

基于禁忌遗传算法的权重发现技术

贾兆红^{1,2}, 唐俊¹, 卢冰原²

(1. 安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039;

2. 中国科学技术大学 信息管理与决策科学系, 安徽 合肥 230026)

摘要:相似范例的检索是范例推理的关键步骤之一,为了准确表达范例之间的本质特征,采用带有权重的特征项方法对范例检索起到重要的作用。在讨论了带有权重的最近邻算法的基础上,提出了一种禁忌遗传算法来获取范例库上的特征项权重,通过利用禁忌算法的自适应性和具有记忆功能的优点来改善遗传算法的全局搜索能力和提高其收敛速度。实验结果表明将这种方法应用于范例推理的案例检索过程中具有可行性,并且可以得到较高的分类精度和搜索效率。

关键词:基于范例的推理;权重;禁忌搜索;遗传算法

中图分类号:TP18

文献标识码:A

文章编号:1673-629X(2006)11-0026-02

Weights Finding Based on GA/TS Combined Algorithms

JIA Zhao-hong^{1,2}, TANG Jun¹, LU Bing-yuan²

(1. Ministry of Education Key Lab. of Intelligent Computing and Signal Processing, Anhui Univ., Hefei 230039, China;

2. Department of Information Management and Decision Science, University of Science and Technology of China, Hefei 230026, China)

Abstract: The attribute weighting has an important effect on the retrieval of similar cases, which is the key step of case-based reasoning, because it can exactly demonstrate the essential characters of cases. After discussing the improved k -Nearest Neighbor method, introduces a new algorithm, which combines the advantages of GA and TS, to find the attribute weights. The experimental result shows its good accuracy and feasibility.

Key words: case-based reasoning; weight; tabu search; genetic algorithms

0 概述

基于范例的推理(Case-Based Reasoning, CBR)是人工智能领域的研究热点之一^[1]。CBR的一般过程包括问题特征的提取、案例检索、案例修改/修正、新案例的存储等步骤。而范例特征项的权重对案例检索的结果起到重要作用^[2]。

遗传算法(Genetic Algorithm, GA)是CBR系统中一种非常有效的随机搜索方法,具有运算简单、并行搜索、鲁棒性强等特点。但是早熟问题和寻优时间较长是GA的缺点。禁忌搜索(Tabu Search, TS)是一种全局性邻域搜索算法^[3]。它通过局部邻域搜索机制和相应的禁忌准则来避免迂回搜索,并通过破禁水平来释放一些被禁忌的优良状态,进而保证多样化的有效搜索,以实现最终全局优化,具有多样化和自适应性的特点^[4]。但是TS的每次迭代

只能产生一个当前最优解,这就造成最终得到的解对初始解的依赖性较强。

鉴于两种方法自身的特点,文中将GA和TS两种方法结合起来使用,着重研究基于禁忌和遗传算法的范例特征项权重获取算法,并通过实例验证本算法的有效性。

1 特征项权重的发现问题

相似性是衡量对象之间相似度的指标,一般通过计算对象在特征空间中的距离获得。在CBR系统中,为避免噪音,可以使用改进的最近邻算法(k -Nearest Neighbor, k -NN)来进行范例检索,即使用带权重的距离度量算法。权重是根据数据集中各属性的相关性高低来给属性赋以一个值^[5]。

假设范例 $X = \{X_1, X_2, \dots, X_n\}$, $X_i (1 \leq i \leq n)$ 是它的特征值, W_i 是其权重。 X 是 n 维特征空间 $D = (D_1 * \dots * D_n)$ 上的一点, $X_i \in D_i$ 。对 $\forall X, Y \in D$, 则 X, Y 对于特征 F 的距离(或者说 X, Y 在特征空间 D 上的距离)为:

$$D_F(X, Y) = \left(\sum_i W_i * d(X_i, Y_i)^r \right)^{1/r} \quad (1)$$

其中:

收稿日期:2006-01-16

基金项目:安徽省教育厅自然科学基金资助项目(2005kj055);安徽省高校青年教师科研基金项目(2005jq1034)

作者简介:贾兆红(1976-),女,安徽巢湖人,讲师,博士研究生,研究方向为商务智能;导师:陈华平,教授,博导,研究方向为智能信息与决策支持系统。

$$d(X_i, Y_i) = \begin{cases} |X_i - Y_i| & \text{如果 } D_i \text{ 是连续的} \\ 0 & \text{如果 } D_i \text{ 是离散的, 且 } X_i = Y_i \\ 1 & \text{如果 } D_i \text{ 是离散的, 且 } X_i \neq Y_i \end{cases}$$

公式(1)中 r 的取值由用户定义, 如 $r = 2$ 时, 则 $D_F(X, Y)$ 为欧拉距离。

2 基于 GA 和 TS 的权重挖掘方法

将 TS 独有的记忆功能引入 GA 的搜索过程, 既可以利用 TS 强大的爬山能力来有效地克服 GA 的早熟缺点, 提高搜索速度, 又因 GA 具有并行搜索的优点, 可以通过 GA 来得到初始解, 以提高解的质量^[3,6]。基于如上讨论, 将 TS 引入 GA, 构造新的交叉算子 TSC 和变异算子 TSM 如下:

在 TSC 中将父代群体适应值的平均值作为破禁水平, 使用一个长度为 L 的禁忌表, 用来记录染色体的适应值。进行 TSR 时, 将经过交叉后的子代的适应值与破禁水平比较, 对优于破禁水平的子代染色体进行破禁, 即存入下一代中; 劣于破禁水平且属于禁忌的子代, 选择最好的那个进入下一代; 若不属于禁忌, 则接受该子代并存入下一代中。

TSM 的禁忌表存放的是最近变异的染色体上的位序号。每次发生变异时, 查询禁忌表, 若禁忌表存在当前变异位, 则禁忌; 但若变异后的新个体适应值优于破禁水平, 则将该新个体存入下一代。

利用改进后的遗传算法来完成特征项空间的直接搜索。算法的关键是 k -NN 的协作, 它被用来作为评估函数, 用在遗传中每个染色体的适应度函数上。算法在开始阶段, 将需要确定属性权值的数据库的数据分成两部分, 即产生参考案例集 REF 和测试案例集 TEST, $\text{ref}[i] \in \text{REF}$, $\text{test}[j] \in \text{TEST}$, $i = 0, \dots, m$; $j = 0, \dots, n$, 其中 $\text{ref}[j]$, $\text{test}[j]$ 分别表示 REF, TEST 中的第 i, j 个案例, m, n 分别为 REF, TEST 中的案例数。

在实验中, 一个染色体就是一个权矢量, 由多个基因组成, 每个基因表示的就是单个基因项的权重, 而基因的个数等于在案例集中独立特征项的个数。对于每个权矢量 $\text{weight}[i]$, 找每个测试案例 $\text{test}[j]$ 距离最近的训练案例, 利用所有的测试案例与它们在参考集中最近邻的距离 $\text{dist}(j, k)$ 之和作为适应度函数 $\text{fitter}[i]$, 其定义如下:

$$\text{fitter}[i] = \sum_{j=0}^n \sum_{k=0}^m d(\text{test}[j], \text{ref}[k]) \quad (2)$$

$$D(\text{test}[j], \text{ref}[k]) = \left(\sum_f W[i]_f * D(\text{test}[j]_f, \text{ref}[k]_f)^2 \right)^{1/2}$$

其中, $D(\text{test}[j], \text{ref}[k])$ 表示第 j 个测试案例与其最近的参考案例的距离; f 表示第 f 个相关特征, $W[i]_f$ 表示第 i 个权矢量第 f 分量的值, 即第 f 特征项的权值。

混合了 GA 和 TS 的权重发现算法 TSGA 的过程描述如下:

输入: n, m , 权矢量数组大小 N , 遗传代数 G , 禁忌表

长度 L

输出: 每代最优的适应度值, 最后得到的最优权矢量过程:

① 读取数据, 初始化 REF 和 TEST, 即产生 $\text{ref}[m]$ 和 $\text{test}[n]$;

② 生成初始权矢量数组 $\text{weight}[N]$;

③ 根据公式(2) 计算 weight 数组中每个对象的适应值 $\text{fitter}[i]$, 并依据适应值对 weight 数组排序;

④ 采用赌轮选择法及改进的 TSC, TSM 算子, 对 weight 数组优化 G 代;

⑤ 算法终止, 最终输出的第 G 代权矢量即为所求。

3 实验与讨论

农业气象数据库中相关的气象要素可以用来对气象灾害进行预测。经过预处理, 从库中选取 5 个观测属性(地区、季节、平均气温、平均降雨量和平均日照量)和 1 个决策属性(气象灾害), 这些属性分别命名为: Field, Fall, Tem, Rain, Sun 和 Disaster。主要工作就是通过新算法来确定这 5 个观测属性的权值, 即它们对于决策属性的影响程度。

首先对属性进行编码, 并将库中的数据分为参考集和测试集两部分。库中的案例采用整数编码, 编码的每个分量表示该案例的相应特征项的值。例如: 案例(25, 1, 5, 53, 26)表示 25 号地区在 1 季度的 $\text{Tem} = 5$, $\text{Rain} = 53$, $\text{Sun} = 26$ 。对权矢量采用浮点数编码, 矢量的每个分量都用一个浮点数来表示, 分别表示对应特征项的权值。

用提出的算法 TSGA 进行实验, 得到实验结果如表 1 所示。

表 1 发现的特征项权重

	Field	Fall	Tem	Rain	Sun
GA	0.003906	0.000977	0.000488	0.005981	0.001465
TS	0.0003976	0.0000984	0.000459	0.006053	0.001418
TSGA	0.003989	0.001004	0.000443	0.006145	0.001392

为了测试 TSGA 的优越性, 以分类精度作为标准, 将 TSGA 的特征项权重获取过程与 GA, TS 的作比较, 结果如表 2 所示。Accuracy 是通过最近邻算法, 对实验中的测试集案例计算所得出的分类精度。

表 2 实验结果比较

方 法	Accuracy(%)	平均运行时间(s)
GA	77.3	952.3
TS	85.1	516.8
TSGA	87.6	610.4

由表 2 可以看出, 在这 3 种赋权方法中, TSGA 的分类精度最高, TS 次之, GA 的最低; TS 的平均运行时间最短, GA 的平均运行时间最长。实验结果表明, TSGA, TS 稳定性和精度都比 GA 好, 而且在搜索速度上也优于 GA。由此可见, 在 GA 中引入 TS 不仅可以有效地避免早熟现

(下转第 31 页)

```

<s:homeothermic>yes</s:homeothermic>
<s:has_habitat>water</s:has_habitat>
<s:has_eggs>0</s:has_eggs>
<s:has_gills>0</s:has_gills>
</rdf:Description>
</rdf:RDF>

```

对 RDF 解析后可以得到以下的三元组集合和类别、类型信息:

```

has_covering(bat, hair) hair(covering) bat(animal)
has_legs(bat, 2) 2(rdfs:literal)
has_milk(bat, yes) yes(rdfs:literal)
homeothermic(bat, yes)
has_habitat(bat, air) air(habitat)
has_habitat(bat, caves) caves(habitat)
has_covering(shark, none) none(covering) shark(animal)
has_legs(shark, 0) 0(rdfs:literal)
has_habitat(shark, water) water(habitat)
has_eggs(shark, yes)
has_gills(shark, yes)

```

Mode declarations(模态声明)

```

:- modeb(1, class(+ animal, # class))?
:- modeb(1, has_gills(+ animal))?
:- modeb(1, has_covering(+ animal, # covering))?
:- modeb(1, has_legs(+ animal, # nat))?
:- modeb(1, homeothermic(+ animal))?
:- modeb(1, has_eggs(+ animal))?
:- modeb(*, habitat(+ animal, # habitat))?
:- modeb(1, has_milk(+ animal))?

```

types(类型)定义:

```

animal(bat), animal(shark), class(mammal), class(fish), covering
(hair), covering(none), habitat(air), habitat(caves), habitat(wa-
ter)

```

Backgroud knowledge(背景知识)即为以上所解析得

(上接第 27 页)

象而且可以大大提高收敛的速度。

4 结束语

遗传算法和禁忌搜索各具特点,文中将两种方法结合在一起进行范例库上特征项权重的发现,给出了一个混合算法。合理地将 GA 和 TS 组合在一起,不仅可以提高寻优速度并且可以提高整体的全局寻优能力。这种新算法不仅适用于 CBR 和最近邻算法,而且在模式识别等其他领域都有一定的应用价值。

参考文献:

- [1] Ashley K D, Bridge D G. Case - based reasoning research and development[C]//Proceedings of the 5th International Conference on Case - based Reasoning. Trondheim, Norway: [s.

到的三元组。

最后可以通过知识学习^[5]得到以下的分类规则:

```

class(A, mammal): - has_covering(A, hair)
class(A, fish): - has_gills(A)

```

4 结束语

语义 Web 采用 RDFMS 作为数据描述的方式,并且在本体层的支持下使数据包含了清楚的语义信息,于是对于这些数据进行挖掘将能够更好地发现所需要的知识。文中在语义 Web、数据挖掘、ILP 的理论基础上,着重进行了归纳逻辑程序设计 ILP 应用于语义 Web 数据挖掘的研究。由于现在的语义 Web 本身还处在一个初级发展阶段,基于它的数据挖掘方面的研究也仅仅是一个开始,要想充分地结合应用 ILP 和语义 Web 数据挖掘还需要更深入的探讨。

参考文献:

- [1] Berners - Lee T. Semantic Web Road map[DB/OL]. 1998 - 09. www. w3. org/DesignIssues/Semantic. html.
- [2] Berners - Lee T. Semantic Web Architecture[DB/OL]. 2000. www. w3. org/2000/Talks/1206 - xml2k - tbl/slide10 - 0. html.
- [3] Michie D, Spiegelhalter D J, Taylor C C. Machine Learning, Neural and Statistical Classification[M]. New York: Ellis Horwood, 1994.
- [4] 郑磊,贾东,刘椿年. 归纳逻辑程序设计综述[J]. 计算机工程与应用, 2003, 39(17): 46 - 49.
- [5] Edwards P. An Empirical Investigation of Learning from the Semantic Web[C]//First International Semantic Web Conference. Sardinia, Italy: [s. n.], 2002.
- [6] Wrobel S. Inductive Logic Programming for Knowledge Discovery in Database[C]//In: Dzeroski S, Lavrac N. Relational Data Mining. [s. l.]: Springer - Verlag, 2001: 74 - 101.

n.], 2003: 157 - 163.

- [2] Vollrath I. Handling vague and qualitative criteria in Case - based Reasoning Applications[C]//Proceedings of the 5th European Workshop on Advances in Case - based Reasoning. [s. l.]: Springer, 2000: 309 - 321.
- [3] 张颖,刘艳秋. 软计算方法[M]. 北京: 科学出版社, 2002.
- [4] Glover F, Laguna M. Tabu Search[M]. Boston: Kluwer Academic Publishers, 1997.
- [5] 贾兆红,倪志伟,赵鹏. 用遗传算法来挖掘范例库中的特征项权重[J]. 计算机工程, 2003, 29(14): 71 - 73.
- [6] Youssef H, Sait S M, Adiche H. Evolutionary algorithms, simulated annealing and tabu search: a comparative study[J]. Engineering Applications of Artificial Intelligence, 2001(14): 167 - 181.