记忆和遗忘策略改进的案例推理方法

张春晓[1] 赵 辉[2]
(北京电子科技职业学院自动化工程学院 北京 100176)[1] (清华大学信息技术研究院 北京 100084)[2]

摘 要 在案例推理(Case-Based Reasoning,CBR)中,随着案例库规模的不断扩大,当检索的时间成本超过案例增多带来的准确率收益时,会出现"覆没问题"。从认知科学的角度研究一种具有选择记忆和有意遗忘功能的案例库维护方法,对新案例进行选择性保存,并对旧案例进行有意识删除。对比实验结果表明了所提方法的有效性,选择记忆和有意遗忘策略在提高分类准确率的基础上,能够显著降低时间复杂度和空间复杂度,从而使CBR的求解性能得以提高。

关键词 案例推理,案例库维护,记忆,遗忘
中图法分类号 TP18 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.048

Case Base Reasoning Method Improved by Memory and Forgetting Strategy

ZHANG Chun-xiao[1] ZHAO Hui[2]
(College of Automatization Engineering,Beijing Polytechnic,Beijing 100176,China)[1]
(Research Institute of Information Technology,Tsinghua University,Beijing 100084,China)[2]

Abstract In the case-based reasoning (CBR),with the continuously growing of the size of case base,there may be so called "swamping problem" when the time cost of retrieval exceeds the benefit of the accuracy. From the perspective of cognitive science,a case base maintenance method with the ability of selective memory and intentional forgetting was proposed,which can selectively save the new cases and intentionally delete the old cases. The contrast experiments show the effectiveness of the proposed method. The selective memory and intentional forgetting policy can significantly reduce the time and space complexity,and preserve or improve the accuracy of CBR classifier,thus improve the performance of CBR.

Keywords Case-based reasoning,Case base maintenance,Memory,Forgetting

CBR作为一种重要的问题求解范式,基于"相似问题具有相似解"的认知假设,通过检索存储于案例库中的相似案例来解决当前的新问题[1],在预测[2]、分类与诊断[3-4]以及应急处理[5]等领域得到实际应用。传统的观点认为,案例库中的案例越多,检索到相似案例的可能性越大,有利于提高问题求解的准确性。然而,当检索的时间成本超过检索的收益时,就会带来所谓的"覆没问题"[6]。另外,当案例库中存在有害或冗余的案例时,问题求解的整体性能会受到影响[7]。因此,迫切需要寻找合适的案例维护方法,在保持较小案例库规模的同时保证较高的问题求解准确率,以提高系统的整体性能。

许多学者对案例库维护方法进行了研究,大致可分为面向求解效率[8-9]和面向整体性能[10-12]的维护方法。对于前者,基本策略是直接或间接控制知识库的规模,如果知识库的容量超过某个预设值,随机性地直接删除一条知识[8]。文献[9]将可拓学与案例推理结合后形成一种知识推理方法,进行案例的添加、删除和替换,以提高检索效率。对于后者,Smyth和Mckenna在基于覆盖度和可达度对案例能力进行分类的基础上,增加案例局部能力的相关性作为案例能力分类的因素,建立了能力组,得到案例库的全局能力[10]。近几年,有一些研究者利用聚类方法将案例库划分为较小的新的案例库,以降低案例检索的成本,并指导案例删除[11-12]。总的来说,上述两类典型的维护方法中,面向效率的维护方法虽然可以阻止问题求解效率的下降,但由于较少考虑问题的求解能力,可能会删除一些有用的关键案例,以致无法求解某些问题;而面向整体性能的维护方法中,由于案例能力分类或案例库聚类都需耗费大量的时间,并且案例的能力分类或聚类正确与否会直接影响删除策略的有效性,实际上限制了CBR的整体性能。因此,案例库的维护策略仍需进一步探讨。

针对上述问题,考虑到CBR与认知科学的相关性,本文将选择记忆和有意遗忘引入CBR的案例库维护中,提出了一种基于"遗忘值"的案例维护方法,即对新案例进行选择性保存,对旧案例进行有意识的删除,以保证案例库在更新的同时

到稿日期:2016-09-05 返修日期:2016-12-24 本文受科技类博士资助课题:记忆与遗忘改进的案例推理分类器研究(YZKB2015010),校级骨干教师培育项目(CJGX2016-JX-26/004),2015校内重点课题:工业控制网络控制命令语义分析(YZK2015044),电气自动化技术教学团队建设(CJGX2016-JX-26/023)资助。

张春晓(1983-),女,博士,讲师,主要研究方向为案例推理及其应用、人工智能等;赵 辉(1988-),男,博士,主要研究方向为人工智能、轨道交通等。

可以控制案例增长的数量。实验结果证明了问题求解的准确率或效率得以提高。

1　具有记忆和遗忘功能的 CBR 模型

在众多描述 CBR 的模型中，应用最为广泛的是由 Aamodt 和 Plaza 提出的“4R”循环[1]（见图 1）及其变型。传统的 4R 包括以下 4 个环节：

检索（retrieve）最相似的案例；

重用（reuse）检索到的结论，尝试解决新问题；

修正（revise）建议的解答；

保存（retrain）新问题和修正的解为一条新案例。

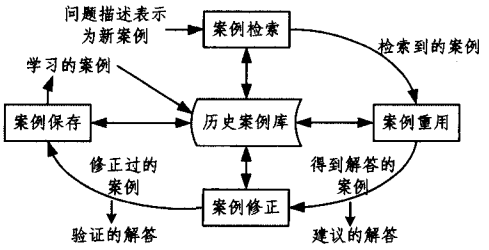

图 1　传统的 CBR 循环

由于 CBR 采用增量学习方式，随着新案例的不断保存，案例库的规模不断扩大。虽然检索到相似案例的可能性增加，但是随之而来的“覆没问题”可能导致案例检索的效率降低，同时案例库中的冗余案例和有害案例不仅增加了案例检索的空间复杂度，也可能损害案例检索的准确率。因此，有必要寻求一种合理的案例库维护方法来提高 CBR 的整体性能。

追溯 CBR 研究的源泉，大部分灵感来源于认知科学对人类记忆的探索[13]。许多研究者使用遗忘对认知模型进行维护[14-16]，他们分别将遗忘用于记忆的不同阶段，使认知模型在任务交换[14]、启发式推理[15]、反应度[16]等许多方面的能力得到提高。本文从记忆和遗忘的角度研究案例维护方法，在传统的“4R”循环的保存环节之后增加回顾环节，该环节主要根据内省学习原理更新检索到的 K 个近邻案例的“遗忘值”，并根据相应的遗忘策略对其进行删除操作。保存环节则根据相应的记忆策略对新案例进行有选择的保存，并对新保存的案例赋予初始遗忘值。图 2 展示了扩展后的具有记忆和遗忘功能的 CBR 循环。

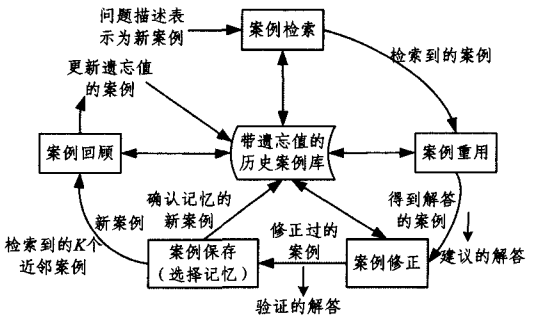

图 2　具有记忆和遗忘功能的 CBR 循环

图 2 中历史案例库中的源案例表示为如下的三元组形式：

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

其中，p 是历史案例的总数；X_k 是每一条源案例的问题描述；Y_k 是该案例的解答；$F_k\in[0,1]$，是该案例的遗忘值，将用于后面的回顾环节，案例库中的每个案例的初始“遗忘值”为 0.5。该初始化的原理与文献[17]对案例有用度的设定类似，是假定所有案例在用于案例分类之前具有相同的遗忘程度；当一条案例被检索出并用于目标问题分类后，该案例的遗忘值会根据其分类效果进行更新。

设目标案例的问题描述集为 X，利用具有记忆和遗忘功能的 CBR 系统推理求出该描述相应的类别 Y 需包括以下环节：

首先按 K-近邻（K-Nearest Neighbor，KNN）检索策略[18]检索出 K 个与新问题最相似的案例，然后统计这 K 个案例的类别，选出其中案例数最多的类别作为检索到的类别。

对于分类问题，重用环节直接将检索到的类别作为目标问题的建议类别。但当 CBR 用于一般问题的求解时，有时还需要改编或修正检索到的结论。

经过上述检索和重用后，目标问题和推理所得的建议类别可被表示为一条新案例。案例保存环节，将根据下文提出的记忆策略确定新案例是否被存储到历史案例库中。

最后的回顾环节采用提出的遗忘策略对检索环节得到的 K 个近邻案例的遗忘值进行更新，并将其中符合删除策略的案例从历史案例库中删除。

至此，历史案例库中案例记录的总数更新为 p' 个：

$$p'=p+p_{New}-p_{Del} \tag{2}$$

其中，当新案例被确认保存时，p_{New} 为 1；否则，p_{New} 为 0；p_{Del} 表示在当次循环的案例回顾环节中被删除的案例个数。

2　案例的记忆和遗忘策略

根据认知心理学理论，人类通过记忆或遗忘将过去的经历变成记忆。Markovitch 和 Scott 曾经讨论了记忆和遗忘在学习中的作用，他指出记忆可能是有害的，即使储存的是正确的知识，遗忘也可以对系统性能产生持续的提高[8]。对认知科学的研究也是 CBR 发展的基础之一，本节将从记忆和遗忘的角度研究 CBR 的问题求解和学习过程，讨论新案例的保存、旧案例的删除等案例维护工作。

2.1　案例的记忆策略

新案例的保存固然可以提高检索到相似案例的可能性，但是持续地增加新案例也会使案例库规模急剧扩大，这必将增加案例检索的复杂度，甚至降低检索的准确度，产生所谓的“覆没问题”。因此，对新案例不能一味增加，而应该根据某个策略进行选择性保存。本文提出以下记忆策略作为保存新案例的判断条件：

（1）新案例被错误分类。这说明当前案例库的信息不足以正确判断新案例的类别，本文将修正后的新案例保存到案例库中，以为日后的案例求解增加相关信息。

（2）新案例被正确分类，但是与最相似的旧案例的相似度过小。如果二者的相似度过小，虽然当前案例被正确分类，但是由于分类信息不是很充分，不能确定当前的案例库是否会对日后的案例求解产生误导，因此需要保存新案例来增加砝码。可以设置合理的相似度阈值 $\delta(\delta\in[0,1])$，当最大相似度小于该阈值时即保存新案例。

当新案例满足上述两者之一时便将其存到案例库中，并添加初始遗忘值。为了保证公平性，与案例库原有的其他案

例类似，新加入案例的初始遗忘值为 0.5。

新案例的记忆策略可被概括如下：

If（新案例被误分类 ‖ 新案例被正确分类，但是其与旧案例的最大相似度$<\delta$） (3)

Then 新案例被保存到历史案例库

2.2 案例的遗忘策略

作为与保持相矛盾的另一面，遗忘参与人类的记忆过程。记忆的内容不能保持或者提取有困难时就是遗忘，遗忘有多种情况，其中一种称为有意遗忘。有意遗忘是指对指定材料进行定向遗忘，是在意识的参与下对要求忘记的材料进行有意识的遗忘[19]。根据有意遗忘的理论，本文将某些无关或有害的案例标记为“遗忘项”，并在案例检索之前将其删除，这样可以降低或消除无关信息的干扰，提高案例检索的准确率；同时，案例删除可以保证案例库的容量在合理的范围，降低 CBR 循环过程的复杂度。

本文设计了一种基于遗忘值的遗忘策略，其由 3 个子策略构成，分别是遗忘触发策略、遗忘值更新策略及案例删除策略。

遗忘触发策略根据内省学习的原理分为 3 种不同的方式，分别是失败驱动、成功驱动和混合驱动。内省学习将内省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发现问题，形成修正自身的学习目标，由此改进自身处理问题的方法。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[20-21]。基于内省学习的案例遗忘触发策略，就是在问题求解性能的基础上决定何时触发案例的遗忘策略，包括以下 3 种触发策略：

1）失败驱动策略。当新案例被错误分类时，启动案例遗忘策略。

2）成功驱动策略。当新案例被正确分类时，启动案例遗忘策略。

3）混合驱动策略。无论对新案例是否正确分类，都启动遗忘策略。

本文将在后面的实验中比较 3 种触发方式对案例库性能的影响，并分析何种触发策略更有效。

一旦遗忘策略被触发，首先要更新选中的案例的遗忘值。为了简化计算，遗忘值的更新策略不是针对整个案例库，而是仅应用于案例检索得到的 K 个近邻案例。对于案例，按照下式更新其“遗忘值”：

$$F_i(t+1)=F_i(t)+\alpha \cdot r_i \cdot F_i(0), i=1,2,\cdots,K \quad (4)$$

其中，α 是遗忘增强因子，根据强化学习理论[22]，一般取为0.1或 0.2；r 是奖惩函数，当检索得到的相似案例 i 与新案例的类别相同时取值为 -1，否则为 1；$F_i(0)$是案例 i 的初始遗忘值，$F_i(t)$是当次迭代更新前的遗忘值，$F_i(t+1)$是当次迭代更新后的遗忘值。可见，如果案例对分类的影响是正面的，则遗忘值会下降，下降越多表示该案例越有用，也越不可能被遗忘；否则，遗忘值会上升，并且遗忘值越大表示该案例越应该被遗忘。

然后，案例删除子策略根据遗忘值对 K 个近邻案例进行删除或保持操作。如果更新后的遗忘值是下降的，则表示该案例是有用的，需保留，并且本次遗忘值的下降可以抵消日后误分类造成的遗忘值上升；如果更新后的遗忘值是增加的，则表示该案例会造成误分类或逐渐失去作用，当更新后的案例遗忘值超出某个阈值 ξ（$\xi\in(0.5,1]$）时便将该案例删除，否则继续保持。具体地，当案例更新后的遗忘值满足下式时，则可将其删除：

$$F_i(t+1)\geqslant\xi \quad (5)$$

如果确定了遗忘触发策略，ξ 的取值就可以进一步控制案例删除的速度。ξ 值越大，则允许遗忘值增加的范围越大，对分类错误的案例会越宽容，因此删除速度也越慢；否则，删除速度加快。

2.3 具有记忆和遗忘功能的 CBR 的分类算法

具有记忆和遗忘功能的 CBR 的分类算法可概括为以下步骤：

Step1 数据分为训练集和测试集。

Step2 训练集作为案例库，并为案例库中的每个案例分配初始遗忘值。

Step3 顺序取出测试集的一条案例，通过 KNN 检索策略检索出 K 个近邻案例。

Step4 从 K 个近邻中找到案例最多的类别作为新案例的建议类别。

Step5 根据式(3)执行新案例的保存：如果新案例的实际类别与建议类别不一致，或者实际类别与建议类别一致但是最大相似度小于阈值，则将正确的类别赋给新案例，为其添加初始遗忘值，并将新案例保存到案例库。

Step6 根据 2.2 节所述的遗忘触发策略，判断是否对 K 个案例进行遗忘。如果否，则转 Step7；如果是，则根据式(4)分别对 K 个案例进行遗忘值的更新，并据式(5)判断更新遗忘值后的近邻案例是否应该被删除。若满足删除条件，则删除该案例；否则保持该案例及其更新的遗忘值。

Step7 判断测试集案例是否已全部取完，若取完，则 CBR 分类结束；否则，转 Step3。

3 实验评估

为了验证记忆与遗忘策略对 CBR 分类器的改进效果，利用 UCI 数据集进行了分类对比实验，然后将改进的 CBR 系统应用于心血管病的辅助诊断。

3.1 UCI 分类实验

3.1.1 实验设置

为了评价提出的记忆和遗忘策略对 CBR 系统的影响，本文采用 UCI 资源库[23]中的 10 个数据集进行分类实验。实验数据集的具体情况如表 1 所列，其中集包含了类别和样本数量的不同情况。

表 1 数据集概况

数据集	缩写	样本	属性	类别
seeds	SE	210	7	3
Glass	GL	214	9	6
heart	HE	270	13	2
BUPA	BU	345	6	2
wdbc	WD	569	30	2
ILPD	IL	583	10	2
PIDD	PI	768	8	2
Steel plates faults	STE	1941	27	7
Cardiotocography	CA	2126	22	3
Image Segmentation	IMA	2310	18	7

本文提出 3 种遗忘触发策略，根据遗忘触发策略的不同分 3 种分类器，分别是采用成功触发遗忘策略的 CBR 分类器（简记为 SD）、失败触发策略的 CBR 分类器（简记为 FD）、混合触发策略的 CBR 分类器（简记为 HD），这 3 种分类器的保存策略都采用式(3)所示的记忆策略，遗忘策略中的遗忘值更新子策略和案例删除子策略都按式(4)和式(5)进行。另外，为了比较本文的选择性保存策略对 CBR 性能的影响，还增设了一种不加遗忘策略的 CBR 分类器（记为 ND），该分类器的其他环节都与上述 3 个分类器相同。为了验证选择性记忆和回顾环节在 CBR 循环中的作用，将上述 4 种 CBR 分类器与 2 种采用传统 CBR 循环的分类器进行比较，分别采用全部不保存（NR）和全部保存（AR）的策略。以上 6 种分类器均采用 KNN 检索的检索策略（为了消除不同 K 值对分类效果的影响，统一取最近邻个数 $K=3$），重用策略为重用近邻案例中案例个数最多的类别。SD，FD，HD，ND 4 种分类器的参数设置如下：相似度阈值 $\delta=0.7$，遗忘度阈值 $\xi=0.6$，遗忘增强因子 $\alpha=0.2$。

每个数据集都采用 5 折交叉验证。分类器的性能评价指标包括准确率、分类时间（包含检索、重用、保存、回顾在内的整个 CBR 循环所需的时间）和案例库变化率（分类后与分类前的案例库样本数量之比，即（训练集样本数＋记忆的样本数－遗忘的样本数）/训练集样本数））。另外，采用 T 检验来验证这些方法是否有显著差异。

3.1.2　各分类器分类性能的比较

为了对比本文提出的记忆和遗忘策略对 CBR 性能的影响，分别使用 AR，NR，HD，SD，FD，ND 分类器对表 1 所列的 10 个数据集进行分类，记录准确率、分类时间和案例库变化率的平均值，并将 NR，HD，SD，FD，ND 的结果与 AR 进行 T 检验（显著性水平为 0.05），结果分别如表 2－表 4 所列（加粗表示性能最优，· 表示性能显著改善，* 表示性能显著降低）。

表 2　平均准确率及显著性分析（$K=3$）

数据集	AR	NR	HD	SD	FD	ND
SE	**93.3333**	92.3810	92.3810	92.8571	92.8571	92.8571
GL	71.6279	71.6279	70.2326	**73.0233**	69.3023	71.6279
HE	78.1481	75.9259	79.2593	78.5185	**79.6296**	77.7778
BU	**62.8986**	62.3188	59.4203	61.4493	61.1594	62.6087
WD	**97.0175**	96.8421	96.8421	96.8421	96.6667	96.8421
IL	48.3761	48.3761	**69.0598**	48.3761	66.1538	48.3761
PI	72.7273	72.8571	**74.5455**	74.0260	72.8571	72.7273
STE	71.4139	**71.8252**	71.0540	71.4139	70.8997	71.3111
CA	**96.7136**	96.5258	**96.7136**	96.6197	**96.7136**	96.6197
IMA	**96.6667**	96.2338	96.3636	96.3636	96.3636	96.3636
平均值	78.8923	78.4914	**80.5872**	78.9490	80.2603	78.7111

由表 2 可知，就所有数据集的平均准确率而言，NR 最低，HD 最高，其他含保存环节的分类器的准确率都高于 NR 的准确率，说明新案例的保存可以提高分类准确率。HD，SD，FD 的平均准确率高于 AR 和 ND，说明旧案例的遗忘对提高分类准确率起到积极作用。ND 的平均准确率低于 AR，但是对这两分类器在相同数据集上的 T 检验表明，ND 和 AR 的分类准确率没有显著变化（显著性水平为 0.05），即相对全部记忆的保存策略，选择性记忆不会显著降低 CBR 的分类准确率。

表 3　平均分类时间及显著性分析（$K=3$）

数据集	AR	NR	HD	SD	FD	ND
SE	0.0137	0.0121	0.0134	0.0156*	0.0122	**0.0120**
GL	0.0143	0.1731	0.0149	0.0152	**0.0139**	0.0144
HE	0.0224	0.0224	0.0237	0.0221	**0.0209**	0.0214
BU	0.0286	0.0263·	0.0284	0.0296	**0.0262**	0.0265
WD	0.0828	**0.0722**·	0.0828	0.0841	0.0763·	0.0767·
IL	0.0646	0.0597	0.0577·	0.0615	**0.0568**·	0.0610
PI	0.0917	**0.0814**·	0.0852·	0.0876·	0.0823·	0.0833·
STE	0.7418	0.6299·	0.6172·	0.6434·	**0.5913**·	0.6186·
CA	0.7498	**0.6398**·	0.6946·	0.6991·	0.6418·	0.6451·
IMA	0.7844	0.7653	0.7010·	0.7035·	**0.6483**·	0.6517·
平均值	0.2594	0.2482	0.2319	0.2362	**0.2170**	0.2211

由表 3 可知，就所有数据集的平均分类时间而言，FD 最优，HD 和 SD 也都优于 AR 和 NR，说明案例遗忘带来的案例库缩减可以减少分类时间。将 HD，SD，FD，ND 的分类时间与 AR 的分类时间进行 T 检验，可见当样本数量较多时（如后 4 个数据集），具有遗忘或选择性记忆功能的分类器的时间性能都显著提高；而当样本数量较少时（如前 3 个数据集），时间优势不明显，这是因为当样本数量较少时，本文提出的维护方法所针对的案例库覆没问题不明显。

表 4　平均案例库大小变化率及显著性分析（$K=3$）

数据集	AR	NR	HD	SD	FD	ND
SE	125	100·	**96.07**·	99.64·	97.62·	101.79·
GL	125	100·	**86.40**·	99.19·	89.42·	107.21·
HE	125	100·	**91.11**·	101.20·	92.96·	106.02·
BU	125	100·	**85.29**·	101.74·	89.13·	109.35·
WD	125	100·	**97.85**·	99.34·	98.86·	100.79·
IL	125	100·	92.18·	107.95·	**90.85**·	112.91·
PI	125	100·	**88.44**·	99.77·	91.53·	106.82·
STE	125	100·	**87.44**·	101.40·	90.01·	107.17·
CA	125	100·	**98.56**·	99.89·	98.97·	100.85·
IMA	125	100·	**98.04**·	99.61·	98.76·	100.91·
平均值	125	100·	**92.14**	100.97	93.81	105.38

由表 4 可知，相对于 AR，其他分类器的案例库空间都显著减小，说明遗忘策略和选择性记忆可以减小案例的储存空间。其中，采用 HD 的空间减小得最多，这是因为 HD 在对每一条新案例进行分类后都会触发案例的遗忘策略；而 SD 和 FD 则分别在新案例分类正确或错误时才触发遗忘策略；ND 则只含有选择性保存，不进行案例遗忘。

综上，本文提出的具有选择记忆和有意遗忘功能的 CBR 分类器可以在提高或不显著降低分类准确率的前提下减小案例库空间，缩短运行时间。这一实验结果可用前文介绍的有意遗忘理论解释，即在案例检索之前将某些无关或“有害”的案例删除，以降低或消除无关信息的干扰，提高案例检索的准确率，同时保证案例库的容量在合理的范围，降低 CBR 循环过程的复杂度。综合考虑准确率、时间和案例库空间因素，在这些改进的分类器中，失败驱动触发遗忘策略的 CBR 分类器在分类时间上的性能最优，混合驱动触发遗忘策略的 CBR 分类器在分类准确率和案例库空间缩减上的性能最优。

3.2　CBR 在心血管病分类中的应用

为了验证改进的 CBR 的应用效果，将其应用于心血管病的分类中。心血管病数据集[24]由北京工业大学生命科学与生物工程学院提供，该机构为了研究不同心血管状态下桡-指脉搏波关系以及脉搏波新特征参数的变化关系，先后在首都医科大学附属北京安贞医院和北京工业大学生命科学与生物工程学院进行了心血管血流动力学检测。该数据集共包括 930 条案例，每条案例包括 17 个描述属性（包括年龄、收缩

压、舒张压、心率、性别以及脉搏波的各种波形参数等)和 1 个类别属性(心血管状态正常组类别为 0(表示阴性),心血管状态异常组类别为 1(表示阳性))。心血管病数据集的详细信息如表 5 所列。

表 5 心血管病数据集概况

项目	数量
案例数	930
属性个数	17
类别	2
阳性数	465
阴性数	465

为了验证改进前后 CBR 分类器的性能,同时研究训练集和测试集的数量对分类器性能的影响,分别采用 AR(代表传统 CBR 分类器)和 HD(代表改进后的 CBR 分类器)分类器对心血管病数据集进行如下两组实验。

实验 1 大样本实验:训练集数量较多,测试集数量较少。测试在较少的测试数据下记忆与遗忘功能对 CBR 分类准确率、分类时间和空间变化率的影响。

实验 2 小样本实验:训练集数量较少,测试集数量较多。测试在较多的测试数据下记忆与遗忘功能对 CBR 分类准确率、分类时间和空间变化率的影响。

为了保证实验的有效性,采用五折交叉验证。数据集分成 5 份,大样本实验的每一折实验中,取数据集中的 4 份作为训练集,剩余的 1 份为测试集;小样本实验与之相反,每一折实验中,取数据集中的 1 份作为训练集,剩余的 4 份为测试集。KNN 检索时的最近邻个数 $K=1$,其他参数的设置和性能评价指标同 3.1 节。

改进前、后的 CBR 分类器(AR 和 HD)在两组实验中的结果如表 6 所列。表 7 列出了 HD 在这些评价指标上相对于 AR 的下降率。

表 6 平均准确率、平均运行时间和平均案例库大小变化率

实验类型	分类器	准确率/%	分类时间/s	空间变化率/%
大样本	AR	91.6129	0.1314	125.0000
	HD	91.1828	0.1167	100.2688
小样本	AR	91.4247	0.3502	500.0000
	HD	91.3978	0.1393	116.4516

表 7 HD 相对于 AR 在各评价指标上的下降率/%

实验类型	准确率	分类时间	空间变化率
大样本	−0.47	−11.16	−19.78
小样本	−0.03	−60.23	−76.71

由表 6 和表 7 可知,在各组实验中,HD 的分类准确率稍低于 AR,但是其分类时间和空间变化率都明显优于 AR。具体地,相对于 AR,HD 的分类准确率下降不超过 0.5%,而分类时间节省率超过 10%,空间下降率更为显著,超过 19%。小样本实验时,由于测试案例数量更多,HD 在分类时间和空间变化率上下降得更为显著。

上述结果表明,在心血管病数据集上,相对于传统的 CBR,记忆与遗忘策略改进的 CBR 可以在不显著降低准确率的前提下减小案例库的空间并节省时间。CBR 对心血管病数据的分类结果可以为医生的诊断提供辅助支持,避免诊断的盲目性。

结束语 为克服案例推理中的“覆没问题”,本文对传统 CBR 系统的“4R”循环进行扩展和修改,在保存环节后增加了回顾环节,使之成为具有选择性记忆和有意遗忘功能的 CBR 系统。

扩展的 CBR 系统通过对新案例的选择性保存和对历史库中旧案例的有意遗忘,合理维护案例库。选择性记忆可以在保证有效学习的同时避免案例库的盲目增长;有意遗忘方式可以动态控制案例库的大小,并去除案例库中的“无用案例”。

UCI 和心血管病数据集的实验表明,具有记忆和遗忘功能的 CBR 系统可以提高或不显著降低分类准确率,减少分类时间,动态控制案例库的增长,当样本数量较多时,这种效果更明显。因此选择记忆和有意遗忘有效地抑制覆没问题,对案例库维护具有积极作用。下一步将继续研究相似度阈值、遗忘度阈值以及 K 值对本文方法的影响。

参 考 文 献

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

[2] CHEN Y W, CHAI T Y. Preprocessing of operation data in heating furnace [J]. Control Theory & Applications, 2012, 29(1): 114-118. (in Chinese)
陈友文, 柴天佑. 加热炉生产数据预处理策略研究[J]. 控制理论与应用, 2012, 29(1): 114-118.

[3] HSU K H, CHIU C C, CHIU N H, et al. A case-based classifier for hypertension detection[J]. Knowledge-Based Systems, 2011, 24(1): 33-39.

[4] HUO D, LI W Z, ZUO Y Z, et al. Detection Analysis of Steel Structural Damage Based on CBR [J]. Journal of Beijing University of Technology, 2013, 39(4): 570-575. (in Chinese)
霍达, 李伟峥, 左勇志, 等. 基于改进案例推理的钢结构破损原因诊断分析[J]. 北京工业大学学报, 2013, 39(4): 570-575.

[5] FENG C, YANG N D, GUI W M, et al. Method for generating emergency alternative based on case-based reasoning[J]. Control and Decision, 2016, 31(8): 1526-1530. (in Chinese)
封超, 杨乃定, 桂维民, 等. 基于案例推理的突发事件应急方案生成方法[J]. 控制与决策, 2016, 31(8): 1526-1530.

[6] FRANCIS A G, RAM A. The utility problem in case-based reasoning [R]. USA: Georgia Institute of Technology, 1993.

[7] SCOTT P D, MARKOVITCH S. Knowledge considered harmful [C]// Proceedings of IEEE Colloquium on Knowledge Engineering. London, 1990: 1-5.

[8] MARKOVITCH S, SCOTT P D. The role of forgetting in learning [C]// Proceedings of the Fifth International Conference on Machine Learning. 1988: 459-465.

[9] WEN T Z, XU A Q, SUN W C. Fault diagnosis method based on extension case-based reasoning [J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(11): 2124-2130. (in Chinese)
文天柱, 许爱强, 孙伟超. 基于可拓案例推理的故障诊断方法[J]. 北京航空航天大学学报, 2015, 41(11): 2124-2130.

[10] SMYTH B, MCKENNA E. Competence models and the maintenance problems [J]. Computational Intelligence, 2001, 17(2): 235-249.

(下转第 289 页)

证明相似度系数为 S_1 的尺度上推算法的效果优于相似度系数为 S_2 的尺度上推算法。通过实验验证了 MSARSUA 算法的可行性和有效性。

结束语　本文针对多尺度数据挖掘研究尚未成熟的现状，对多尺度关联规则挖掘方法进行了分析和完善，提出了一种基于高斯金字塔法的多尺度关联规则尺度上推算法。本文基于包含度的相似度理论提出了频繁项集处理方法，以便于尺度上推算法的实现；详细介绍了高斯金字塔法的加权平均思想，在此基础上提出了多尺度关联规则尺度上推算法。最后，通过实验验证了 MSARSUA 算法的可行性和有效性。未来工作中，将多尺度思想和分类、聚类等数据挖掘方法相结合，进一步研究多尺度分类和聚类挖掘算法；探索更加优秀的相似度理论和计算权重的方法，将分布式并行的思想引入多尺度关联规则算法中，以进一步提高算法的执行效率。

参 考 文 献

[1] LI H J, LI H Y, LI A H. Analysis of multi-scale stability in community structure[J]. Chinese Journal of Computer, 2015, 38(2): 301-312. (in Chinese)
李慧嘉，李慧颖，李爱华. 多尺度的社团结构稳定性分析[J]. 计算机学报，2015，38(2)：301-312.

[2] CHEN J J, SU S B, XU H L. Decision tree optimization algorithm based on multiscale rough set model[J]. Journal of Computer Applications, 2011, 31(12): 3243-3246. (in Chinese)
陈家俊，苏守宝，徐华丽. 基于多尺度粗糙集模型的决策树优化算法[J]. 计算机应用，2011，31(12)：3243-3246.

[3] LIU M M, ZHAO S L, CHEN M, et al. Scaling-up mining algorithm of multi-scale association rules mining[J]. Application Research of Computers, 2015, 32(10): 2924-2929. (in Chinese)
柳萌萌，赵书良，陈敏，等. 多尺度关联规则挖掘的尺度上推算法[J]. 计算机应用研究，2015，32(10)：2924-2929.

[4] WANG Z, ZHANG L, FANG T. A Multiscale and Hierarchical Feature Extraction Method for Terrestrial Laser Scanning Point Cloud Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(5): 2409-2425.

[5] ALEX J, HENRI P, SYLVIE, et al. Interactive Multiscale Classification of High-Resolution Remote Sensing Images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2013, 6(4): 2020-2034.

[6] YING D, TAO G, LEI L. Self-Organizing Map Based Multi-scale Spectral Clustering for Image Segmentation[C]// 2012 International Conference on Computer Science and Electronics Engineering. 2012: 329-333.

[7] ZHANG W X, XU Z B, LIANG Y, et al. Inclusion degree theory [J]. Fuzzy Systems and Mathematics, 1996, 10(4): 1-9. (in Chines)
张文修，徐宗本，梁怡，等. 包含度理论[J]. 模糊系统与数学，1996，10(4)：1-9.

[8] YU Z M, GAO F. Laplacian pyramid and contrast pyramid based image fusion and their performance comparison[J]. Application Research of Computers, 2004, 21(10): 128-130. (in Chinese)
玉振明，高飞. 基于金字塔方法的图像融合原理及性能评价[J]. 计算机应用研究，2004，21(10)：128-130.

(上接第 284 页)

[11] SMITI A, ELOUEDI Z. WCOID-DG: an approach for case base maintenance based on weighting, clustering, outliers, internal detection and dbsan-gmeans [J]. Journal of Computer and System Sciences, 2014, 80: 27-38.

[12] YAN X, FU H, TU N W. Dynamic Prediction of Coal and Gas Outburst Based on Clustering and Case-Based Reasoning[J]. Journal of Transduction Technology, 2016, 29(4): 545-551. (in Chinese)
阎馨，付华，屠乃威. 基于聚类和案例推理的煤与瓦斯突出动态预测[J]. 传感技术学报，2016，29(4)：545-551.

[13] MÁNTARAS R L D, MCSHERRY D, BRIDGE D, et al. Retrieval, reuse, revision and retention in case-based reasoning [J]. The Knowledge Engineering Review, 2005, 20(3): 215-240.

[14] SCHOOLER L J, HERTWIG R. How forgetting aids heuristic inference [J]. Psychological Review, 2005, 112: 610-628.

[15] KENNEDY W G, TRAFTON J G. Long-term symbolic learning [J]. Cognitive Systems Research, 2007, 8, 237-247.

[16] DERBINSKY N, LAIRD J E. Effective and efficient forgetting of learned knowledge in Soar's working and procedural memories [J]. Cognitive Systems Research, 2013, 24: 104-113.

[17] SALAMÓ M, LÓPEZ-SÁNCHEZ M. Adaptive case-based reasoning using retention and forgetting strategies [J]. Knowledge-Based Systems, 2011, 24(2): 230-247.

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

[19] BAI X J, ZHANG L H, LI H X. The experimental study on the intentional forgetting between two paradigms [J]. Journal of Psychological Science, 2011, 34(1): 2-6. (in Chinese)
白学军，张丽华，李红霞. 两种范式下有意遗忘的实验研究[J]. 心理科学，2011，34(1)：2-6.

[20] COX M T, RAM A. Introspective multistrategy learning: on the construction of learning strategies [J]. Artificial Intelligence, 1999, 112: 1-55.

[21] ZHANG Z, YANG Q. Feature weight maintenance in case bases using introspective learning[J]. Journal of Intelligent Information Systems, 2001, 16: 95-116.

[22] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. Cambridge: The MIT Press, 1998.

[23] FRANK A, ASUNCION A. UCI machine learning repository. Irvine, CA: University of California[OL]. http://archive.ics.uci.edu/ml.

[24] YANG L. Discussions of evaluation methods of cardiovascular function by pulse wave[D]. Beijing: Beijing University of Technology, 2010. (in Chinese)
杨琳. 脉搏波评价心血管功能的方法探讨[D]. 北京：北京工业大学，2010.