用遗传模拟退火算法挖掘特征项权重的研究

齐 平,贾瑞玉,贾兆红,王会颖
(安徽大学 计算机学院,安徽 合肥 230039;
安徽大学 计算智能与信息处理教育部重点实验室,安徽 合肥 230039)

摘 要:能否在范例库中检索和选择出最为相似的范例决定了范例推理系统性能。文中介绍了遗传算法和模拟退火算法,比较了两种算法的特性,提出一种混合遗传模拟退火算法。该算法不但具有强的局部搜索能力,还缩短了搜索时间。将该算法用于发掘范例库上特征权重,理论分析和实验结果表明了这种混合遗传模拟退火算法优于普通的遗传算法。

关键词:遗传算法;模拟退火算法;权重;范例推理
中图分类号:TP18 文献标识码:A 文章编号:1673-629X(2007)02-0143-03

Using Genetic-Simulated Annealing Algorithm to Find Attribute Weighting

QI Ping,JIA Rui-yu,JIA Zhao-hong,WANG Hui-ying
(School of Computer Science,Anhui University,Hefei 230039,China;
Ministry of Education Key Lab. of Intelligent Computing and Signal Processing,Anhui Univ.,Hefei 230039,China)

Abstract:This article introduces two algorithms,genetic algorithm and simulated annealing algorithm,and puts forward one weighting method by using genetic-simulated annealing algorithm. This algorithm not only has the strong partial searching ability,moreover also reduces the searching time. The theoretical analysis and experimental results show that this method has better performance than other methods,by using this algorithm to find the characteristic weighting of case base.

Key words:genetic algorithm;genetic-simulated annealing algorithm;weighting;case-based reasoning

0 引 言

范例推理(Case-Based Reasoning,CBR)是由目标范例的提示而得到历史记忆中的源范例,并由源范例来指导目标范例求解的一种策略,它是一种重要的机器学习方法[1]。CBR 是区别于基于规则推理的一种推理和学习模式,它是指借用旧的事例或者经验来解决问题,评价解决方案,解释异常情况或理解新情况。

CBR 的显著优点有:信息的完全表达,增量式学习,形象思维的准确模拟,知识获取较为容易,求解效率高[2]。与传统专家系统相比,它的最大优点在于动态知识库,即通过增量学习而不断增加知识。然而,CBR 也存在一些问题,主要体现在 CBR 对噪声数据比较敏感和范例工程的自动化程度不够,同时,知识的获取也存在一定程度的瓶颈[1]。

范例库推理系统的学习和推理性能,是由它从范例库中检索和选择与给予的范例最相近范例的能力决定的。范例之间的相似性是检索的关键,其中范例库的特征项权重对检索的质量与速度都起到重要的作用。文中提出用遗传模拟退火算法来发现特征项权重的思想与算法。

1 遗传模拟退火算法

1.1 遗传算法

遗传算法(Genetic Algorithms)是 J. Holland 于 1975 年受生物进化论的启发而提出的,它是模拟生物在自然环境中的遗传和进化过程而形成的一种自适应全局化概率搜索法,遗传算法模拟自然进化中"优胜劣汰,适者生存"的原理进行自学习和寻优[3]。

遗传算法一般通过初始化、选择、交叉和变异四步来完成最优解的搜索过程。在初始化阶段,产生一个初始群体,每个个体代表了问题的一个可能解,它们随机地分布在问题的解空间。这个群体是搜索的起点,

收稿日期:2006-05-18
基金项目:安徽省教育厅科研项目(2005kj055;2005kj056)
作者简介:齐 平(1981-),男,安徽枞阳人,硕士研究生,研究方向为机器学习、知识发现;贾瑞玉,副教授,硕士生导师,研究方向为专家系统、神经网络、知识发现。

在初始化后,由用户所定义的适应度函数来评价每个染色体,以使每个染色体有一个初始的适应度,从而将每个个体的进化能力数字化。

复制规则是为了让具有高适应度的个体被保留,而且它们的重要属性可以一代一代地被复制下去,以使搜索朝着最优的方向进行。适应度值高的个体可以多次被选中并复制到下一代中,而适应度值低的可能一次也不被选中[4]。

交叉是指在随机选择的两个“父母”之间,通过交换双亲染色体的对应基因段,产生一个新的个体。交叉操作扩大了搜索空间,使算法能在更加广阔的空间里寻找新解。同时,它也用交叉概率来限制交叉发生的可能性。

虽然复制和交叉产生了许多新的串,但它们没有在位上引入任何新的信息,这一功能是由变异完成的[1]。变异的遗传原理是随机选择染色体群中的一个个体,随机选择该染色体上的一位并改变它,如果变异后新个体比原先的优良,它就取代原个体[3]。

遗传算法用简单的编码技术和繁殖机制来表现复杂的现象,不受搜索空间的限制性假设约束,而且它具有良好的全局搜索能力,利用它的内在并行性可以方便地进行分布式计算,方便求解。但是遗传算法的局部搜索能力比较差,容易“过早收敛”,陷入局部最优解[5]。

1.2 模拟退火算法

模拟退火算法(Simulated Annealing)又称为模拟退火法。模拟退火算法源于对固体退火过程的模拟,它能够以随机搜索从概率的意义上找出目标函数的全局最优点。模拟退火算法特别适合于解决大型组合优化问题,算法的核心在于模拟热力学中液体的冻结与结晶或金属溶液的冷却与退火过程[6,7]。

经典退火方式为:

$$T(t) = T_0 / \ln(1 + t)$$

该方案特点是温度下降很缓慢,因此算法的收敛速度也很慢。

快速退火方案为:

$$T(t) = T_0 / 1 + \alpha t$$

该方案在高温区的温度下降是比较快的,而在低温区的降温速率较小[8]。

模拟退火算法能够以随机搜索技术从概率意义上找出目标函数的全局最优解,但它也有很多不足:它对整个搜索空间的了解不多,不便于搜索过程进入最有希望的搜索区域,从而使模拟退火算法的效率不高[8]。

1.3 遗传模拟退火算法

分析遗传算法和模拟退火算法的基本原理,可以发现:遗传算法和模拟退火算法各自都有很多的优点,但也存在着诸多不足之处,它们两者之间有很强的互补性[9]。

遗传算法是模拟生物遗传和进化过程中选择、交叉、变异机理而形成的一种自适应全局优化概率搜索算法,但遗传算法最为严重的缺陷是过早收敛问题,在搜索的初期由于优良个体急剧增加使种群失去多样性,从而造成程序陷入局部,达不到全局最优解的现象。

模拟退火算法是基于金属退火的机理而建立起的一种全局最优化方法,虽然它能够以随机搜索技术从概率意义上找出目标函数的全局最优解,但由于计算速度和时间的限制,在优化效果和计算时间之间存在矛盾,因而难以保证计算结果为全局最优点,效果不理想[10]。

遗传算法和模拟退火算法有互补性,遗传算法把握总体的能力比较强,但局部搜索能力比较差[9];模拟退火算法具有比较强的局部搜索能力,但全局搜索能力不如遗传算法。因此将遗传算法与模拟退火算法结合起来,可以克服遗传算法和模拟退火算法各自的缺点,发挥它们的优势[11]。

2 用遗传模拟退火算法挖掘特征项权重

2.1 特征项赋权技术

给定目标范例,在范例库中检索和选择出最为相似的范例决定了范例推理系统性能。相似性是衡量对象之间相似度的标准,一般通过计算对象在特征空间中的距离获得。在 CBR 系统中,大多数的范例检索都使用近邻算法(K Nearest Neighbor)。可是,距离度量以及相似性函数对不相关的噪音特征很敏感,因此,如果有这些噪音特征,使用近邻算法就会有比较高的出错率。所以给数据集每个属性赋予一定的权重,权重的大小表示属性具有高低不同的相关性[2]。

给特征项赋予权重的算法有多种,如类映射技术、条件概率技术、神经网络方法、遗传算法等方法。文中采用结合遗传算法与模拟退火算法的遗传模拟退火算法给特征项赋予权重。

2.2 使用遗传模拟退火算法赋予特征项权重的算法

(1)用遗传算法进行全局搜索。

利用遗传算法的全局随即搜索能力,利用近邻法做为评估函数,做为适应度函数。在算法开始阶段将属性值的数据库数据分成两部分,产生参考范例集 REF 和测试范例集 TEST,使用 ref[i],test[j] 表示参考集和测试集中的第 i 和 j 个范例。

在实验中一个染色体代表一个权矢量,一个染色

体的每个基因表示的就是单个基因项的权重。对于每个权矢量,找个与每个测试范例距离最近的训练范例,利用所有测试范例与它们在参考集中最近邻的距离之和做为适应度函数,并以此来评估每个一权矢量的遗传能力[12]。

(2)退火过程。

用混合遗传模拟退火算法是独立地对选择交叉变异等遗传操作所产生的一组新个体进行模拟退火过程。对一组新个体随机选择各个个体中的基因做为扰动点,经过扰动的个体所得到的适应度增强则接受新个体,而新个体适应度减少,则以一定概率接受[12]。

(3)构造的遗传模拟退火算法流程。

①初始化 REF 参考范例集与 TEST 测试范例集。随即产生一个权矢量数组并进行评估。设置初试退火温度 t_0。

②计算当前权矢量的适应度值。

③选择两个个体,执行个体的交叉操作。

④以一定概率执行个体的变异操作。

⑤由模拟退火状态函数产生新个体。

⑥以一定概率接受新个体,执行个体模拟退火操作。

⑦判断模拟退火是否稳定,若不稳定则返回步骤⑤,否则执行退温操作。

⑧个体复制,判断条件是否可以终止,若可以则算法结束,否则返回步骤②。

3　实验结果

利用农业专家系统中'小麦苗情实例库'作为实验数据,来验证上述算法。

从库中选择了 4 个观测属性和 1 个决策属性。其中,'土壤肥力'分为高肥力、中肥力、低肥力;气象类型分为暖年、常年、冷年;灌溉水平分为 4 次以上、3－4 次、1－2 次;灌溉方式分为喷灌、渠灌。得到如下范例库信息:

土壤肥力	气象类型	灌溉水平	灌溉方式	产量
1	2	1	1	高
2	2	1	2	高
2	1	2	1	中
…	…	…	…	…

通过遗传算法来确定这 4 个观测属性的权值。其中每个测试属性的权值表示了该属性对于产量的决定能力,结果如下表所示:

土壤肥力	气象类型	灌溉水平	灌溉方式
0.1206	0.0977	0.1881	0.0365

通过遗传模拟退火算法确定 4 个观测属性的权值。结果如下:

土壤肥力	气象类型	灌溉水平	灌溉方式
0.1371	0.0936	0.1703	0.0271

从库中选取记录,代入进行推理,用分类精度作为标准,比较两种算法的优越性。结果如下:

方法	遗传算法	遗传模拟退火算法
分类精度	72.5%	79.3%

实验结果表明,遗传模拟退火算法在性能上要优于单纯的遗传算法。

4　结束语

理论上来讲,遗传算法和模拟退火算法两种算法都属于基于概率分布机制的优化算法。模拟退火算法的优化机制是通过赋予搜索过程最终趋于零的概率突变性来避免陷入局部极小而达到全局最优;遗传算法则通过概率意义下的'优胜劣汰'思想的群体遗传操作实现优化。

文中提出的遗传模拟退火算法的基本思想是将遗传算法与模拟退火算法相结合而构成的一种优化算法。与基本遗传算法的总体运行过程相类似,遗传模拟退火算法也是从一组随机产生的初始解开始全局最优解的搜索过程。它通过选择、交叉、变异等遗传操作来产生一组新的个体,然后再独立地对所产生出的个体进行模拟退火过程,以其结果作为下一代群体中的个体,反复进行这个过程,直到满足某个终止条件。

遗传模拟退火算法不但具有强的局部搜索能力,克服了早熟现象,而且还保留了优胜劣汰的替换策略,缩短了搜索时间。理论分析和实验结果表明混合遗传模拟退火算法在性能上要优于单纯的遗传算法。

参考文献:

[1] 杨善林,倪志伟.机器学习与智能决策支持系统[M].北京:科学出版社,2004.

[2] 贾兆红,倪志伟,赵　鹏.用遗传算法挖掘范例库中的特征权重的方法[J].计算机工程,2003,29(14):586－589.

[3] 刘怀亮,刘　淼.一种混合遗传模拟算法及其应用[J].广州大学学报:自然科学版,2005,4(2):79－82.

[4] 倪志伟,蔡庆生,贾瑞玉.范例库中特征项权重发现技术[J].厦门大学学报:自然科学版,2002,41(1):158－162.

[5] 周　明,孙树栋.遗传算法原理及应用[M].北京:国防工业出版社,1999.

[6] 刘志刚,王建华,耿英三,等.一种改进的遗传模拟退火算法及其应用[J].系统仿真学报,2004,3(5):125－129.

[7] 靳利霞,唐焕文,李　斌,等.一类连续函数模拟退火算法及其收敛性分析[J].计算数学,2005,2(1):22－28.

[8] 陈华根.模拟退火算法机理研究[J].同济大学学报:自然科学版,2004,32(6):347－352.

(下转第 148 页)

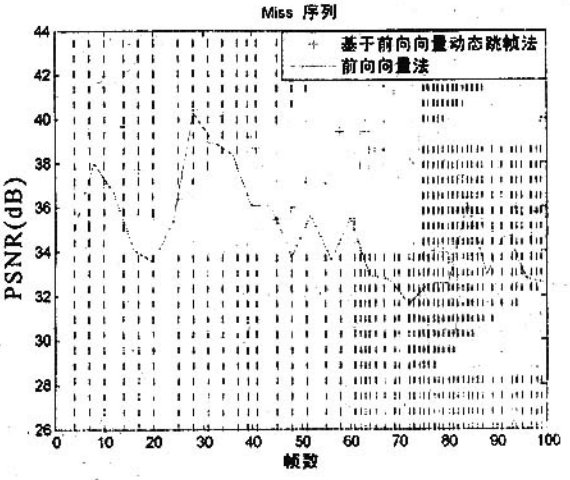
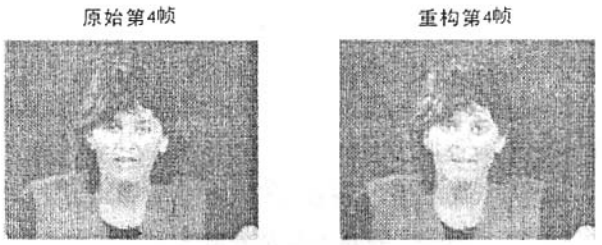


图 4 MISS 序列动态跳帧情况

图 5 和图 6 中,分别给出了 Miss 序列第 0 帧到第 4 帧和第 31 帧到第 34 帧的跳帧情况,可以看出,重构的第 4 帧和原始第 4 帧图像有些差异,但没有带来明显的失真,而重构的第 34 帧和原始的第 34 帧图像基本一样,得到较好图像的视觉效果。这说明在进行动态跳帧的同时也要兼顾视频质量,这就需要选择合适的阈值。对同一序列,丢失的帧数越多,重建图像的平均信噪比越小。

(a) 原始的第 0 帧到第 4 帧序列图像

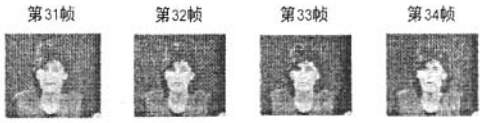


(b) 左图为原始第 4 帧,右图为重建第 4 帧

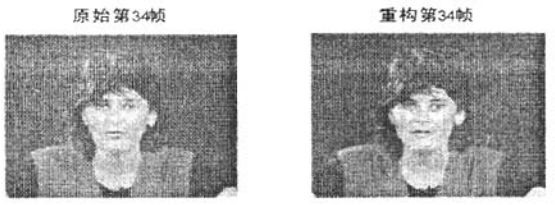
图 5 第 0 帧到第 4 帧动态跳帧与运动矢量的动态修正

4 结束语

在转换编码中,根据帧与帧之间累计的运动矢量大小,文中提出了一种动态跳帧机制。在动态跳帧转换编码的基础上进一步提出了动态运动矢量修正问题,用来调整修正运动矢量的范围。仿真结果表示,基于前向向量法,文中仿真了动态跳帧和运动矢量的修正。结果显示比双线性内插法、FDVS 和前向向量法能够提高图像的信噪比。



(a) 原始的第 31 帧到第 34 帧序列图像



(b) 左图为原始第 34 帧,右图为重建第 34 帧

图 6 第 31 帧到第 34 帧动态跳帧与运动矢量的动态修正

参考文献:

[1] Vetro A, Christopoulos C, Sun huifang. Video Transcoding Architectures and Techniques: An Overview[J]. IEEE SIGNAL PROCESSING MAGAZINE, 2003, 20(2): 18 - 29.

[2] Hwang J N, Wu T D, and Lin C W. Dynamic frame - skipping in video transcoding[C]// In Proceedings of the IEEE Second Workshop on Multimedia Signal Processing. Redondo Beach, CA: [s.n.], 1998: 616 - 621.

[3] Shen Bo, Sethi I K, Bhaskaran V. Adaptive Motion Vector Resampling for Compressed Video Down - Scaling[C]// Proc. IEEE Conference on Image Processing (ICIP). Santa Barbara, CA: [s.n.], 1997: 771 - 775.

[4] 金庆学,贾明华,李晓辉. 基于降低时间分辨率转换编码模型的运动估值方法[J]. 安徽大学学报:自然科学版, 2004(6): 48 - 52.

[5] Fung K - T, Chan Y - L, Siu W - C. Dynamic frame skipping for high - performance transcoding[C]// in IEEE Proc. On ICIP. [s.l.]: [s.n.], 2001: 425 - 428.

(上接第 145 页)

[9] 倪志伟,蔡庆生,贾瑞玉. 用神经网络来实现基于范例的推理系统[J]. 厦门大学学报:自然科学版, 2002, 7(1): 32 - 35.

[10] 刘素华,侯惠芳,李小霞. 基于遗传算法和模拟退火算法的特征选择方法[J]. 计算机工程, 2005, 31(16): 458 - 463.

[11] Fogel D B. System identification through simulated evolution: a machine learning approach to modeling[M]. America: Gnn Press, 1994.

[12] 倪志伟,李龙澍,贾瑞玉. Data mining and neural network techniques in case based system[J]. 武汉大学学报, 2001, 6(1): 24 - 30.