DNA计算在信息安全中的应用

刘汝正
(广东海洋大学网络与教育技术中心 广东湛江 524088)

摘 要 随着Internet的飞速发展,计算机技术的应用走进千家万户和大中小企业,开放的信息系统必然存在众多的安全隐患,信息安全防范备受关注,与此同时,新兴的智能计算方法的研究也逐渐被应用到信息安全防范中来。本文介绍了当前DNA计算方法在信息安全中的应用,阐述了其工作原理和特点,同时对DNA计算方法的发展进行了展望。

关键词 遗传算法,DNA算法,信息安全

Application of the DNA Algorithm in Information Security

LIU Ru-zheng
(Network and Education Technology Center, Guangdong Ocean University, Guangdong Zhanjiang 524088, China)

Abstract With the rapid development of internet, application of computer technology has walked into thousands of households and the large and medium-sized enterprises, and open information systems must exist many security risks, so information security has been concerned. Meanwhile, emerging Intelligent method of calculation is gradually applied to the security of information. This paper introduces the application of the DNA algorithm in information security currently, and describes the working principle and characteristics of it, at the same time the development of the DNA algorithm is prospected.

Keywords Inheritance algorithm, DNA algorithm, Information safety

1 前言

随着Internet的迅速发展,电子商务已成为潮流,人们可以通过互联网进行网上购物、银行转账等许多商业活动。现在商业贸易、金融财务和其他经济行为中,不少已经以数字化信息的方式在网上流动着。在下个世纪伴随着电子商务的不断发展和普及,全球电子交易一体化将成为可能。"数字化经济"(Digital Economy)初具规模,电子银行及电子货币的研究、实施和标准化开始普及。

然而,开放的信息系统必然存在众多潜在的安全隐患,黑客和反黑客、破坏和反破坏的斗争仍将继续。在这样的斗争中,安全技术作为一个独特的领域越来越受到全球网络建设者的关注。

2 DNA遗传算法

遗传算法(Genetic Algorithms or GAs)是基于自然选择和自然遗传机制的搜索算法,它是一种有效的解决最优化问题的方法。遗传算法最早是由美国Michigan大学的John Holland和他的同事及学生提出的。类似于自然界演化的基本法则,"适者生存"是遗传算法的核心机制,同样,"复制(reproduce)"、"杂交(crossover)"、"变异(mutation)"等自然界的生物演化规则在遗传算法中都得到类似的体现。

用遗传算法解最优化问题,首先应对可行域中的个体进行编码,然后在可行域中随机挑选指定群体大小的一些个体组成作为进化起点的第一代群体,并计算每个个体的目标函数值,即该个体的适应度。接着就像自然界中一样,利用选择机制从群体中随机挑选个体作为繁殖过程前的个体样本。选择机制保证适应度较高的个体能够保留较多的样本;而适应度较低的个体则保留较少的样本,甚至被淘汰。在接下去的繁殖过程中,遗传算法提供了交叉和变异两种算法对挑选后的样本进行交换和基因突变。交叉算法交换随机挑选的两个个体的某些位,变异算子则直接对一个个体中的随机挑选的某一位进行突变。这样通过选择和繁殖就产生了下一代群体。重复上述选择和繁殖过程,直到结束条件得到满足为止。进化过程最后一代中的最优解就是用遗传算法解最优化问题所得到的最终结果。

与其他算法相比,遗传算法主要有以下四个方面的不同:遗传算法所面向的对象是参数集的编码,而不是参数集本身;遗传算法的搜索是基于若干个点,而不是基于一个点;遗传算法利用目标函数的信息,而不是导数或者其他辅助信息;遗传算法的转化规则是概率性的,而不是确定性的。

2.1 DNA计算基本思想

生物进化中采用的许多处理模式已经被人们应用于智能系统,较成熟的有神经网络、演化算法、蚁群算法等软计算技术。利用DNA计算的巨大并行性、高的能量效率和存储容量,尝试开发实际的DNA计算,能够获得更灵活更高效的计算技术。

DNA计算的操作对象是DNA(脱氧核糖核酸)。DNA的基本元素是核苷酸,由化学结构的不同,核苷酸被分为四类碱基(bases):腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(c)、胸腺嘧啶(T)。DNA由两条核苷酸链组成,每条染色体是一段双股螺旋的DNA。一股的碱基序列与另一股的碱基序列互补;A与

刘汝正 讲师,主要从事面向对象技术研究。

T互补，c与G互补。在mRNA中连续排列着由三个连续的碱基组成密码子，这些密码子是氨基酸的代码，60种密码子对应20种氨基酸。氨基酸被用于合成蛋白质。图1给出了碱基互补配对图示意(图1)。

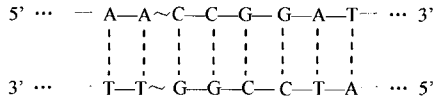

图1 碱基互补配对图

DNA算法解决计算问题的基本思想是：利用DNA特殊的双螺旋结构和碱基互补配对原则对问题进行编码，把要运算的对象映射成DNA分子链，在DNA溶液的试管里，在生物酶的作用下，生成各种数据池(Data Pool)，然后按照一定的规则将原始问题的数据运算高度并行地映射成DNA分子链的可控的生化过程。最后，利用分子生物技术如聚合酶链反应PCR(Polymerize chain reaction)、聚合重叠放大技术POA(Parallel Overla PAssembly)、超声波降解、亲和层析、克隆、诱变、分子纯化、电泳、磁珠分离等，破获运算结果。

2.2 DNA计算与遗传算法

遗传算法(GA)是模拟生命的自适应和进化过程设计的一种软计算方法。遗传算法中，问题的最优解的搜索是通过问题进行编码组成染色体空间，然后对染色体进行交叉和变异算子(或者还有倒位算子)等操作，通过不断获得新的染色体空间而得到问题的最优解。遗传算法在算法实现上具备结构上的隐含并行性、计算原理上的随机性和自适应性，对非线性复杂问题有全局搜索能力，简单通用、鲁棒性强，已成为学术界和工程界研究重点。

近年来，遗传-模糊和遗传-神经混合方式的研究和应用非常活跃。遗传算法在这类混合系统的设计中存在以下特点：

1)遗传算法作为一种全局搜索和优化技术融入系统中，它本身不能表达知识，但具有较强的学习能力和优化能力，基于遗传算法的混合系统设计必须考虑能力互补的问题；

2)遗传算法的编码方法、遗传操作算子需要进行适当的调整；

3)尽可能将现有的近似推理技术和优化技术有机地集成到混合系统中。

DNA计算是新型的智能计算方法。DNA计算一般可概括为3个基本步骤，即：

1)分析要解决的问题，采用特定的编码方式，将该问题反映到DNA链上，并根据需要合成DNA链；

2)根据碱基互补配对的原则进行DNA链的杂交，由杂交或连接反应执行核心处理过程；

3)得到的产物即为含有答案的DNA分子混合物，用提取法或破坏法得到产物DNA。在提取前通常采用PCR技术或克隆技术扩增DNA的量，提取分析后得到的结果，常用的提取方法有凝胶电泳。破坏法的目的是排除非答案DNA分子，如使用限制性核酸酶切断干扰DNA链。如果待计算的问题比较复杂，经一轮的核心处理和提取分析只能得到中间结果，那么可重复步骤2)，3)直到得到满意结果为止。

从DNA计算与遗传算法所依赖的进化机理角度看，遗传算法最适合于采用DNA计算技术，设计基于DNA计算技术的混合遗传算法。图2给出了DNA计算与遗传算法的计算过程异同。

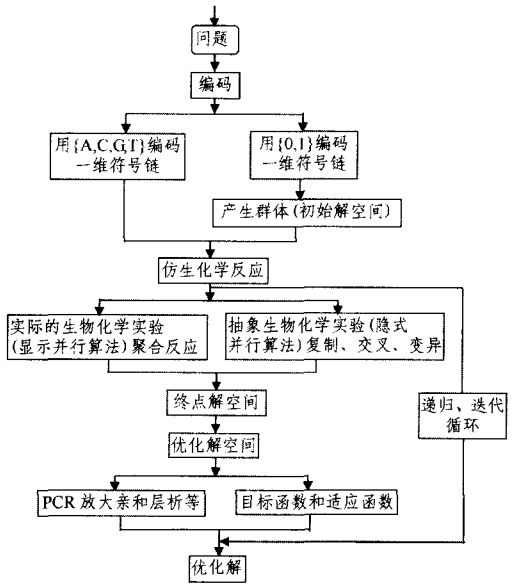

图2 DNA计算与遗传算法比较

3 DNA算法在信息安全中的应用

3.1 TEIRESIAS算法

TEIRESIAS算法最初用于DNA的识别，该算法的运行可划分两个阶段：扫描(scanning)和卷积(convolution)。在扫描阶段，现频率大于某特定阈值的基本模式(elementary pattem)被识别出来。这些基本模式将被组合成越来越大的模式，直到所有存在的最大模式(maximal pattem)都被构造出来。

设∑为字母表(比如，所有英文字母)。一个模式被定义为符合下列形式的任意字符串：

∑(∑{'.'}) * ∑

即任何以字母开始和结束的字符串，并且包含一个可选的由字母和"."组成的字符集。这里，'.'(任选字符)表明此位置可以被任意字母占据。

关于模式的定义：

(1)一个模式P实际是一个正则表达式。因此它定义了一个语言$G(P)$。这个语言的元素是所有可通过用字母表∑中的字符去替换模式P中任选字符而得到的字符串。

(2)对任意模式P，它的每个本身也构成模式的子字符串叫作P的子模式。

(3)一个模式P被称为是一个(L,W)模式($L \leqslant W$)，如果P的每个长度大于或等于W的子模式包含至少L个字符。

(4)给定一个模式P以及一个字符串集合$S=\{s_1,s_2,\cdots,s_n\}$我们这样定义P关于S的偏移序列：

$Ls(P)=\{(i,j)|s_i$ 在位置j 匹配$P\}$

(5)如果P'可以通过如下方式从P中得到。则认为模式P'比模式P更具体：将P中一个或多个任选字符('.')更换成确定的字符；或者添加一个由字符和任选字符组成的字符串到P的左边和(或)右边。

(6)给定一个输入字符串集合S，一个模式P被称为对S是最大的。如果不存在另一个模式P'比P更具体。换句话说，如果P是最大模式，除非同时缩减其偏移序列的大小。

(下转第118页)

gent系统涌现出一定的智能性，能够解决传统技术无法解决的复杂问题，其思想十分适合大规模故障诊断系统的设计和实现。为解决复杂装备日渐突出维修保障问题，作者研究了MAS故障诊断系统，重点探讨了任务分解与分配策略。与传统故障诊断系统相比，MAS故障诊断系统具有诸多优点：集体智能性、可扩充性、资源重用能力和容错性等。发展基于多Agent系统的远程故障诊断系统也符合复杂装备维修保障需求的网络化和智能化趋势。在某国家部委项目的应用中，该原型系统显示出了较高的准确性和智能性。

参 考 文 献

[1] Wooldridge M. 石纯一，等译. An Introduction to MultiAgent Systems [M]. 北京：电子工业出版社，2003

[2] 高庆，李田，魏震生. 层次分析在故障诊断中的应用[J]. 火力与指挥控制，2006(2)

[3] 蔡自兴，姚莉. 人工智能及其在决策系统中的应用[M]. 长沙：国防科技大学出版社，2006

[4] 李千目，戚湧，张宏，等. 基于粗糙集神经网络的网络故障诊断新方法[J]. 计算机研究与发展，2004(10)

[5] Leonardi E, Mellia M, Neri F. Bounds on Average Delays and Queue Size Averages and Variances in Input-queued Cell-based Switches [J]// Proc. IEEE INFOCOM. 2001(2): 1095-1103

[6] 刘建辉，张俊利，王爽. 基于Agent的远程协同故障诊断系统研究[J]. 计算机测量与控制，2006(1)

[7] 严建峰，李伟华，杜北. 基于规模压缩的混合蚁群算法[J]. 控制与决策，2007(9)

[8] Guerin R, Peris V. Quality-of-service in Packet Networks: Basic Mechanisms and Directions [J]. Computer Networks, 1999, 31(3): 169

(上接第 104 页)

否则它不能被更具体化。

使用上面的定义，我们可以简洁地描述 TEIRESIAS 问题：

问题定义：给定一个输入字符串集合 $s=\{s_1,s_2,\cdots,s_n\}$ 以及参数 L,W,K。找到所有最大的〈L,W〉模式，且这些模式最少支持 K(即：它们至少出现在属于 S 的 K 个独立字符串中)。

3.2 DNA 算法在邮件过滤上的应用

DNA 算法是解决搜索问题的一种通用算法，对于各种通用问题都可以使用。搜索算法的共同特征为：①首先组成一组候选解；②依据某些适应性条件测算这些候选解的适应度；③根据适应度保留某些候选解，放弃其他候选解；④对保留的候选解进行某些操作，生成新的候选解。

大量的病毒通过邮件这个载体感染计算机，于是反垃圾邮件技术就成为了信息安全防范的一个重要措施和研究热点，目前常用的垃圾邮件过滤技术包括白名单和黑名单、内容过滤、身份验证等，基于上述 DNA 算法的特性，下面就用 DNA 算法中的 TEIRE1SIAS 算法来实现在邮件中对垃圾邮件的过滤。具体做法是从训练集合中识别出正常邮件和垃圾邮件的模式集合，采用模式匹配的方法识别垃圾邮件。

3.2.1 邮件的表示

采用贝叶斯分类，邮件 D_i 被表示为一个向量 $D_i=\{W_{1i},W_{2i},\cdots,W_{ni}\}$，这里 $W_{1i},W_{2i},\cdots,W_{ni}$ 分别表示邮件 D_i 在属性 $X_1,X_2,\cdots,X_n$ 上的权重。将邮件转换成这样一组向量需要三个步骤：从训练集中得到所有符合一定条件的属性；根据某算法，选择一定数量的属性构成属性集；根据邮件对属性集中所有属性的匹配情况，计算邮件对各属性的权值。下边分别介绍这三个步骤。

(1)传统的贝叶斯垃圾邮件过滤中，采用词作为属性。对于英文，词与词之间有空格隔开，词语的抽取非常容易。处理中文邮件则需要引入中文词法分析系统，以从邮件中抽取中文词语。基于模式的贝叶斯垃圾邮件过滤是以模式(正则表达式)作为属性。本文采用第 2 节所介绍的 TEIRESIAS 算法从邮件中获取模式。首先获取出现次数超过一定阈值 K 的基本模式，再通过卷积得到所有最大模式。

(2)本文采用信息增益算法来选择属性。计算每个属性的信息增益值，选取具有最大增益值的Ⅳ个属性构成用于过滤的属性集。信息增益值，G 的计算公式如下：

$$IG(X)=\sum_{x\in\{0,1\},c\in\{SPAM,HAM\}}P(X=x,C=c)g\log\frac{P(X=x,C=c)}{P(X=x)gP(C=c)}$$

(3)最简单的权值计算办法是布尔方法：邮件中出现某个属性。邮件对其权为 1，否则为 0。采用布尔方法。

3.2.2 训练和分类算法

根据贝叶斯公式，邮件属于某类的概率可以通过下式计算：

$$P(C=c/D_i)=\frac{P(C=c)P(D_i/C=c)}{P(D_i)}$$

上式中，$P(C=c)$和 $P(D_i)$都容易根据训练集得出，问题的关键在于计算 $P(D_i|C=c)$。朴素贝叶斯假设所有属性是相互独立的，于是有：

$$P(D_i/C=c)=\prod_{i=1}^{n}P(X_i/C=c)$$

一般认为将一封正常邮件错判为垃圾邮件所造成的损失要远大于漏过一封垃圾邮件。假没错判一封正常邮件的代价是漏过一封垃圾邮件代价的 A 倍，则一封邮件 D_i 被分类器判断

为垃圾邮件应该满足：

$$\frac{P(C=SPAM/D_i)}{P(C=HAM/D_i)}>\lambda$$

由于 $P(C=SPAM/D_i)=1-P(C=HAM/D_i)$，设 $t=\frac{\lambda}{1+\lambda}$，则将 D_i 判断为垃圾邮件的前提是 $P(C=SPAM/D_i)>t$。

结束语 信息安全是一个综合性的课题，涉及技术、管理、使用等许多方面，既包括信息系统本身的安全问题，也有物理的和逻辑的技术措施，一种技术只能解决一方面的问题，而不是万能的。目前更多更新的算法方法都被应用到信息安全的各个方面中来，只有严格的保密政策、明晰的安全策略以及高素质的人才才能完好、实时地保证信息的完整性和确证性。

参 考 文 献

[1] Adleman L M. Molecular computation of solutions to combinatorial problems[J]. Science, 1994, 266(5187): 1021-1023

[2] Lipton R J. DNA solution of hard computational problems[J]. Since, 1995, 268(5210): 542-545

[3] 胡忠望. 互联网上隐私信息的安全问题研究[J]. 微计算机信息，2006(7): 94-97

[4] 陆正伟. 一种应用免疫原理的入侵检测原型系统[J]. 微计算机信息，2006(7): 97-100

[5] 李士勇. 模糊控制神经控制和智能控制论. 哈尔滨工业大学出版社，1998: 116-119

[6] 任立红，等. 采用 DNA 遗传算法优化设计的 TS 模糊控制系统. 控制与决策，2001, 16(1): 16-24