第16卷 第8期
2006年8月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.8
Aug. 2006

遗传算法在决策支持系统和数据挖掘中的应用

王淑静,贾兆红,王 亮,俞 磊
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:决策支持系统和数据挖掘技术的应用已深入到企业管理、金融、办公及日常生活等各个领域。使用原先决策支持系统中的传统方法已远远不能满足决策者的需要,从而出现许多新技术新方法来辅助和完善决策支持过程。文中在介绍遗传算法的基础上,提出了一种将遗传算法与决策支持系统相结合的观点,阐述了基于遗传算法的决策支持系统的模型设计,且深入讨论了遗传进化技术在数据挖掘中的应用。

关键词:遗传算法;决策支持系统;数据挖掘;关联规则
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2006)08-0035-03

Application of Genetic Algorithms in Decision Support System and Data Mining

WANG Shu-jing, JIA Zhao-hong, WANG Liang, YU Lei
(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: At present, the application of decision support system and data mining technology have penetrated into every field of daily life, including enterprise managenent, finance, office business and so on. The traditional methods which were adopted for previous decision support system can not satisfied the decision maker, so new technology and technique appeared for assisting and perfecting the process of decision-making. Under the discussion of genetic algorithm, propose a new viewpoint about a novel decision support systems based on genetic algorithms. In this paper elaborate on the modeling design of decision support systems based on genetic algorithms, and discuss the applications of genetic and evolution technology in data mining.

Key words: genetic algorithm; decision support system; data mining; association rule

0 引 言

遗传算法(Genetic Algorithm,GA)是近几年发展起来的一种崭新的全局优化算法,它借用了生物遗传学的观点,通过自然选择、遗传、变异等作用机制,来实现解决适应性问题的计算。由于具有鲜明的生物背景,遗传算法尤其适用于求解大规模、高度非线性、不连续的最优化问题。

决策支持系统(Decision Support Systems,DSS)是在管理信息系统和运筹学的基础上发展起来的,主要由计算机自动组织和协调多模型的运行和存取,处理数据库中的大量数据,达到高层次的决策能力。目前决策支持系统的应用已深入到企业管理、金融、办公及日常生活等各个领域,为经济发展、社会进步做出了重大的贡献。然而随着应用的不断拓展,决策过程中出现的信息越来越多,系统也越来越复杂,使用原先决策支持系统中的传统方法已远远不能满足决策者的需要,从而出现许多新技术新方法来辅助和完善决策支持过程。

将GA与DSS相结合,可以充分发挥GA的特点,为DSS的发展提供了一个全新的方向。此外,随着数据挖掘(Data Mining,DM)方法的不断发展,已经有越来越多的方法应用到数据挖掘中,其中包括遗传算法,实验证明将GA应用到DM中将会产生良好的效果。文中将介绍一种基于遗传算法的决策支持系统的建模方法,并简要分析其优缺点,最后对遗传算法在决策支持系统和数据挖掘中的应用特点分别进行了深入讨论,并在此基础上指出有待进一步研究的问题。

1 基于遗传算法的决策支持系统

1.1 遗传算法概述

遗传算法是一类借鉴生物界自然选择和自然遗传机制的随机化搜索算法。该算法是一种群体型操作,这种操作以全体中的所有个体为对象,其操作包括:选择(selection)、交叉(crossover)、变异(mutation)3个主要遗传算子,它们使得遗传算法具有了其他传统方法所没有的特性[1]。遗传算法的核心内容是它的5个基本要素,它们是:参数编码、初始群体的设定、适应度函数的设计、遗传操作的设

收稿日期:2005-11-11
基金项目:安徽省教育厅自然科学基金项目(05010703,2005kj055)
作者简介:王淑静(1982-),女,安徽巢湖人,硕士研究生,研究方向为机器学习与智能软件开发;导师:倪志伟,副教授,研究方向为专家系统、机器学习与知识发现。

计、控制参数的设定。

遗传算法的特点可以概括成下面的 4 点。

(1)利用变量的编码方式,而不直接使用变量本身;

(2)在解空间中从多出发点搜索问题的解,而不像某些传统的搜索方法从一点出发搜索问题的解;

(3)直接利用目标函数的函数值信息,而不使用函数的导数或其他的辅助信息;

(4)使用概率转移规则,而不采用确定性的转移规则。

1.2 决策支持系统概述

决策支持系统(DSS)是以信息论、管理科学、运筹学和行为科学为基础,以计算机和仿真等技术为手段,综合利用现有的数据和模型,通过人机交互方式辅助解决半结构化和非结构化决策问题的集成系统。决策支持系统主要由以下 3 个部分组成[2],如图 1 所示。

a. 数据库及其管理系统数据部件。

b. 模型库及其管理系统(模型部件)。

c. 对话系统(对话部件)。

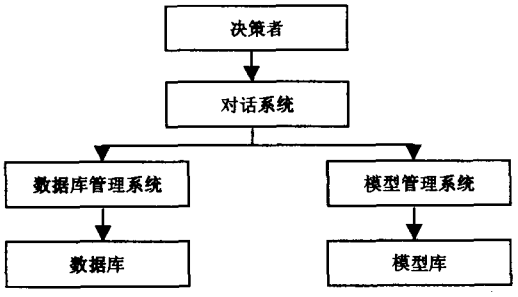

图 1 决策支持系统的基本结构

(1)数据部件。

数据是辅助决策的关键,因此数据部件是 DSS 不可缺少的重要部分。数据库子系统包括数据库和数据库管理系统,主要负责存储和管理 DSS 使用的各种数据以及系统之外的数据,并实现各种不同数据源间的相互转换。

(2)模型部件。

模型库子系统是决策支持系统的核心,同时也是较难实现的部分。这部分包括模型库和模型库管理系统。模型库中主要存放各种已有的模型,也存放由用户使用建模语言建立的模型。模型库管理系统支持决策问题的定义、概念模型化和模型的运行、修改、增删等。模型库子系统与对话子系统的交互作用,可以使用户控制对模型的操作;它与数据库子系统与对话子系统的交互作用,可以提供模型所需的数据,从而实现了模型输入、输出和中间结果的存取。

(3)对话部件。

对话子系统是决策支持系统的人机接口界面,它负责接收和检验用户的请求,协调数据库子系统和模型库子系统之间的通信,为决策者提供信息收集、问题识别以及模型构造、使用、改进、分析和计算等功能。

1.3 基于遗传算法的决策支持系统模型设计及其讨论

在 DSS 模型设计中,采用遗传程序设计可以解决传统建模方法的部分缺点和局限性,更灵活地处理遗传算法中的表示问题,不像常规遗传算法那样采用确定长度的染色体串。这种程序设计由适合问题域的函数和端点构成[2,3]。

模型库与数据库类似,都是一种共享资源,数据库是将众多的数据按一定的结构形式组织起来,模型库是将众多的模型按一定的结构形式组织起来,并通过模型库管理系统对各个模型进行有效的管理和使用。通过模型库可以将多个模型组合起来构成更大的模型。模型库中的模型种类很多,有数学模型、数据处理模型、图形和图像模型、报表模型、智能模型等[1]。

一般模型的建立分为两步[4,5]:

(1)建立模型的数学结构,即针对具体某个领域具体某个问题建立起可行模型中的变量之间的模型方程形式,如线性方程、非线性方程、微积分方程等。

(2)确定模型的参数。其中需要确定的参数主要包括模型中方程的个数、不同参数的变化范围、变量的数目、变量的系数等。然后将确定了的一批可行模型方程 $Y_i = f_i(X_j)$(其中 X_j 为可能参数及实体数据参数)放入模型库中,通过遗传算法的运算,在模型库中匹配出最相容的模型。

智能决策支持系统(Intelligent Decision Support System,IDSS)是人工智能技术和决策支持系统相结合的产物,是在 DSS 的基础上集成专家系统后而形成的。

由于 IDSS 的对象是非结构化的决策问题,通常很难找到每一个问题的所谓"标准解",此外,为了提高 IDSS 的适应性(即随着环境改变自身性能的能力),IDSS 需要经常检测解的质量,不断修正知识库的内容。在 IDSS 应用 GA,有以下优点:

①GA 具有自组织性、自适应性和智能性。应用 GA 在求解问题时,算法将利用进化过程中所获得的信息自行组织搜索。GA 的这一特性,使它具有能根据环境的变化而自动发现环境的特性和规律的能力。从而使遗传算法可以用来解决一些复杂的非结构化问题。

②GA 不是采用确定性规则,而是采用概率的变迁规则来指导它的搜索方向。在优化过程中,使搜索的每一步向最终结果靠近的机制或智能性称为搜索的探索性或启发性。GA 采用以适应值为标尺,以概率作为一种工具来引导搜索过程,是一种导向随机搜索方法。

③GA 对给定的问题,可以产生很多的潜在解,最终的选择可以由使用者确定。在某些特殊情况下,如多目标优化问题,不止一个解存在,而是有一组近似最优解,这时 GA 对于确认可替代解集而言特别适合。

2 遗传算法在数据挖掘中的应用

2.1 数据挖掘概述

数据挖掘是指从数据集中抽取和精化新的模式,是从大型数据库或数据仓库中提取人们感兴趣的知识,这些知

识是隐含的、事先未知的潜在有用信息。数据挖掘的范围非常广泛,其对象可以是数据库、文本、Web信息、空间数据、图像和视频数据等。知识发现过程可粗略地理解为3步:数据准备(data preparation)、数据挖掘(data mining)、结果解释和评价(interpretation and evaluation)。知识发现的结果可以表示成各种形式,包括规则、法则、科学规律、方程或概念网。

2.2 遗传算法在数据挖掘中的应用

目前已有的数据挖掘方法有很多种,比如统计方法、机器学习方法、神经计算方法等[6]。这里主要介绍遗传算法在数据挖掘中的应用。

在数据挖掘领域,可以将数据挖掘的问题看作是搜索问题,数据库则被看作为搜索空间,发现算法看作是搜索策略。因此将遗传算法引入到数据挖掘中,可以在数据库中进行搜索,对随机的一组规则进行进化,直到数据库可以被该组规则覆盖,从而挖掘出人们需要的隐含在数据库中的规则。

数据挖掘的一个主要的目标就是关联规则的挖掘。在数据挖掘的知识模型中,关联规则是比较重要的一种。关联规则模式属于描述型模式,发现关联规则的算法属于无监督学习的方法[6]。基于遗传算法的方法是运用遗传算法的自适应寻优及智能搜索技术,获取与客观事实最相容的问题解[7,8]。

关联规则发现的主要对象是事务数据库。在数据库的数据挖掘中,关联规则就是描述这种在一个事务中,物品之间同时出现的规律的知识模式,通过对数据库中数据的分析处理,发现不同数据项间所存在的相互依赖关系,以描述数据库中数据项之间的潜在关系的规则。这里使用4个参数来描述关联规则的属性,即可信度(Confidence)、支持度(Support)、期望可信度(Expected Confidence)、作用度(Fit)。可信度表示规则的强度,支持度表示规则出现的频度,期望可信度是理想状态下规则的强度,作用度是可信度与期望可信度的比值。

一个关联规则的形式可表示为:$X_1 \wedge X_2 \wedge \cdots \wedge X_m \Rightarrow Y_1 \wedge Y_2 \wedge \cdots \wedge Y_n$,其中:$X_i(i = 1,2,\cdots,m)$,$Y_j(j = 1,2,\cdots,n)$都是数据库中的数据项。数据项之间的关联规则即是,根据一个事务中某些项的出现可推出另一些项也在同一事务中出现。如果 $X_1 \wedge X_2 \wedge \cdots \wedge X_m$ 出现,那么 $Y_1 \wedge Y_2 \wedge \cdots \wedge Y_n$ 必定出现,这表明数据项 $X_1 \wedge X_2 \wedge \cdots \wedge X_m$ 和数据项 $Y_1 \wedge Y_2 \wedge \cdots \wedge Y_n$ 必然存在着某种联系。

发现关联规则的任务就是从数据库中发现那些强规则,也就是说,发现关联规则的问题就是提出这样一些规则,它们的支持度和置信度分别大于用户指定的支持度和置信度的最小值。

使用遗传算法进行数据挖掘,大致分为下面几个主要步骤:

(1)编码。对实际问题进行编码,编码方法可以是二进制编码,也可以是十进制编码。

(2)定义遗传算法的适应度函数。由于算法用于规则归纳,因此适应度函数由规则覆盖的正例和反例来定义。随机产生一组规则,对每一个规则应用数据库中给定的例子进行判断,根据适应度函数计算其适应度。

(3)选择、交叉、变异操作。应用交叉、变异运算对该组规则进行进化,再利用选择运算产生下一代规则,这样经过若干次迭代后,遗传算法满足终止条件,从而得到一组理想规则。

(4)规则的提取。通过一个规则评价函数来计算给定规则的支持度和可信度,然后根据计算出的支持度和可信度来判定该规则是否满足需要。

(5)规则的优化。应用规则优化算法对所得规则进行优化,淘汰冗余的规则,以得到最简规则。

将该算法应用到农业气象数据中做试验,试验结果证明该算法发现的规则与实际情况相吻合。

3 结束语

简要阐述了遗传算法基本原理、DSS基本架构和数据挖掘的基本概念,给出了一种基于遗传算法的决策支持系统模型的建模方法,并讨论了在IDSS设计中引入遗传算法的优点,最后讨论了遗传算法在数据挖掘中的应用。

参考文献:

[1] 韩世欣,黄梯云,李一军.基于机器学习理论的智能决策支持系统模型操纵方法的研究[J].决策与决策支持系统,1996,6(1):10-18.

[2] 黄梯云.智能决策支持系统[M].北京:电子工业出版社,2001.

[3] 米凯利维茨.演化程序——遗传算法和数据编码的组合[M].周家驹,何险峰译.北京:科学出版社,2000.

[4] 张　雷,郑泽席,宋万德.一种基于遗传算法的决策支持系统建模方法[J].空军工程大学学报(自然科学版),2000,1(3):27-29.

[5] 赵志刚.遗传算法在决策支持系统智能化过程中的应用研究[D].天津:河北工业大学,2002.

[6] 史忠植.知识发现[M].北京:清华大学出版社,2002.

[7] Holte R C. Very simple classification rules perform well on most commonly used datasets[J]. Machine Learning, 1993(11):63-90.

[8] Han Jiawei, Cai Yandong, Nick C. Data-driven discovery of quantitative rule in relation database[J]. IEEE Transactions on Knowledge and Data Engineering, 1993, 5(1):29-40.