: 10.13195/j.kzyjc.2015.0270

基于可信度阈值优化的案例推理评价分类方法

严爱军[a,b,c], 赵 辉[a,c], 王 普[a,c,d]

(北京工业大学 a. 电子信息与控制工程学院，b. 计算智能与智能系统北京市重点实验室,
c. 数字社区教育部工程研究中心，d. 城市轨道交通北京实验室，北京 100124)

摘 要: 为了提高案例推理 (CBR) 分类器的性能, 提出一种基于可信度阈值优化的 CBR 评价分类方法. 首先, 通过一种可降低时间复杂度的改进型可信度评价策略对案例重用得到的建议解的可信度进行计算; 然后, 通过遗传算法 (GA) 对可信度阈值进行迭代寻优; 接着, 根据得到的优化阈值将目标案例及其建议解划分为可信集或不可信集; 最后, 对不可信集按多数重用原则进行分类结论的调整, 从而实现可信的 CBR 评价分类. 对比实验表明, 改进的可信度评价策略能有效提高分类性能, 从而可提高 CBR 分类器的决策与学习能力.

关键词: 案例推理；评价分类；可信度阈值；遗传算法

中图分类号: TP18 **文献标志码:** A

Trustworthiness evaluation method with threshold optimization for case-based reasoning classification

YAN Ai-jun[a,b,c], *ZHAO Hui*[a,c], *WANG Pu*[a,c,d]

(a. College of Electronic Information and Control Engineering，b. Beijing Key Laboratory of Computational Intelligence and Intelligent System，c. Engineering Center of Digital Community of Ministry of Education，d. Beijing Laboratory for Urban Mass Transit，Beijing University of Technology，Beijing 100124，China．Correspondent：YAN Ai-jun，E-mail：yanaijun@bjut.edu.cn)

Abstract: To improve the performance of a case-based reasoning (CBR) classifier, a trustworthiness evaluation method with threshold optimization for case-based reasoning classification is proposed. Firstly, an improved trustworthiness evaluation(TE) strategy is adopted to calculate the trustworthiness value of the suggested solutions achieved in reuse step. Then, the optimal threshold value of the trustworthiness is obtained by using the genetic algorithm(GA). Subsequently, the target case and its suggested solution is divided into the trustworthy set and the untrustworthy set in accordance with this threshold value. Finally, the majority reuse strategy is adopted to adjust the suggested solutions in the untrustworthy set so as to fulfill an overall CBR evaluation classification process. The experimental results show that the proposed method can effectively increase the classification performance and improve the learning ability for a CBR classifier.

Keywords: case-based reasoning；evaluation classification；trustworthiness threshold；genetic algorithm

0 引 言

案例推理 (CBR) 是一种起源于认知心理学的问题求解和机器学习方法, 它解决问题可依次按照案例的检索、重用、修正和存储 4 个环节进行, 即 4R 循环[1]. 目前, CBR 的应用领域非常广泛[2-5], 但大多关注如何提高其问题求解的综合性能[6-7]. 以模式分类任务为例, 提高 CBR 的分类准确率和降低时间复杂度是衡量 CBR 分类器性能高低的两个重要指标[8-10].

CBR 分类器根据目标案例与源案例的相似程度得到目标案例的类别, 并存储正确分类的案例以完成分类和学习过程[11]. 在此过程中, 评价分类结论是否可信的方法影响着 CBR 分类是否准确[12]. 对此, 基于信心评估的评价策略得到了广泛关注[13-15]. 文献 [15] 针对 CBR 的重用过程提出一种分类建议解的可信度评价 (TE) 方法, 通过计算目标案例分属于所有类别的可信度大小, 从中选择可信度值最大的类别作为目标案例的分类结论. 虽然 TE 方法可提高 CBR 的求解质量, 但时间复杂度较高, 限制了整体性能的提升.

收稿日期: 2015-03-08；**修回日期:** 2015-05-19.

基金项目: 国家自然科学基金项目 (61374143)；北京市自然科学基金项目 (4152010).

作者简介: 严爱军 (1970−), 男, 副教授, 博士, 从事人工智能及应用、过程建模与优化控制等研究；赵辉 (1988−), 男, 博士生, 从事人工智能及应用的研究.

为了简化评价策略的时间复杂度, 本文引入可信度阈值参与问题求解. 由于阈值的选取会直接影响分类器的综合性能, 而且人工设定[16]的随意性较大, 如何得到合理的可信度阈值需要进一步研究.

为了提高CBR分类器的分类准确率和效率, 本文提出一种基于可信度阈值优化的CBR评价分类方法(TETOCBR). 该方法可改进可信度评价的计算过程, 并利用遗传算法(GA)进行可信度阈值的迭代寻优. 根据优化的可信度阈值将建议类别划分为可信集和不可信集, 再对不可信集中的类别进行修正调整. 对比实验结果表明了本文方法的有效性.

1 CBR分类的问题分析

当有待分类的目标案例T出现时, CBR分类器从案例库中检索出与目标案例T最相似的源案例X, 并将X的类别作为目标案例T的分类结论Y_T. 为了得到分类结论Y_T, 需要评估案例库中每个源案例$X_i(i=1,2,\cdots,m)$与目标案例T的相似度, 可采用基于欧氏距离的相似度评估策略去计算[17], 即

$$\begin{cases} s_i = s(T,X_i) = 1 - D(T,X_i),\ i=1,2,\cdots,m; \\ D(T,X_i) = \sqrt{\sum_{j=1}^{n}(t_j - x_{ij})^2}. \end{cases} \tag{1}$$

其中: $s_i \in [0,1]$为第i个源案例X_i与目标案例T的相似度; $D(T,X_i) \in [0,1]$为欧氏距离; x_{ij}, t_j分别表示第i个源案例与目标案例中第j个属性的归一化特征值. 在得到分类结果Y_T的过程中若出现下面两种情况, 则有可能导致分类失败:

1) 目标案例T与某个类别的源案例问题描述X的距离$D(T,X)$最小, 但两案例之间的相对距离仍然较远, 此时若将X与T视为同一类别, 则可能会导致结论不可信;

2) 目标案例T与多个类别的源案例问题描述的距离均比较接近, 即T位于不同类别源案例的重叠区域, 此时的分类建议解也有可能不可信.

在上述情况下, 直接重用该结论会影响CBR的分类准确率与可信度, 因而, 亟需确定一个评价分类建议解是否可信的策略, 用以指导结论的重用.

目前出现的评价策略包括对结论的信心评估[13]、可信区域划分[14]和可信度值计算等[15]. 其中: 信心评估的投票权重难以确定; 可信区域划分适用于对结论的定性评价; 而可信度值的计算效率不高, 并且对不可信的结论没有作后续修正处理, 导致分类准确率较低. 因此, 针对此问题, 本文引入可信度阈值参与到CBR分类过程.

2 基于可信度的案例评价方法

本节设计了一种改进型的CBR评价分类器, 并描述简化的可信度计算过程、可信度阈值的遗传优化方法以及不可信结论的修正.

2.1 CBR评价分类器的功能

在传统CBR模型的案例重用环节后, 增加案例评价环节与阈值寻优过程, 可得到一种CBR评价分类器. 其主要功能如下: 首先, 采用基于欧氏距离的相似度评估策略得到与目标案例最为相似的K个源案例; 然后, 使用最大相似度重用方式($K=1$)获得目标案例T的建议解; 接着利用所提出的案例评价策略对此建议解的可信度进行评价, 并根据优化的阈值, 将建议解划分为可信解和不可信解, 当出现不可信解时, 对结论进行调整; 最后, 将目标案例集及相应的确认解或可信解存储于案例库中, 用于下一个分类问题的推理求解.

2.2 可信度计算

下面通过计算建议解的可信度数值[15]来判断这个解是否可信. 文献[15]采用了5个可信度的评价指标, 并分别计算了目标案例属于所有类别的$2K$个近邻(K个同类别的近邻案例和K个异类的近邻案例), 时间复杂度较高. 为了降低算法的时间复杂度, 选取了其中的3个可信度评估指标进行计算, 且减少了近邻的数量, 将$2K$改为K, 具体算法如下: 首先, 利用最大相似度重用方式获得目标案例T的建议解, 记为类别c; 然后, 以类别c为依据对K个建议解进行归类, 与类别c不同的p个源案例描述组成异类近邻集(DN), 分别记为$\{\mathrm{DN}_1,\mathrm{DN}_2,\cdots,\mathrm{DN}_p\}$, 与类别$c$相同的$q$个源案例描述组成同类近邻集(SN), 分别记为$\{\mathrm{SN}_1,\mathrm{SN}_2,\cdots,\mathrm{SN}_q\}$, 其中$p+q=K$; 最后, 利用可信度指标计算得到结论的可信度值. 该可信度指标定义如下:

1) 异类平均距离. 计算目标案例T与DN中的p个异类案例间的平均欧氏距离

$$M_1 = \frac{1}{p}\sum_{i=1}^{p} D(T,\mathrm{DN}_i). \tag{2}$$

2) 同类平均相似度. 计算目标案例T与SN中的q个同类案例间的平均相似度

$$M_2 = \frac{1}{q}\sum_{j=1}^{q} s(T,\mathrm{SN}_j). \tag{3}$$

3) 目标案例T与同类、异类案例的相似度比率. 首先, 计算出目标案例T与同类案例间的相似度和, 再计算目标案例T与异类案例间的相似度和; 然后求

出两者相似度和的比值

$$M_3 = \frac{\sum_{j=1}^{q} s(T, \mathrm{SN}_j)}{\sum_{i=1}^{p} s(T, \mathrm{DN}_i)}. \tag{4}$$

为了消除3个指标的量纲差别, 将式(2)～(4)中的 $M_1 \sim M_3$ 统一归一化为区间[0, 1]中的数值, 然后对这3个指标求平均值, 得到目标案例建议解的可信度大小, 即

$$\lambda = \frac{1}{3}\sum_{k=1}^{3} M_k. \tag{5}$$

为了定量判别建议的类别 c 是否可信, 可事先确定一个可信度阈值 $T_\lambda \in (0,1)$. 当 $\lambda \geqslant T_\lambda$ 时, 类别 c 为可信解, 将其作为目标案例的类别, 并划入可信集(TS)中; 否则, 作为不可信解并划入不可信集(UTS)中去修正. 至此, 完成对建议解类别 c 的评价过程. 其中: 可信集中的案例可直接重用当前结论, 并将该目标案例的描述与分类结论存储于历史案例库中; 而对于不可信集中的案例, 采用多数重用法进行结论的调整, 并将调整后的结论与案例描述存储到历史案例库中, 即

$$Y_T = \arg\max_{c}(\mathrm{Num}(Y_T = c)). \tag{6}$$

其中Num()表示 K 个近邻案例类别为 c 的个数.

注 1　时间复杂度分析. 改进的案例评价策略通过式(1)～(3)对检索得到的 K 个近邻案例计算其可信度, 其时间复杂度为

$$T_1 = O(2 \times K \times n)$$

其中 n 为特征属性的个数; 而传统的可信度计算过程的时间复杂度为

$$T_2 = N_c * O(2 \times 2K \times n + m \times n),$$

其中: N_c 表示类别个数, m 表示历史案例的个数. 因此 $T_2 > T_1$. 由此可知, 改进的可信度的评价策略能够减少计算的时间复杂度, 从而可提高分类器的运行效率.

2.3　可信度阈值优化

采用GA进行可信度阈值的优化, 主要步骤如下.

Step 1: 定义适应度函数

$$F = \frac{N'_{\mathrm{uts}} + N'_{\mathrm{ts}}}{N_{\mathrm{uts}} + N_{\mathrm{ts}}}. \tag{7}$$

其中: N_{uts}, N_{ts} 分别表示根据当前可信度阈值 T_λ 划分出的不可信集UTS和可信集TS中案例的个数; N'_{uts} 表示在 N_{uts} 个案例中利用多数重用修正能够正确分类的个数; N'_{ts} 表示在 N_{ts} 个可信案例中实际正确分类的案例个数. 该适应度函数旨在确保能够得到最优的可信度阈值.

Step 2: 选择. 设种群大小为 P_s, 其中第 i 个阈值个体的适应度为 F_i, 则该阈值被选择的概率为

$$P_i = \frac{F_i}{\sum_{i=1}^{P_s} F_i}. \tag{8}$$

Step 3: 交叉. 根据预先设定的交叉概率 P_c, 采用单点交叉方式生成两个新的阈值个体.

Step 4: 变异. 以预先设置好的变异概率 P_m 对阈值个体的二进制编码的某位进行取反操作.

按照设定的迭代次数Iter进行迭代寻优, 可得到可信度阈值 T_λ; 然后, 根据 T_λ 即可将目标案例划分为可信集和不可信集, 并在不可信集中采用多数重用方式($K > 1$)完成整体修正.

注 2　算法收敛性分析. 设 F_t 表示第 t 代阈值种群 $P_o(t)$ 的最优适应值, F^* 表示全局最优值. 由于GA的迭代寻优过程是齐次遍历的马尔可夫链[18], 第 t 代的阈值种群 $P_o(t)$ 从任意状态向 F_t 的转移概率大于0, 而从 F_t 向不含最优适应值的转移概率为0, 于是满足 $\lim\limits_{t\to\infty} P\{F_t = F^*\} = 1$[19], 故保留具有最大分类准确率所对应的可信度阈值能够依概率收敛到全局最优解 F^*.

3　实验结果与分析

3.1　实验设计

实验数据选用UCI资源库中的10个分类数据集, 基本信息如表1所示.

表 1　实验数据集基本信息

序号	数据集	案例数	属性个数	类别数
1	Yeast	1 299	8	4
2	Heart	297	13	5
3	Glass	214	9	6
4	Statlog (Heart)	270	12	2
5	German Credit	1 000	24	2
6	Seeds	210	7	3
7	Page Blocks	5 473	10	5
8	Iris	150	4	3
9	Wine	178	13	3
10	Image Segment	2 310	18	7

采用5折交叉验证方法, 将每一个数据集按平均原则分为5份, 其中的4份存储于案例库中, 余下的1份用于分类测试, 轮流测试完毕后完成5折交叉实验. 通过以下两个实验进行验证.

实验 1　确定参数. 1) 确定 K 近邻检索策略中近邻个数 K 和可信度阈值 T_λ. 当分别取 $T_\lambda = 0.3, 0.5, 0.8, 0.9$ 时, 观察 $K = 5, 7, 9, 11$ 的分类准确率, 最高准确率对应的 K 和 T_λ 为所求; 2) 确定GA的种群

数量 P_s. 设置不同种群数量 P_s, 观察分类准确率的收敛性, 当准确率不再增长时, 收敛速度最快的种群数量 P_s 为所求.

实验 2 对比实验. 对比TETOCBR与径向基函数网络(RBFNN)、支持向量机(SVM)、CBR, 以及采用TE 策略的CBR(TECBR)方法的分类准确率.

实验中一些参数的设置情况是: RBFNN中的径向基函数采用Gaussian函数 $y = \mathrm{e}^{-x^2}$; TETOCBR中, GA寻优阈值时二进制编码位数为7, 变异概率 $P_m = 0.05$, 交叉概率 $P_c = 0.4$, 迭代次数 $\mathrm{Iter} = 20$.

3.2 确定参数

在可信度计算中由式(2)～(5)可知, 检索步骤中 K 值的选取影响着可信度值的计算, 进而影响着TECBR的分类准确率. 当分别取 $K = 5, 7, 9, 11$ 和 $T_\lambda = 0.3, 0.5, 0.8, 0.9$ 时, 10个数据集的平均分类准确率如图1所示. 由此可见, 当 $K = 7$, $T_\lambda = 0.3$ 时平均分类准确率达到了最高.

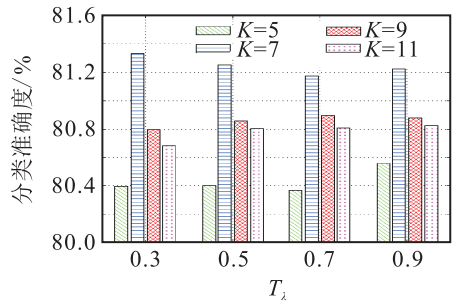

图 1 不同 K 和 T_λ 下TETCBR的平均准确率

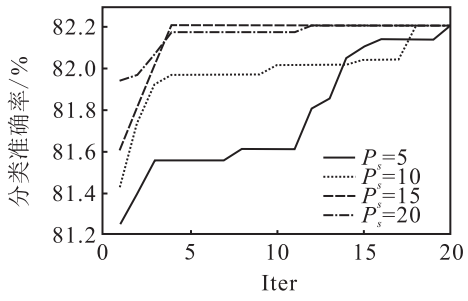

图 2 不同种群数量情况下的平均分类准确率

在TETOCBR阈值寻优环节中, GA中不同种群 P_s 对阈值寻优的收敛速度和分类准确率的影响如图2所示. 在设定的迭代次数Iter内, 10个数据集的平均分类准确率均达到了收敛, 随着种群数量的增加, 平均分类准确率达到收敛的迭代次数Iter, 有减少的趋势, 且在 $P_s = 15$ 时, 收敛速度最快.

3.3 对比实验

下面主要考察在加入GA进行阈值优化后, TETOCBR分类器在分类性能上的优越性. 根据实验2拟定的方案, 通过GA寻优分别得到10个数据集的最优阈值; 根据此阈值计算出TETOCBR的分类准确率, 并与传统CBR、TECBR、RBFNN和SVM分类器的平均分类准确率进行对比, 结果如表2所示. 由表2可以看出, 平均分类准确率的大小顺序依次为TETOCBR、TECBR、SVM、CBR和RBFNN, 其中, TETOCBR的平均分类准确率达到了82.36%, 说明本文方法在提高分类准确率方面具有优越性.

表 2 TETOCBR与其他方法的准确率对比

序号	CBR 准确率/%	RBFNN 准确率/%	SVM 准确率/%	TECBR 准确率/%	TETOCBR 准确率/%	TETOCBR 阈值
1	54.36	60.43	59.73	58.21	60.07	0.95
2	51.16	55.89	56.56	56.56	59.44	0.08
3	67.27	68.69	56.54	64.96	67.36	0.98
4	75.19	81.11	83.7	81.11	81.71	0.22
5	68.5	73	76.8	71.7	71.7	0.05
6	92.86	90.95	93.33	93.33	95.61	0.02
7	95.49	94.76	92.75	94.35	95.58	0.03
8	96	96.67	96.67	96	96.53	0.34
9	94.9	56.74	98.31	96.01	98.82	0.33
10	97.27	87.58	92.68	96.36	96.82	0.02
平均值	79.3	76.58	80.71	80.86	82.36	

4 结 论

为了提高CBR分类器的分类准确率和效率, 本文在案例重用后增加了案例评价环节, 提出一种基于可信度阈值优化的CBR评价分类方法. 改进的可信度评价策略在保证分类准确率不降低的情况下, 可以有效减少时间复杂度; 采用遗传算法对可信度阈值进行寻优, 并对分类结论作整体修正, 使得分类准确率比常见的分类器要高, 避免了人工设定阈值的主观性, 有效提高了CBR分类的综合性能.

由于遗传算法寻优的随机性, 并不能保证本文方法对所有分类问题均能适用, 下一步的研究重点在于根据具体分类问题的特点, 研究CBR与其结合的机理. 在此基础上, 给出改进的可信度阈值优化方法, 划分出更为合理的可信集与不可信集, 从而促进CBR分类器准确率的进一步提高.

参考文献(References)

[1] Aamodt A, Plaza E. Case-based reasoning: Foundational issues, methodological variations, and system approaches[J]. AI Communications, 1994, 7(1): 39-59.

[2] 严爱军, 柴天佑, 王普. 基于案例推理的竖炉故障预报系统[J]. 控制与决策, 2008, 23(2): 177-181.
(Yan A J, Chai T Y, Wang P. Fault prediction system using case-based reasoning for shaft furnace status[J]. Control and Decision, 2008, 23(2): 177-181.)

[3] Lee S W, Seo K K. Intelligent fault diagnosis based on a hybrid multi-class support vector machines and case-based

reasoning approach[J]. J of Computational and Theoretical Nan Science, 2013, 10(8): 1727-1734.

[4] Chuang C L. Application of hybrid case-based reasoning for enhanced performance in bankruptcy prediction[J]. Information Sciences, 2013, 236(7): 174-185.

[5] Yan A J, Chai T Y, Yu W, et al. Multi-objective evaluation-based hybrid intelligent control optimization for shaft furnace roasting process[J]. Control Engineering Practice, 2012, 20(9): 857-868.

[6] 韩敏, 沈力华. 基于FCM与神经网络的案例推理方法[J]. 控制与决策, 2012, 27(9): 1421-1424.
(Han M, Shen L H. Case-based reasoning based on FCM and neural network[J]. Control and Decision, 2012, 27(9): 1421-1424.)

[7] Xu X, Wang K, Ma W, et al. Improving the reliability of case-based reasoning systems[J]. Int J of Computational Intelligence Systems, 2010, 3(3): 256-265.

[8] Guo Y, Hu J, Peng Y H. Research on CBR system based on data mining[J]. Applied Soft Computing, 2011, 11(8): 5006-5014.

[9] Li Y, Shiu Simon C K, Pal S K. Combining feature reduction and case selection in building CBR classifiers[J]. IEEE Trans on Knowledge and Data Engineering, 2006, 18(3): 415-429.

[10] Salamó M, López-Sánchez M. Rough set based approaches to feature selection for case-based reasoning classifiers[J]. Pattern Recognition Letters, 2011, 32(2): 280-292.

[11] Rezvan M T, Hamadani A Z, Shalbafzadeh A. Case-based reasoning for classification in the mixed data sets employing the compound distance methods[J]. Engineering Applications of Artificial Intelligence, 2013, 26(9): 2001-2009.

[12] Massie S, Craw S, Wiratunga N. When similar problems don't have similar solutions[C]. Proc of 7th Int Conf on Case-Based Reasoning. Heidelberg, 2007, 4626: 92-106.

[13] Muhlbaier M D, Topalis A, Polikar R. Learn^{++}, NC: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes[J]. IEEE Trans on Neural Networks, 2009, 20(1): 152-168.

[14] Chua J J, Tischer P E. Determining the trustworthiness of a case-based reasoning solution[C]. Proc of the Int Conf on Computer intelligent for Modeling, Control and Automation. Gold Coast, 2004: 952-962.

[15] Garcia F, Oroaco J, Gonzàlez J, et al. Assessing confidence in cased based reuse step[C]. Proc of the 10th Int Conf of the Catalan Association for Artificial Intelligence. Amsterdam, 2007: 161-168.

[16] 赵辉, 严爱军, 王普. 提高案例推理分类器的可靠性研究[J]. 自动化学报, 2014, 40(9): 2029–2036.
(Zhao H, Yan A J, Wang P. On improving reliability of case-based reasoning classifier[J]. Acta Automatica Sinica, 2014, 40(9): 2029-2036.)

[17] Liao Z, Mao X, Hannam P M, et al. Adaptation methodology of CBR for environmental emergency preparedness system based on an improved genetic algorithm[J]. Expert Systems with Applications, 2012, 39(8): 7029-7040.

[18] Othman M A, Zaer S A, Adnan M A, et al. A robust and efficient genetic algorithm for solving a chemical reactor problem: Theory, application and convergence analysis[J]. Trans of the Institute of Measurement and Control, 2012, 34(5): 594-603.

[19] Greenhalgh D, Marshall S. Convergence criteria for genetic algorithms[J]. Siam J on Computing, 2000, 30(1): 269-282.

(责任编辑：滕 蓉)