基于禁忌搜索与遗传算法的案例检索技术

贾兆红[1,2]，贾瑞玉[1]，倪志伟[3]，唐 俊[1]
（1.安徽大学，安徽 合肥 230039；
2.中国科学技术大学，安徽 合肥 230026；
3.合肥工业大学，安徽 合肥 230009）

摘 要：案例的检索和提取是案例推理系统的一个关键步骤，案例检索结果的优劣直接影响到案例重用、修改以及整个系统的性能。遗传算法是一种基于进化思想的全局优化方法，但是存在搜索速度慢以及早熟收敛等问题；禁忌搜索是一种局部优化技术，具有搜索速度快等优点。文中将禁忌算法和遗传算法结合在一起提出了一种新的聚类方法，并将该聚类方法引入大型案例推理系统的案例检索过程中。实验结果表明使用这种方法能够达到较理想的搜索效果。

关键词：基于案例的推理；案例检索；禁忌搜索；遗传算法；聚类
中图分类号：TP18 **文献标识码**：A **文章编号**：1673－629X(2007)04－0147－03

Case Retrieval by Algorithm Based on Tabu Search and Genetic Algorithms

JIA Zhao-hong[1,2], JIA Rui-yu[1], NI Zhi-wei[3], TANG Jun[1]
(1. Anhui Univ., Hefei 230039, China;
2. Univ. of Sci. and Techn. of China, Hefei 230026, China;
3. Hefei Univ. of Techn., Hefei 230009, China)

Abstract: Case retrieval and selection is the key step of a case－based reasoning system. The results of case retrieval directly affect case reuse and revisal, even the performance of the system. Based on evolution strategy, genetic algorithm is a global search method, but it has the shortcomings, such as slow speed of convergence, premature convergence. Tabu search is a technique of local optimization with high search speed. Combining the advantages of GA and TS, a new approach for clustering is proposed and introduced in the process of case retrieval of a big case－based reasoning system. The experimental result shows that very nice effects are obtained with this new method.

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

0 引 言

基于案例的推理(Case－based Reasoning, CBR)是通过复用以前的实例来进行问题求解和学习的方法，而案例检索是 CBR 的关键步骤之一，并直接影响 CBR 系统的性能[1]。案例的检索是对一个待求解的新案例，利用案例库索引机制，根据相似性度量方法，从案例库中找出一组与新案例匹配较好的旧案例，并从中选择出一个最佳的案例。在实际应用中，案例库随着新案例的添加会不断增大，这样就加大了检索的负担，导致系统效率降低。为此，可以考虑在检索前将一个大的案例库划分成多个小的案例库，以减小搜索空间，达到提高检索效率的目的[2]。

遗传算法(Genetic Algorithms, GA)是一种非常有效的随机搜索方法，具有运算简单、鲁棒性强等特点。最近已有不少学者将 GA 成功地应用于聚类问题的研究中，实践证明 GA 具有解决此类问题的良好性能[3]。GA 虽然具有较强的全局搜索能力，但是局部搜索能力较弱，且早熟问题和收敛速度慢仍然是 GA 的缺点，因此在求解问题时，可以将一些局部搜索性能较好的算法与 GA 相结合以提高 GA 的搜索性能。禁忌搜索(Tabu search, TS)是一种搜索速度快、局部搜索能力强、不容易陷入局部极值的全局性邻域搜索算法[4]，具有多样化和自适应性的特点，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，以最终实现全局优化[5]。但是 TS 的初始解只能有一个，且算法的性能对初始解的依赖较强。因而文中将 GA 和 TS 结合起来，优势互补，提出一种

收稿日期：2006－06－22
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(2005jq1034，2006jq1034)
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、专家系统。

基于 GA 和 TS 的混合算法来进行聚类学习，并将此方法应用于大型案例库的案例检索过程，通过实验结果验证了算法的有效性。

1 基于禁忌遗传的聚类算法

聚类分析是将样本空间中的点集按照它们之间的“相似度”划分成若干个不同的类，使得属于同一类的个体之间的相似性尽量大，而属于不同类的个体间的相似性尽量小。聚类的方法很多，c－均值是其中常用的一种[6]。该方法首先选定某种距离度量作为模式间相似性度量，然后确定某个评价聚类划分结果质量的准则函数，在给出初始聚类中心点后，用迭代法找出使准则函数取极值的最好聚类划分结果。该方法收敛速度较快，但是容易陷入局部最优解，且对初始解的依赖性较强[6]。

GA 是一种全局搜索方法，具有并行搜索能力，适合求解大规模的全局优化问题，但其收敛速度较慢，且局部搜索能力较差，当种群中个体的相似性较大时，遗传操作很难引入新的基因，从而容易导致算法收敛于一个局部极值，出现早熟问题。TS 是一种全局逐步寻优的“爬山”算法，它通过引入一个灵活的存储结构和相应的禁忌准则来避免迂回搜索，并通过破禁准则来释放一些被禁忌的优良状态，具有较强的“爬山”能力，使搜索能够跳出局部极值，搜索到解空间的其他区域，以保证搜索到全局最优解。但是 TS 对初始解的依赖性较强，不同的初始解会导致差异较大的搜索结果。将 TS 独有的记忆功能引入 GA 的搜索过程，一方面可以利用 TS 强大的爬山能力来提高 GA 的搜索速度和避免其早熟的缺点，另一方面也能利用 GA 给 TS 提供一些较好的初始解，从而提高 TS 的局部搜索性能。因而，将 TS 引入 GA 的遗传操作中，构造新的交叉算子 TSC 和变异算子 TSM。两个算子分别使用一个长度为 L 的禁忌表来记录解的禁忌条件：TSC 的禁忌表 TL1 记录当前种群中 L 个最好染色体的适应度值；TSM 的禁忌表 TL2 存放最近变异的 L 个染色体上的位序号。将父代群体的平均适应度值作为破禁水平，即：$\mathrm{Asp}(s)=\sum_{g=1}^{p} f_{(s-1)g}/p$

设第 $s-1$ 代第 g、h 个染色体交叉后产生新个体 g'、h'，对每个新个体进行如下 TSC 操作：

If $f_{sg'} > \mathrm{Asp}(s-1)$ then accept g'

Else

  If $f_{sg'} \notin$ TL1 then accept g'

  Else choose the better of $\{g,h\}$ to generation s;

Update TL1;

对 $s-1$ 代的染色体 g 上第 r 位变异后产生的新个体 g'' 进行如下 TSM 操作：

If $f_{sg''} > \mathrm{Asp}(s-1)$ then accept g''

Else

  if $r \notin$ TL2 then accept g''

  Else forbid this mutation operation;

Update TL2;

聚类学习的对象是案例，其解是各个聚类的中心。在用 GA 求解这类问题时，可以用染色体上的一个基因表示一个聚类中心，c 个聚类中心就表示为一个染色体，一个染色体就代表了一个聚类解。

编码对于算法的性能影响很大，常用的编码方式有二进制编码和非二进制编码，一般来说，二进制编码比非二进制编码的搜索能力强，此外二进制编码还具有交叉、变异操作简单的优点，因而，文中采用二进制编码方式。

遗传算法在进化搜索求解问题最优解时，基本上不利用外部信息，仅以适应度函数为依据，利用种群中每个个体的适应度值来进行搜索[7]。文中按照案例间的相似度来进行聚类分析。案例间的相似度描述了案例在多维空间中的距离，相似度越大，则距离越近，属于同一类的可能性就越大。假定 R^s 将空间中 n 个点 $\{x_1,x_2,\cdots,x_n\}$ 分为 c 个类，第 s 代第 g 个体的适应度函数定义为：

$$f_{sg}=\left(\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}D_{ij}(y_i,x_j)\right)^{-1} \tag{1}$$

$$D_{ij}(y_i,x_j)=\sum_{k=1}^{m} w_k * D^2(y_{ik},x_{jk}) \tag{2}$$

其中，(1) 式中 $u_{ij}=0$(或 1) 表示案例 x_j 不属于(或属于) 第 i 个聚类中心 y_i，$\sum_{i=1}^{c} u_{ij}=1$；(2) 式表示案例 x_j 到聚类中心 y_i 的欧氏距离，m 是案例所包含的属性个数，w_k 是第 k 个属性的权重，权值可以通过文献[8]的方法来确定。

综上所述，将混合 TS 和 GA 的聚类算法(TSGAC)的主要过程描述如下：

输入：样本个数 n，种群大小 p，聚类个数 c，最大遗传代数 G，禁忌表长度 L，交叉概率 pc，变异概率 pm

输出：每代最优个体的适应度值，最后一代得到的最优聚类中心

过程：

① 读取数据。从案例库中随机选取 n 个案例并编码，作为样本进行聚类分析，即产生 $X[n]$；

②$s=0$，随机产生初始群体，规模为 p，即产生 $Y_s[p]$；禁忌表 TL1、TL2 置空；

③ 根据公式(1)计算 Y 数组中每个对象 $Y_s[g]$ 的适应值 $f_{sg}, g = 1, \cdots, p$;

④ 采用赌轮选择法及 TSC、TSM 算子,对 Y 数组优化,产生新一代种群;

⑤ $s = s + 1$;如果 $s < G$,转 ③;否则,输出最后一代中的最佳聚类中心;

⑥算法终止。

2　基于 TSGAC 的案例分类和检索

基于上述方法,将案例检索的过程分步进行,其过程如图 1 所示。

具体可以描述为:

(1)在进行案例检索前,利用 TSGAC 算法对案例库进行分类,将案例聚类的最优结果记录下来;

(2)将新案例与各聚类中心案例进行相似度计算,相似度最大的那个聚类就是新案例所在的分类;

(3)采用基于最近邻的搜索方法计算新案例与该聚类中所有案例的相似度,找到的相似度最大的源案例即为所求的候选案例。

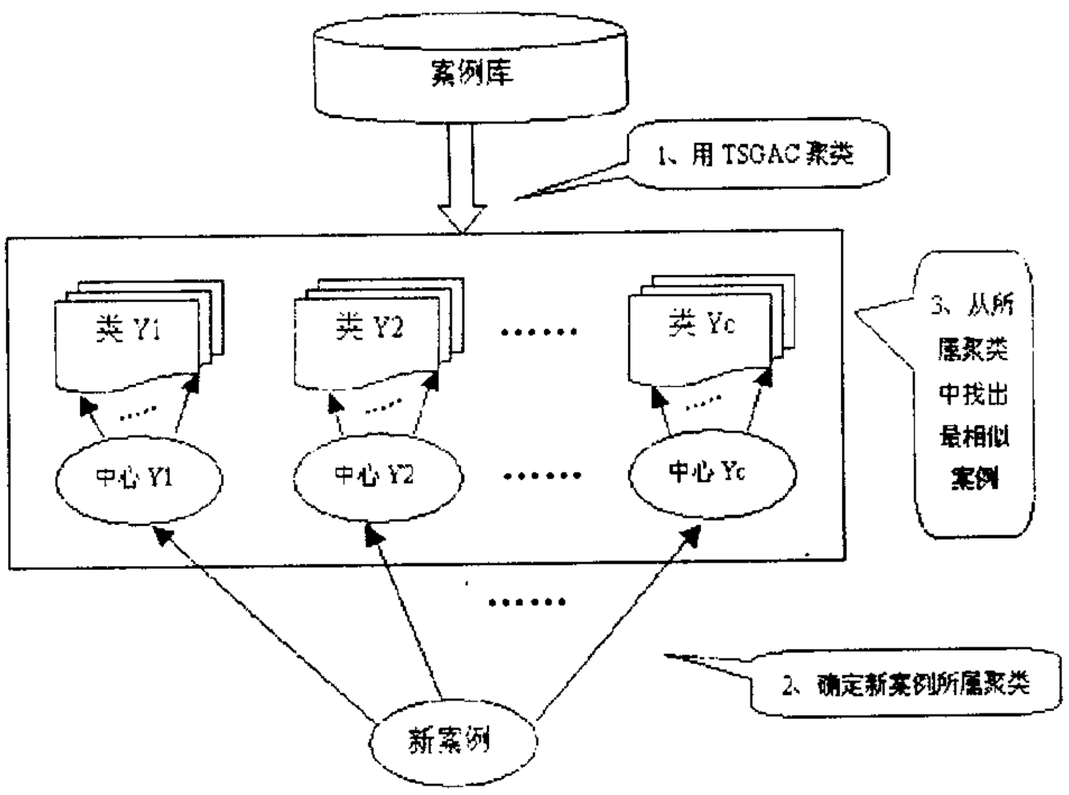

图 1　案例库检索系统

3　实验与讨论

采用农业气象数据[8](取自“安徽省农业气象灾害数据库”)进行实验来验证上述算法的性能。农业气象数据库存有大量的气象记录,相关的气象要素可以用来对气象灾害进行预测。这批数据都是关于气象指标方面的,包含地区、季节、气温、降雨量、日照量以及气象灾害等共计 16 个属性。文中的主要工作就是通过新算法来确定目标案例的最佳匹配案例。

首先从库中提取部分案例组成一个样本集。样本中的案例采用二进制编码,编码的每个分量表示该案例的相应特征项的值。在样本集中,随机产生 c 个聚类中心作为初始种群,采用 TSGAC 进行聚类处理,然后利用分类结果进行案例提取,检索出新输入的案例的最近似案例并输出。

为了测试文中算法的优越性,以分类精度和平均运行时间作为标准,将基于 TSGAC 的案例检索过程与最近邻算法的比较,结果如表 1 所示。

由表 1 可以看出,TSGAC 的分类精度与最近邻方法相当,但是 TSGAC 的平均运行时间却明显少于最近邻算法。实验结果表明,在案例检索过程中引入 TSGAC 预先分类的方法可以在保证检索速度的前提下提高搜索速度。

表 1　实验结果比较

案例数 / 方法	Accuracy(%)		平均运行时间(s)	
	400	800	400	800
最近邻	97.7	99.2	85	163
TSGAC	96.8	99.5	11	23

4　结束语

遗传算法和禁忌搜索各具特点,将两种方法结合在一起提出一种应用于案例库的有效分类方法 TSGAC,并在此基础上进行案例库中的案例检索。合理地将 GA 和 TS 组合在一起,不仅可以提高寻优速度,并且可以保证整体的全局寻优能力。这种新算法不仅适用于 CBR,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] 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.

[2] Han Jiawei, Kamber M. 数据挖掘:概念与技术[M]. 范明,孟小峰,译. 北京:机械工业出版社,2001.

[3] Hall L O, Ozyurt I B, Bezdek J C. Clustering with a genetically optimized approach [J]. IEEE Transaction on Evolutionary Computation, 1999, 3(2):103 - 112.

[4] 张　颖,刘艳秋. 软计算方法[M]. 北京:科学出版社,2002.

[5] Glover F, Laguna M. Tabu Search[M]. Boston: Kluwer Academic Publishers, 1997.

[6] Selim S Z, Ismail M A. K - Means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1984, 6(1):81 - 87.

[7] 傅景广,许　刚,王裕国. 基于遗传算法的聚类分析[J]. 计算机工程,2004,30(4):122 - 124.

[8] 贾兆红,倪志伟,赵　鹏. 用遗传算法来挖掘案例库中的特征项权重[J]. 计算机工程,2003,29(14):71 - 73.