

基于并行遗传算法的规则发现研究

周 勇, 刘 锋

(安徽大学 计算机科学与技术学院, 安徽 合肥 230039)

摘 要:阐述了传统遗传算法的基本思想、原理和步骤及其在数据挖掘(规则集发现)中的应用,给出了基于遗传算法的知识规则挖掘算法的基本思想和关键问题,包括知识规则表示、适应度函数定义等,继而提出多种群并行进化结构,利用精英重组策略,产生池进化模型以及自适应参数的手段调整并行遗传算法进行数据挖掘。在算法具体实现过程中,采用了动态变异交叉概率率等方法,有效避免了并行遗传算法中早熟现象的发生。以北美香菇数据为例,进行并行遗传算法挖掘分类规则,实验说明了该算法在发现和进化规则方面的有效性。

关键词:遗传算法; 并行遗传算法; 规则发现; 数据挖掘

中图分类号: TP18

文献标识码: A

文章编号: 1673-629X(2008)08-0137-03

Research on Rule Discovery Based on Parallel Genetic Algorithm

ZHOU Yong, LIU Feng

(School of Computer Science and Engineering, Anhui University, Hefei 230039, China)

Abstract: Presented the traditional genetic algorithm, the principles and the processing steps of the data mining (rule set discovery). Then proposed the basic thinking and the key problem of this algorithm, including the representation of the rule and the definition of the fitness - function etc. Then used parallel genetic algorithm which added the application of multiple and parallel evolutionary group structure, the elite reorganization strategy, the productive - pool strategy and adaptive parameter adjustment methods for data mining. Meanwhile, in the process of algorithm realization, used the dynamic variation of crossover probability to effectively prevent the genetic algorithm phenomenon of precocious puberty. Finally, used the North - American - mushrooms data as examples to try to use genetic algorithms finding classification rules, and successfully proved the validity of the algorithm in discovering and evolving the rules.

Key words: genetic algorithm; PGA; rule discovery; data mining

1 并行遗传算法

1.1 传统遗传算法

遗传算法(Genetic Algorithm, GA)是1975年由美国Michigan大学教授J. Holland提出的借鉴大自然物竞天演、优胜劣汰的自然选择和遗传机理的人工智能技术^[1],其本质是一种求解问题的高效并行全局搜索方法。它能在搜索过程中自动获取和积累有关搜索空间的知识,并自适应地控制搜索过程以求得最优解。然而,随着问题规模与复杂度的不断提升,GA在应用上仍有许多问题有待研究,一个突出的问题是收敛速度与收敛性之间的矛盾^[2,3]。因此,人们提出并行GA

的模型以望改进算法。

遗传算法作为一种寻优手段,现在已被广泛应用于交通、通信、电力、工程结构优化、计算数学、电子学、材料科学等领域^[4]。

1.2 并行遗传算法的几种进化模型

众所周知,在自然进化过程的任何时刻,总是同时有大量的物种在彼此独立地向前进化,显然,自然界进化过程本身就是一个并行过程,而遗传算法是人们对自然进化过程的机器模拟,其本质上就继承了自然进化所固有的并行性。

目前典型的并行遗传算法(Parallel Genetic Algorithm, PGA)主要有:

- (1)全局单群体主从式PGA(见图1);
- (2)全局单群体细粒度PGA(见图2);
- (3)多群体粗粒度PGA(见图3)。

另外还可以在PGA的算子设计上做出各种改进^[5-8]。

收稿日期:2007-11-10

基金项目:国家自然科学基金(60273043);安徽省教育厅自然科学基金重点科研项目(2006KJ013A)

作者简介:周 勇(1967-),男,安徽合肥人,硕士,讲师,研究方向为机器学习、数据挖掘;刘 锋,博士,教授,研究领域为并行分布计算、计算机网络。

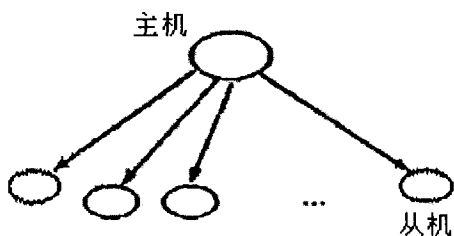


图 1 MS-PGA 的拓扑模型

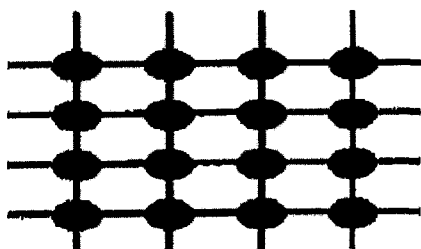


图 2 FG-PGA 的拓扑结构

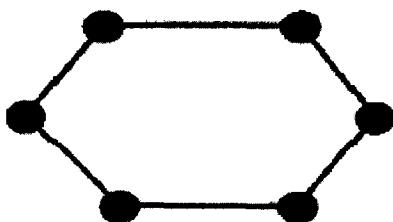


图 3 粗粒度拓扑结构

2 改进的并行遗传算法

2.1 编码方案

使用遗传算法必须把优化问题解的参数形式转换成基因码串的形式,即编码,文中希望发现的对象是分类规则,根据分类规则的特点,采用二进制编码来实现规则(个体)基因码串,一条分类规则分两部分:规则特征部分(前提条件)、规则类别部分(结论)。方便起见,使用枚举型数据集进行规则的发现,一般枚举型数据项可表示为 $\text{if}(\text{attr}_1 = A \& \& \text{attr}_2 = B \& \& \dots) \text{then result} = Z$ 的形式,于是使用如下方法对规则进行编码:

假设某数据集数据有 n 个属性 $\text{attr}_1, \text{attr}_2, \dots, \text{attr}_n$, 属性 attr_i 可取值种数为 $|\text{attr}_i|$, 结果为 result , 结果可取值种数为 $|\text{result}|$, 则使用长度为 $\sum_{i=1}^n |\text{attr}_i| + |\text{result}|$ 的染色体进行编码,其中,若 i 属性的第 j 位为 1 则表示 i 属性取 j 类值,而若 i 属性有 j, k 两位均为 1 ($j \neq k$) 时则表示 i 属性可取 j 类值或 k 类值。另外,如果某特征属性 i 所有位均为 1, 说明该属性无论取何值,不影响规则的成立与否。

2.2 适应度函数的设计与初始种群的生成

遗传算法实现知识库中分类规则挖掘的过程是“好的规则”得以生存,并作为父代规则来进行繁殖、交叉、变异,直至找到最优规则集为止。所谓“好的规则”

是指规则与测试数据集中各实例匹配程度高,适应度函数应能反映出规则对数据集的匹配程度。从以下四个因素设计适应度函数以综合评估规则:

1) 正确分类率 α : 符合该染色体规则且结果一致的数据项个数 / 总个体数,适应度函数值与正确分类率呈正相关;

2) 错误分类率 β : 符合该染色体规则但结果错误的数据项个数 / 总个体数,适应度函数值与正确错误率呈负相关;

3) 染色体有效长度 χ : 为避免过度拟合,希望规则的有效长度越短越好。

有效长度 len 计算方法如下:

χ : = 染色体长度;

for $i = 1$ to 属性个数 n

if (属性 i 全为 1)

then χ : = χ - 属性 i 的长度

染色体有效长度与适应度函数值呈负相关;

4) 染色体规则覆盖度 δ : 即符合该染色体规则的数据项数目,与适应度函数值呈正相关,故最终设计适应度函数 $\text{fitness} = w_1 \times \alpha - w_2 \times \beta - w_3 \times \chi + w_4 \times \delta$, 其中 w_i 为参数因子。

3 基于并行遗传算法的规则发现算法

应用并行遗传算法进行规则发现的算法如下:

输入: 测试数据集、遗传算法控制参数(进化代数、交叉变异率等)。

输出: 进化后的最优分类规则集。

Step1 初始化群体: 随机生成 N 条染色体(二进制基因串), 对规则进行有效性处理(检测纠错);

Step2 适应度值计算: 计算当代群体中各个体(规则)的适应度;

Step3 若当前进化代数 gen 到达设定最大进化代数 maxgen 或某种控制参数符合设定要求, 转至 Step7, 否则继续;

Step4 对该代进行选择、交叉、变异操作, 若该代数为 10 的整数倍, 则进行迁移操作, 进行检测纠错处理并生成子代群体;

Step5 用子代群体中适应度高的个体替代父代群体中适应度低的个体, 形成新一代;

Step6 转至 Step2;

Step7 输出最终规则集。

4 实验结果及分析

实验平台为 Windows XP SP2, CPU PIV 3.0GHz, 1G 内存, Dev-C++ 4.9.9.2 下编译运行。以北美

香菇数据为例,数据项有 22 个分类属性和 1 个结果属性,结果属性是用于区别香菇是否有毒(可食),测试集大小为 8124,其中的 2480 个测试数据有缺失属性项。实验种群大小取 50,交叉概率 0.45,变异率 0.12(高)、0.015(低),进化代数分别取 200、500、1000。从 200 代进化实验挖掘出的结果规则,可以看到,在筛选出的大小为 5 的规则集中,最优的规则适应度为 151.48,它一共覆盖了 2352 个数据项,所有数据项均被正确分类,分类正确率为 100%,而次优规则也分别覆盖了 2040 个数据项和 1920 个数据项,分类正确率同样均为 100%。挖掘出规则后可以再进行解码,将二进制的规则序列解释为文字信息,例如图 4 和图 5 中的最佳规则可以解释为:

```

y
1001111111111001110111111011001011011111010110110011100111011111110110101111001
101111000101001111111110001110111111101100111 151.48
1101101111011000100111110010011110111110110010011100111011111110110111010100
10111010010100111011111100011101111111001000011 139
1101101111011000101111110010010110111110110010011100111011111110110111010100
10111010010100111011111100011101111111001000011 139
1101101111011000101111110010010110111110110010011100111011111110110111010100
10111010010100111011111110011101111111001000011 139
11011011101001000100111101001101011011011011001001111011111101111101101111001
10111010010100011001111101101111111101100111 131.8
want to see details?Y/N
y

```

图 4 200 代进化结果规则

```

Rule 0's condition :
computing...
This chrom can cover 2352 test rules
2352 of them are right,whereas 0 of them are wrong
this chrom's fitness is: 151.48

Rule 1's condition :
computing...
This chrom can cover 2040 test rules
2040 of them are right,whereas 0 of them are wrong
this chrom's fitness is: 139

Rule 2's condition :
computing...
This chrom can cover 2040 test rules
2040 of them are right,whereas 0 of them are wrong
this chrom's fitness is: 139

Rule 3's condition :
computing...
This chrom can cover 2040 test rules
2040 of them are right,whereas 0 of them are wrong
this chrom's fitness is: 139

Rule 4's condition :
computing...
This chrom can cover 1920 test rules
1920 of them are right,whereas 0 of them are wrong
this chrom's fitness is: 131.8

请按任意键继续...

```

图 5 200 代进化规则

if ((cap - shape = bell or conical or convex or flat)
and (cap - color = brown or cinnamon or gray or green or

red or white or yellow) and) then 此类蘑菇无毒
由此实验结果可以看出:

1) 随着进化代数的增加,群体的平均适应度得到了优化;

2) 文中设计的编码方案和适应度函数能够较好地应用于规则发现问题上,进化出的规则在分类正确率和覆盖度上均令人满意;

3) 对两个数据集进行运算后,不仅得到了可行解,而且得到了令人满意的最优解;

4) 从以上的仿真结果也可以看出,利用并行遗传算法来自动地生成规则是用来解决规则发现问题的一个切实可行的方法,相信在实际中的应用将会得到更好的发展。

5 结束语

提出一种基于并行遗传算法发现分类规则的方法。通常其它分类算法如决策树算法在运行后往往可以得到多条分类规则,从而使用户在规则选择上无所适从;而本算法在运行后只得到一个

相互无关联的规则集(当然也可以只有一条“最佳规则”),从而使用户明确了对规则的选择。算法的另一个优点是用户可以通过适应度函数的恰当设计得到不同性质的分类规则,既可以选择发现差错率小的规则,也可以发现在一定容错范围内覆盖面广的规则。传统分类算法通常强调分类规则的准确性,如决策树算法、粗糙集算法在运行后都能得到高准确率分类规则,但发现其它性质的规则如有趣分类规则和易于理解的分类规则却具有一定的难度。然而文中方法只需通过对适应度函数的简单设计来发现这些性质的分类规则。另外,由于采用并行、精英选择以及产生池等策略,使得整体算法在收敛速度、避免局部解等方面性能突出。

参考文献:

- [1] 蒙祖强,蔡自兴.一种新的基于遗传算法的数据分类方法[J].小型微型计算机系统,2004,25:690-693.
- [2] 肖勇,陈意云.用遗传算法构造决策树[J].计算机研究与发展,1998,35(1):49-52.
- [3] Yang Qing, Yang Yuexiang. Learning algorithm based on decision tree[J]. Journal of Xiangtan Normal University, 1999, 20(3):56-60.
- [4] Holland J H. Adaptation in natural and artificial systems[M].

(下转第 181 页)

using CrystalDecisions.Shared;

using CrystalDecisions.ReportSource;

using CrystalDecisions.CrystalReports.

Engine;

(12)在 prixueji.aspx.cs 中的部分关键编码为:

//设置 SQL 语句

```
Session["bindSQL"] = String.Format
("SELECT. 课程编号, 课程名, 学分, 学时, 类别, 临时_成绩表" + Session["学号"].ToString().Trim() + ". 成绩, 临时_成绩表" + Session["学号"].ToString().Trim() + ". 姓名, 临时_成绩表" + Session["学号"].ToString().Trim() + ". 研究方向, 临时_成绩表" + Session["学号"].ToString().Trim() + ". 专业变更, 临时_成绩表" + Session["学号"].ToString().Trim() + ". 入学日期 FROM 信息_课程设置 JOIN 临时_成绩表" + Session["学号"].ToString().Trim() + "ON 信息_课程设置. 课程编号 = 临时_成绩表" + Session["学号"].ToString().Trim() + ". 课程编号");
```

//设置连接信息

```
SqlConnection conn = new SqlConnection(Db.ManagerConnectionString);
```

```
SqlDataAdapter da = new SqlDataAdapter ( Session ["bindSQL"].ToString(), conn);
```

```
DataSet ds = new DataSet();
```

//连接到数据库,从数据库中获取数据然后断开数据库连接

接

```
da.Fill(ds, "prixueji_d");
```

```
prixueji_m report = null;
```

```
report = new prixueji_m();
```

//使用“报表引擎”对象模型将填充的数据集传递给报表

```
report.SetDataSource(ds);
```

//将带有数据的报表对象绑定到 Web 窗体查看器

```
CrystalReportViewer1.ReportSource = report;
```

(13)编译该解决方案,运行结果如图 3 所示。

4 结束语

采用 Crystal Reports for Visual Studio. NET 开发

打印学籍 Microsoft Internet Explorer

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

地址(D): http://localhost/msg/sd2/stud/prixueji.aspx

安徽大学硕士研究生学籍表

学科专业: 计算机软件与理论 研究方向: 数据库与web开发技术
研究生姓名: 沈亚萍 入学日期: 2006年9月

	课程名称	学时	学分	成绩
公共课	英语课程	180	4	82
	科学社会主义理论	36	1	94
	自然辩证法概论(理科)	54	2	84
	英语学位			提前通过
	组合数学	72	4	88
专业基础课	高级人工智能	54	3	90
专业必修课	计算机语言理论	54	3	93

图 3 运行结果

Web 报表,利用它本身提供的报表设计器和.NET 提供的丰富特性,可以极大地提供开发效率,使得开发人员可以从一些底层烦琐的编程任务中解脱出来,专注于实现用户特定的要求,且不需要用户的参与。该设计用动静结合的办法设计和实现了动态报表,并且已经应用到安徽大学研究生部培养与管理系统中,并取得了很好的效果。

参考文献:

- [1] 崔亮,郭忠文.基于.NET平台的Web报表打印方法研究[J].计算机应用,2003,23:332-334.
- [2] 李云亮,李相枢..NET环境下两种Web报表解决方案的对比分析[J].计算机应用研究,2004(6):212-214.
- [3] 唐敏.Web报表工具及其支撑框架的设计与实现[D].北京:北京航空航天大学,2001.
- [4] 陈传波,黄刚,刘清慧.一种基于ASP.NET的自定义报表的设计与实现[J].计算机工程与科学,2006(6):112-114.
- [5] 章立民.用实例学 crystal Report for visual studio. net[M].北京:电子工业出版社,2004:10-60.

(上接第139页)

Ann Arbor, MI: The University of Michigan Press, 1975.

- [5] 赖鑫生,张明义.基于渗透原理迁移策略的并行遗传算法[J].计算机学报,2005,28(7):1146-1152.
- [6] 管宇,徐宝文.基于模式迁移策略的并行遗传算法[J].计算机学报,2003,26(3):294-301.

- [7] 欧阳森.一种新的改进遗传算法[J].计算机工程与应用,2003,39(11):13-15.
- [8] 戴晓明,陈昌领,邵惠鹤,等.粗粒度并行遗传算法收敛性分析及优化运算[J].上海交通大学学报,2003,37(4):499-502.